THIS WEEK

Antarctic Treaty is cold comfort

Researchers need to cement the bond between science and the South Pole if the region is to remain one of peace and collaboration.

"All the day dreams must go; it will be a wearisome return." Written almost exactly a century ago, in the diary of doomed Antarctic explorer Robert Scott, those words mark the moment at which the British naval officer realized that the game was up. He had lost the race to the South Pole to his Norwegian rival Roald Amundsen, and the return journey was to prove worse than Scott imagined. By the end of March 1912, Scott and his two remaining companions had perished, just 18 kilometres from life-saving supplies. Their ill-fated expedition had a scientific slant too, and the zoologist Edward Wilson, who died alongside Scott, was the first scientist to see the South Pole.

It is perhaps a fitting legacy that the hostile landscape of Wilson's final months is now the scene of unprecedented scientific cooperation. Amid conflicting territorial disputes, the 1959 Antarctic Treaty handed the continent to the world's researchers, with the explicit goal of ensuring "in the interests of all mankind that Antarctica shall continue forever to be used exclusively for peaceful purposes and shall not become the scene or object of international discord". Some 30 countries now operate research bases in Antarctica, and the Iranian Students' News Agency reported last week that Iran intends to open one within three years.

The international rush to Antarctica in the name of science has not gone unnoticed by some with interests beyond research. In an article published in *The Australian* newspaper on 31 December, Sam Bateman and Anthony Bergin of the Australian Strategic Policy Institute in Canberra raised the provocative prospect that countries such as China and India could use bases recently opened there to improve satellite communications to military forces that increasingly depend on space-based infrastructure. "To do so would be at odds with the Antarctic Treaty, but the sparse use of the treaty's inspection mechanisms means that such activity could go undetected," Bateman and Bergin say. "If Antarctic sites take on military significance, we could see a move towards destabilisation of Antarctica as a zone of peace."

That may seem a far-fetched scenario, but events may not have to take such a dramatic turn to undermine the treaty and weaken the monopoly that science currently has on Antarctica. The late Christopher Joyner of the Department of Government at Georgetown University in Washington DC identified three potential challenges to the cooperative spirit of the agreement, driven by the changing global political and economic climate. States might implement national continental-shelf claims in offshore Antarctic waters in pursuit of energy resources, he suggested, or tensions could escalate between Japan and Australia over whaling in Antarctic waters. Joyner's third scenario — widespread and unregulated bioprospecting — is a topic already identified as problematic by treaty members, and one discussed at their annual meeting last summer, held in Buenos Aires. Nearly 200 research organizations from 27 states are carrying out research for commercial purposes in the Antarctic, Joyner said, and one big goal is sifting its plants and animals for beneficial genetic and biochemical resources.

➲ **NATURE.COM**
How exploring launched Antarctic science:
go.nature.com/ppfdmu

Although the Antarctic Treaty seems under no immediate threat, one need only look at the escalating political importance of the Arctic to see the potential for change in the south. Just this week, members of the Arctic Council were due to discuss requests from China and India for official observer status. And even under the strict conditions of the treaty, the appeal of Antarctica to some nations searching for new mineral and energy resources has never been too far from the surface.

"There remains great appeal in the spirit of Scott's expedition — of scientists as explorers."

Scientists can play an important part in preserving the treaty and protecting their unique Antarctic playground, and the promise it continues to offer for research. Interest in the poles is currently sky-high — witness the success and international syndication of the BBC's *Frozen Planet* series — and the more that scientists can promote the work that they do there, the more the bond between the two will be cemented in the public's mind. There remains great appeal in the spirit of Scott's expedition, of scientists as explorers, and in that vein, sometime this year, researchers in Antarctica are likely to be the first to penetrate a sub-glacial lake. The long-standing Russian effort to drill into Lake Vostok stalled again this month, raising the prospect that, if they fail to break through this month, then a British team seeking to investigate Lake Ellsworth could beat them to it in December. Not that priority — the loss of which so devastated Scott — matters in 2012. "It's not a race," a spokeswoman for the British project insists. We believe them, just. ■

Face up to fraud

The UK government and funding agencies must address research misconduct.

Many people in science would rather not talk about the problem of research misconduct, much less act on it. After all, who directly involved would benefit from a serious crackdown? Certainly not the institutions at which the misconduct takes place — they are nominally responsible, but can face legal repercussions, embarrassing headlines and a public-relations disaster if they expose cheating academics. It is much easier to shuffle miscreants out of the side door with vague references and a promise of silence, effectively pushing the problem somewhere else, and onto someone else.

So it is perhaps a sort of progress that the *British Medical Journal* and the international Committee on Publication Ethics were able to

organize a meeting on the subject in London last week, gathering representatives from universities, funders, journals and lobby groups to discuss how the problem could be tackled in the United Kingdom (see *Nature* http://doi.org/hmx; 2012). The meeting broke little new ground, but its organizers do, at least, deserve credit for trying.

A big part of the problem is the lack of perceived risk associated with misconduct. Some fraudulent researchers might be sociopaths who don't care about the rules, but many others simply believe that they can anticipate the outcome of a research project, and see no downside to fabricating the required results to save time, or tweaking results to achieve a stronger signal. Either way, stronger action and punishments are needed to discourage such misbehaviour. (Meanwhile, for colleagues considering blowing the whistle, the risks are glaringly huge — witness the plight of scientists, such as cardiologist Peter Wilmshurst, who have raised questions and have faced the full force of Britain's ludicrous libel laws as a result.)

Could publications such as this one do more to deter cheats? Unfortunately, we are often in no position to flag up even proven cases of misconduct, and thereby highlight the risks that miscreants run with their careers. Yes, it is a journal's primary job to clean up the literature, but when papers are retracted owing to misconduct, the libel laws (again) often prevent our editors from saying so. We know that this leaves the affected communities frustrated and in the dark. It leaves us frustrated, too.

So, with journals unable to push towards greater integrity and universities often unwilling to do so, should funding agencies be leading the charge? It is, after all, their money that is wasted if misconduct does occur.

Funding agencies in the United States do sometimes investigate misconduct. Research funded by the National Institutes of Health and some other government agencies falls under the remit of the Office for Research Integrity (ORI), which has the power to bar researchers from receiving future funding. However, as Nicholas Steneck, director of the research-ethics programme at the Michigan Institute for Clinical and Health Research in Ann Arbor, told the London meeting, this process probably misses most major misconduct. And the ORI can't initiate investigations: institutions must conduct their own inquiries first.

"Stronger action and punishments are needed to discourage misbehaviour."

In the United Kingdom, there seems to be little appetite for launching an overarching ORI-type regulator. Certainly, the existing independent advisory group, the UK Research Integrity Office in Falmer, is clear that it has no desire to take on such a role. British funding councils — in collaboration with the country's universities — have chosen instead to produce a 'concordat' detailing good practice, to which institutions will be expected to sign up. This is laudable, but unlikely to strike fear into fraudsters and fabricators.

So, how can Britain highlight cases of misconduct and discourage it in future? Ultimately, the incentives probably need to come from on high, and the government could get the ball rolling by commissioning an anonymous survey on misconduct that UK researchers have witnessed and perpetrated. An official audit would offer a strong platform for others to build on — perhaps with a parliamentary inquiry and subsequent report on the damage done to UK science by misconduct, and an assessment of the options for tackling it and the investment needed. Funders and universities could then work together to establish common definitions of what counts as misconduct, and how it will be punished. And if a reform of the libel laws goes ahead, journals and other scientists would be able to do more to highlight and expose miscreants.

Sounds ambitious? If the solutions were easy, there wouldn't be a problem to discuss. But there is, so we must face it. ■

Cap in hand

A word to the wise on getting that much-needed research funding.

Would you be willing to spend weekends on the yacht of a friendly billionaire in the name of science? Or insist to airport check-in staff that your life-saving research demands that you be upgraded from economy to business class? Perhaps you would be happy to see your face on a T-shirt? Or for folks you met on the Internet to traipse through your lab, taking photographs?

Welcome to the cold reality of science in a global recession. As the flow of public money slows across the world, academic researchers are increasingly turning to private funders and wealthy individuals-turned-philanthropists to pay for their work. There is a strong tradition of such support already, of course, especially in the biomedical field. But as government cuts around the world begin to bite, more and more scientists will be looking for alternative sources of income.

So, in a short series of articles this week, *Nature* focuses on where that money is and how you can access it. Do not feel too proud to ask. Some of your competitors are doing it already.

In our News Features, we look at the two extremes of research paid for by private individuals — from the billionaires willing to set up entire laboratories and pay for the work done there for years, to the web-based begging bowls that can take just a few dollars each from thousands of different people.

On page 254, we talk to those at the top: the big spenders, the entrepreneurs — and those scientists who have benefited from their largesse. How did they do it? Partly by being in the right place at the right time, although it helps to know where the right places are. And it helps even more to have something to say when you get there. As Thomas Pierson of the SETI Institute in Mountain View, California, says, "People give money to people." One secret seems to be to think big. "If I ask for $100,000 and they say 'yes' right away, then I didn't ask for enough," one university fund-raiser tells us. "It's a common mistake."

At the other end of the scale are the crowd-funding websites, explored on page 252, where scientists can post details of their projects and ask many individuals to collectively cough up for them. From a low-cost robot for tackling oil spills to a project to map water quality along the Mississippi River, some researchers are already adept at tapping the potential of the masses. And although the target donors may be different, one key strategy remains the same: tell a good story. Sell yourself and sell your science.

Still, be wary of selling yourself short, warns a Comment piece on page 260. Patrick Aebischer, president of the École Polytechnique Fédérale de Lausanne in Switzerland, complains that too many donations to university research from charities and foundations come with a catch — they don't pay for the associated costs, such as salaries and utility bills. "Institutions with many privately funded projects are effectively 'punished' for their success," he writes. "To meet the higher research-infrastructure costs, universities may drain resources from education, or diminish 'expensive' disciplines such as physics, chemistry or engineering, in which philanthropic support is scarce." The solution, he says, is for institutions to identify the full cost of research activities and pass it on. "Private bodies should not hijack university resources. They should contribute a fair share of the expense."

The money is certainly out there. Just look at the billions poured into football teams. And if the pages of the glossy magazines can be believed, the luxury-goods market remains strong. The money must be spent on something, so why not science? And although those weekends on luxury yachts may be a tough way to make it happen, someone has to do it. ■

NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

SEVEN DAYS *The news in brief*

EVENTS

Mars probe crash

Russia's Phobos-Grunt spacecraft, which failed to make it out of Earth orbit in its attempt to reach a moon of Mars last year, splashed down in the southern Pacific Ocean on 15 January. As *Nature* went to press, exact details of the probe's crash and location were unclear, but it is expected to have scattered fragments over thousands of square kilometres. On-board tanks containing unused, toxic fuel probably broke up well before the probe re-entered Earth's stratosphere. See go.nature.com/qdv3to for more.

POLICY

NOAA to move?

US President Barack Obama has proposed moving the National Oceanic and Atmospheric Administration (NOAA) from the Department of Commerce to the Department of the Interior, as part of a broader agency reorganization that is intended to save taxpayers about US$3 billion over 10 years. Obama made the proposal on 13 January; what the move would mean for NOAA is unclear. See go.nature.com/efbprv for more.

Stem-cell lawsuit

Two scientists seeking to block US government funding for research using human embryonic stem (ES) cells filed their arguments to a US appeals court on 12 January. James Sherley and Theresa Deisher, who work on adult stem cells, are appealing a July 2011 decision in which a federal judge ruled against their case. The appeal is set to be heard on 23 April. On 12 January, the US National Institutes of Health made four more human ES-cell lines eligible for use in federally funded research, bringing the approved total to 146. See go.nature.com/wtsnxn for more.

MEGHDAD MADADI, FARS NEWS AGENCY/AP

Iranian chemical engineer assassinated

A chemical engineer who worked at Iran's Natanz uranium-enrichment facility was killed on 11 January in Tehran — the latest victim of a string of assassinations and attacks apparently aimed at Iran's nuclear programme. Mostafa Ahmadi Roshan Behdast, 32, was assassinated by a magnetic bomb that had been attached to his car (pictured). He was identified in the Iranian press as a deputy director of marketing at the Natanz facility. See page 249 for more.

US reactor safety

The United States may follow France in recommending that nuclear power plants install back-up equipment for containing a serious accident, in the wake of the meltdowns at Japan's Fukushima plant. On 11 January, the Nuclear Regulatory Commission (NRC) said that it considered a plan put forward by the Nuclear Energy Institute, a body based in Washington DC that represents the nuclear industry, to be "a reasonable starting point". The plan involves placing portable pumps, generators and other equipment around reactors for use in emergencies. The NRC is expected to announce new safety measures before the 11 March anniversary of the Fukushima accident. See go.nature.com/saeouu for more.

Emissions data

The US Environmental Protection Agency has released a public database of greenhouse-gas emissions from the nation's largest power plants and industrial facilities. The agency requires more than 6,200 facilities to report annual emissions, and their 2010 emissions cover around half of the nation's direct emissions. See page 247 for more.

FUNDING

San Raffaele saved

The debt-ridden San Raffaele Scientific Institute in Milan, one of Italy's most prestigious biomedical research institutes, will not be taken over by the Vatican after all. The institute filed for bankruptcy last October, after accumulating debts of €1.5 billion (US$1.9 billion), and a consortium led by the Vatican Bank put in a bid to rescue it (see *Nature* **478,** 296–297; 2011). But Milanese businessman Giuseppe Rotelli, head of the private-health group San Donato, which runs 18 hospitals in northern Italy, unexpectedly outbid the consortium on 10 January. Scientists had been concerned that Vatican influence could have distorted research agendas.

Low odds at the NIH

The success rate for scientists seeking grants from the US National Institutes of Health in Bethesda, Maryland, fell to a historic low of 18% in 2011, the agency announced on 13 January. The decline was driven by an 8% increase since

2010 in applications submitted for competitive awards, to almost 50,000. At 15%, women's success rates were equal to those of men for new grant applications, but when competing for the renewal of existing grants, women's 33% success rate trailed men's 36%.

Clean-energy spend

Global government spending on research and development in clean energy fell by 22% from 2010 to US$13.2 billion last year, according to figures released by analysts Bloomberg New Energy Finance on 12 January. Corporate research spending also slipped back, falling by 14% from 2010 to $13.2 billion. But total investment in the sector — which includes the financing of energy projects such as large solar installations and rooftop photovoltaics — surged by 5% from 2010 to reach $260 billion worldwide.

RESEARCH

Frogs in size war

Barely a month after a frog took the title of world's tiniest tetrapod, an even smaller frog has emerged to steal the crown. With an average length of just 7.7 millimetres from snout to vent, *Paedophryne amauensis* (**pictured**) is 1 mm shorter than the previous record-holder, *P. dekot*. Both frogs live in leaf litter in the forests of

C. AUSTIN/LOUISIANA STATE UNIV.

Papua New Guinea and make up two of the six minuscule *Paedophryne* species so far identified on the island. The latest find was reported on 11 January (E. N. Rittmeyer *et al. PLoS ONE* http://doi.org/hmw; 2012).

Transgenic flight

The German chemical giant BASF is shifting its transgenic-plant operations from Europe to the Americas, it says, because of widespread opposition to the technology from consumers, farmers and politicians at home. On 16 January, the company said that it would move its plant-science headquarters from Limburgerhof, Germany, to Raleigh, North Carolina, cutting 140 jobs in Europe, although it would retain research centres in Berlin and Ghent, Belgium. Jonathan Jones, a plant researcher at the Sainsbury Laboratory in Norwich, UK, said the move was "a sign that Europe is not open for business in this area". See go.nature.com/nramri for more.

Nuclear security

Australia is the most secure of the 32 countries that hold at least one kilogram of weapons-useable nuclear material, according to a first-of-its-kind study that asked experts to assess risks of nuclear theft. North Korea is listed last, and the United Kingdom ranks highest among nuclear-armed states, in tenth place. The 11 January study (www.ntiindex.org) from the Nuclear Threat Initiative, a non-profit organization in Washington DC, also examined security conditions in 144 non-nuclear states that might be used as safe havens or transit points for stolen material.

PEOPLE

Astrophysicist dies

Steven Rawlings, an astrophysicist at the University of Oxford, UK, who helped to instigate and coordinate the proposed Square Kilometre Array (SKA) telescope project, died in mysterious circumstances on 11 January, aged 50. His body was found at the house of a friend and police are now investigating the death. Shocked scientists added that it was a great loss to the SKA project. Telescope construction would not start until 2016, but a pivotal decision about whether to locate it in South Africa or Australia is expected by March. See go.nature.com/c8bhhn for more.

Red-wine fraud

A three-year investigation into a biology laboratory at the University of Connecticut has found its chief guilty of falsifying and fabricating data in at least 23 papers and 3 grant applications. Dipak Das, director of the Cardiovascular Research Center at the University of Connecticut Health Center (UCHC) in Farmington, is a cardiologist whose studies included work on the red-wine chemical resveratrol. On 11 January, the UCHC issued a 60,000-page report detailing extensive misconduct; on the same day, Das was dismissed as co-editor-in-chief of *Antioxidants & Redox Signaling*, and two of his articles in the journal were retracted. See go.nature.com/faxbyj for more.

COMING UP

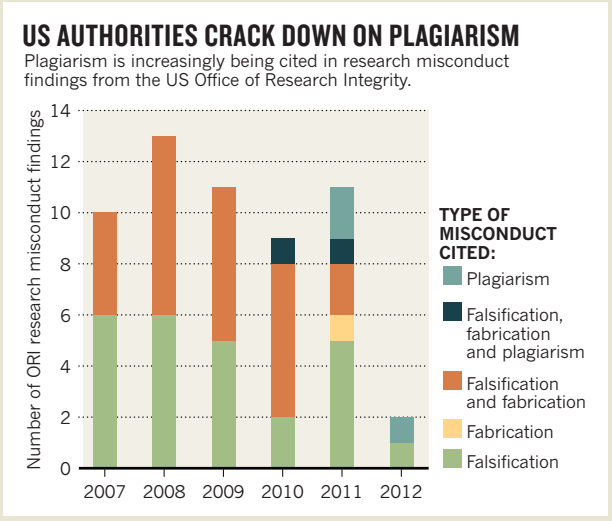
22–27 JAN
The sixth 'Arctic Frontiers' conference, in Tromsø, Norway, discusses the potential of energy resources in the Arctic.
www.arcticfrontiers.com

23 JAN–17 FEB
The World Radiocommunication Conference of the International Telecommunication Union in Geneva, Switzerland, will vote on whether to abolish the leap second, pulling our reference time out of sync with the Sun (see *Nature* **479,** 158; 2011).
go.nature.com/qq1zi7

TREND WATCH

The US Office of Research Integrity (ORI) is sharpening its eye for plagiarism. In the past, the office has censured researchers mostly for fabricating (making up) and falsifying (manipulating or omitting) data, but two of its three latest findings involved plagiarism. John Dahlberg, head of investigative oversight at the ORI, says that such cases are surfacing because of increased use of plagiarism-detection software, so the office is toughening its stance. See go.nature.com/qvp8qg for more.

SOURCE: ORI

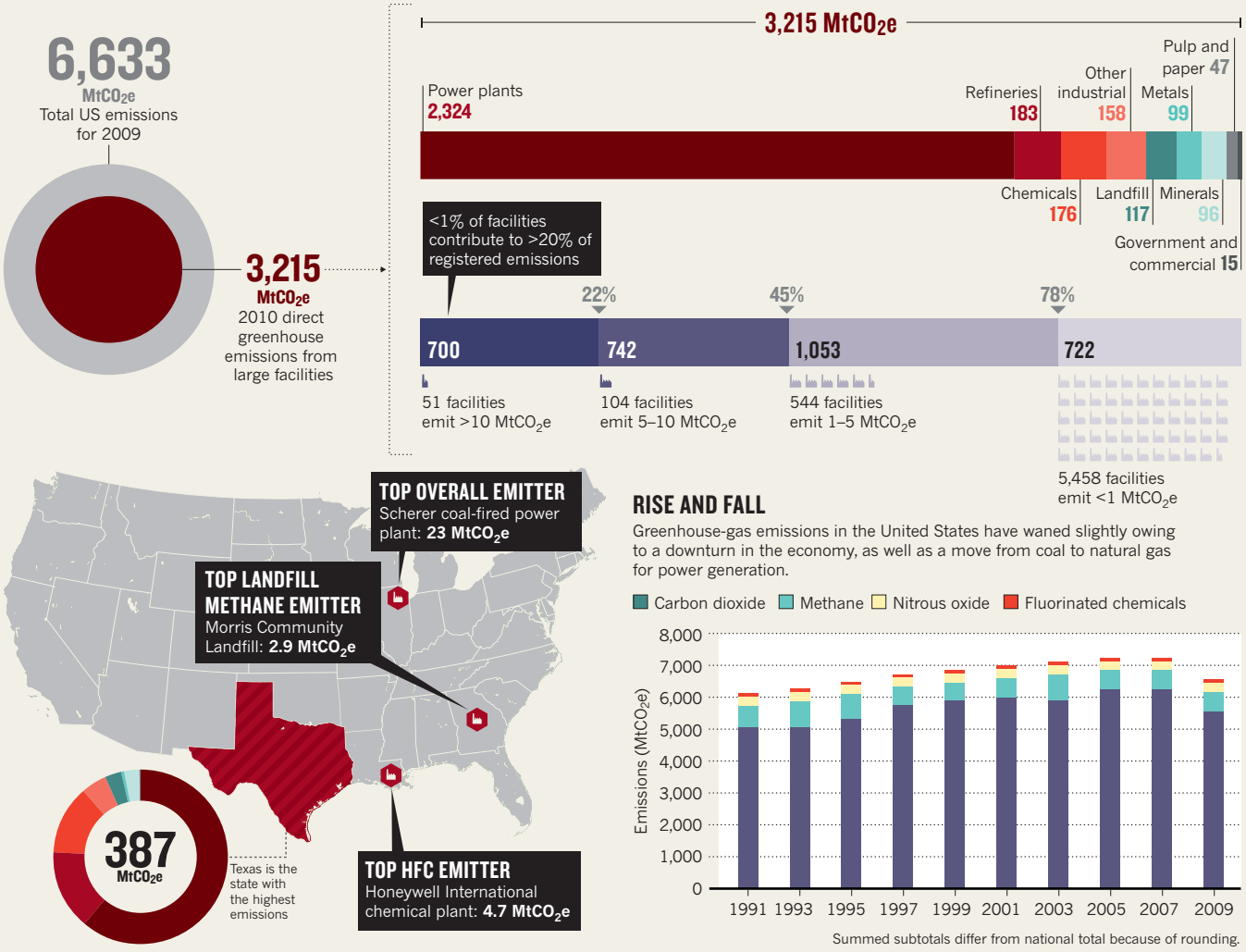


NATURE.COM
For daily news updates see:
www.nature.com/news

SOURCE: EPA

THE GAS TRACKER

The US Environmental Protection Agency's online greenhouse-gas database covers about half of the country's 2010 direct emissions. About three-quarters of those are produced by power plants; around 45% come from 155 facilities (just 2.5% of those that submitted data), each of which emits at least 5 million tonnes of carbon dioxide equivalent ($MtCO_2e$) a year.



CLIMATE CHANGE

Database tallies US emissions

Environment agency launches searchable public log of major greenhouse-gas emitters.

BY JEFF TOLLEFSON

Keeping track of a country's greenhouse gases is an accounting problem of epic proportions. In the United States, scientists have relied on a mix of methods to build up their national emissions inventory, including monitoring the electricity output of a power plant and assessing the quality of the fuel that powers it.

Now they have a new resource: official data from the companies themselves, collated into a user-friendly online database that was launched on 11 January by the US Environmental Protection Agency (EPA).

"It's a great resource, and I'm sure people will find interesting things to do with it," says Gregg Marland, a geologist at Appalachian State University in Boone, North Carolina, who led the development of guidelines for the Intergovernmental Panel on Climate Change on how to construct and use national emissions inventories.

The inventory covers industrial, commercial and government facilities that emit more than 25,000 tonnes of carbon dioxide equivalent per year, and was prompted by a law passed in 2008 that was meant to support the growing efforts towards regulating greenhouse gases. Plans for that regulatory programme collapsed in 2010, but EPA officials continued to work on a database, which would still provide investors and consumers with statistics to help them pressure industry to cut emissions.

The inventory covers facilities that directly emit about half of the country's total emissions, but does not include emissions from agriculture or land use. The online database allows users to compare and rank about 6,200 facilities by state, sector, and type and volume of greenhouse gases, including the ubiquitous ▶

carbon dioxide and methane and the more esoteric fluorinated chemicals.

The data should help researchers trying to measure and track greenhouse-gas emissions, as well as those studying natural carbon cycles. As scientists pin down exactly where anthropogenic carbon emissions are coming from, Marland says, they can separate out carbon uptake and emissions by plants and soils. "The better we understand both the magnitude and distribution of human emissions, the better we understand what is happening in the biosphere."

THE USUAL CULPRITS

Some database searches produce very familiar results. Carbon dioxide makes up about 95% of the greenhouse-gas emissions logged (although if emissions from agriculture and other sectors in the full US inventory are included, that proportion would drop to around 83%). The state with the largest overall emissions is Texas, which logged 387 million tonnes of carbon dioxide equivalent in 2010. And power plants overshadow any other stationary sources of greenhouse gases, accounting for about three-quarters of emissions in the inventory (see 'The gas tracker').

But there are also some eye-opening statistics. Just 2.5% of the facilities that have submitted data to the EPA are responsible for 45% of the emissions, for example. And almost all of the United States' emissions of trifluoromethane (also known as HFC-23), a potent greenhouse gas, come from just two facilities — the Honeywell International plant in Baton Rouge, Louisiana, and the Dupont manufacturing plant in Louisville, Kentucky. Since 1990, US producers have voluntarily reduced by about 85% their emissions of trifluoromethane, which is a by-product of the manufacture of the refrigerant and chemical feedstock chlorodifluoromethane.

The flood of data does not mean that scientists can stop measuring greenhouse-gas emissions in the atmosphere. Pieter Tans and his team at the National Oceanic and Atmospheric Administration's Earth System Research Laboratory in Boulder, Colorado, measure greenhouse-gas plumes from major facilities through a network of tall monitoring towers, for example. The annual totals in the database will certainly help to improve their atmospheric models, he says, but monitoring how emissions vary over the course of hours and days is still vital.

Totalling those real-time measures should also provide a way to verify the companies' annual estimates. "The hypothesis is that these are correct emissions estimates," Tans says. "We can test that to see whether what is being reported is consistent with actual observations." ■

EDUCATION

Evolution advocate turns to climate

Education centre known for battling creationists aims to help science teachers convey understanding of global warming.

BY SUSAN YOUNG

Students walking out of classrooms when global warming is mentioned; teachers pressured to change lesson plans to avoid the subject or portray it as speculative rather than a matter of scientific consensus. For Eugenie Scott, the stories and anecdotes fit a familiar pattern. Scott is executive director of the National Center for Science Education (NCSE) an organization based in Oakland, California, with a reputation for doggedly defending the teaching of evolution in US classrooms. But a growing impression that climate science is facing a similar struggle, together with entreaties from educators and textbook authors, has helped to convince her that the NCSE should expand its mandate to include the politically charged issue of global warming.

"I think we can make an important contribution," says Scott. "If teachers understand that there is a place that they can go to for help, we can use some of the expertise that we've gained over the years dealing with evolution to apply to this related problem."

A recent survey in the United States suggests that there is indeed a problem. From August to October 2011, the nation's National Earth Science Teachers Association (NESTA) queried Earth- and space-science teachers about their experiences of climate-change education. Depending on the region, 25–30% of respondents reported that students, parents, administrators or other community members had argued with them that climate change is not happening or that it is not the result of human activity. Some school boards and state legislators have threatened to require educators to 'teach the controversy' about climate change — a term coined in relation to evolution that amounts to presenting a scientific theory as one of various possible viewpoints.

Many of the teachers surveyed have strong science backgrounds and professional experience related to climate change. But "when you look at what the teachers are facing across the country, it goes way beyond science", says Roberta Johnson, executive director of NESTA, based in Boulder, Colorado. "It goes into areas of political debate."

In a 16 January announcement, the NCSE says that it will offer support to educators facing ideological opposition when teaching climate change, providing advice on how to present the underlying science. The strategy mirrors its approach to evolution, which includes clarifying for students why science is an appropriate tool for understanding the natural world. "This perspective is also important in helping people to understand the reasons why scientists overwhelmingly accept climate change," the NCSE says in a mission statement describing the new effort.

The statement also says that the NCSE will not take a position on what, if anything, should be done to counteract global warming or mitigate its effects. "What to do about it ranges widely and gets outside of the strict science and into policy issues in which many, many variables are going to have to be considered," says Scott. "We are not a policy think tank; we don't have expertise in this area."

But such policy neutrality may not prevent science teachers from being challenged in the classroom. "The core issue is not whether global warming is happening, or whether humans are involved, but whether it is a crisis," says James Taylor, a research fellow at the Heartland Institute, a libertarian think tank based in Chicago, Illinois, that opposes the regulation of carbon emissions (see *Nature* **475,** 440–441; 2011).

"When you look at what the teachers are facing across the country, it goes way beyond science."

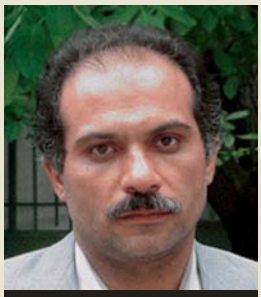
Scott acknowledges that there is more to teaching climate change than explaining the science clearly. "We need to be aware of the fact that people are very emotionally concerned about these issues," she says. If people feel threatened ideologically, politically or economically, "all the science in the world won't convince them". She adds that the NCSE will also help teachers to understand the views of parents and others who oppose the teaching of climate change.

"Knowing the motivations behind a parent's views can help a teacher come up with a solution or response that might assuage that parent's concerns and let their kid remain in the classroom," says Scott. ■

L TO R: FARS NEWS/REUTERS; B. MEHRI/AFP/GETTY; A. KENARE/AFP/GETTY; AP; FARS NEWS/AP

NUCLEAR FALLOUT

Several scientists have been targeted for assassination in Iran over the past two years, although some have no verifiable connection to the country's nuclear programme.



Masoud Alimohammadi
Particle physicist at the University of Tehran, no link to nuclear programme.
Killed 12 January 2010

Fereydoun Abbasi-Davani
Senior nuclear scientist at defence ministry, now head of nuclear programme.
Injured 29 November 2010

Majid Shahriari
Nuclear physicist, possibly involved in nuclear programme.
Killed 29 November 2010

Dariush Rezaei-Nejad
Electrical-engineering student, no link to nuclear programme.
Killed 23 July 2011

Mostafa Ahmadi Roshan
Behdast
Deputy director of marketing at Natanz uranium facility.
Killed 11 January 2012

NUCLEAR POLITICS

Murders unlikely to slow Iran's nuclear efforts

Experts say international sanctions are the best way to stall the weapons programme.

BY SHARON WEINBERGER IN BEIRUT

Last week's assassination of an official working at Iran's uranium-enrichment facility was worthy of a James Bond film: the 32-year-old was killed by a magnetic bomb placed on his car by a passing motorcyclist.

But the murder of Mostafa Ahmadi Roshan Behdast is unlikely to result in a neat filmic denouement. His death is the latest in a string of assassinations and other attacks seemingly aimed at Iran's nuclear programme over the past few years (see 'Nuclear fallout').

Although experts agree that at least some of the killings are part of an organized foreign campaign to slow Iran's efforts to enrich uranium, they are sceptical that the strategy will work. "The immediate effect is very small," says Olli Heinonen, a senior fellow at Harvard University's Belfer Center for Science and International Affairs in Cambridge, Massachusetts. "If I have a project that is important for national security, I never count on one single person," he says, suggesting that international sanctions are a more effective way of slowing Iran's efforts.

Heinonen, former deputy director-general of the International Atomic Energy Agency in Vienna, points out that there is still no proof of who ordered the assassination, or why. Ahmadi Roshan was named in many Western press reports as a nuclear scientist (his degree was actually in chemical engineering), but the Iranian press identified him as deputy director of marketing at the uranium-enrichment facility in Natanz, suggesting that he was not a crucial figure in Iran's nuclear programme.

Peter Zimmerman, a physicist and emeritus professor at King's College London, says that killing key personnel could delay a nuclear programme in the right circumstances. For example, killing Leslie Groves, the military chief of the Manhattan Project during the Second World War, might have had serious consequences if it had happened in 1942 or 1943 when the atomic-bomb project was in doubt and Groves' advocacy was crucial. But "if I had killed [J. Robert] Oppenheimer in 1944, I don't think it would have delayed Hiroshima by even a month, and maybe not a week", suggests Zimmerman. In Iran, he says, none of the victims so far is crucial to the nuclear programme.

Even identifying key personnel in the programme can be difficult, because scientists suspected of involvement often have academic affiliations. In an interview in 2007 with this reporter, Fereydoun Abbasi-Davani introduced himself as a university professor and head of the Nuclear Society of Iran, an academic association for promoting nuclear knowledge. But he was also on a United Nations sanctions list for his alleged involvement in Iran's clandestine military programme. In 2010, an attempt to assassinate Abbasi-Davani failed, and he was subsequently appointed as the head of Iran's Atomic Energy Organization in Tehran.

At the time, Abbasi-Davani described efforts to slow Iran's nuclear programme as "fruitless", and resulting in only short-term setbacks. For example, he said, the United States allegedly led a ban on selling software to Iran that could be used for calculating nuclear reactions. Abbasi-Davani said that Iranian scientists eventually developed their own software, in part using information from contacts abroad. "We were slowed down, but we could get the knowledge and know-how that we needed in other ways," he said.

A similar recovery was seen after the Stuxnet computer worm wreaked havoc with the uranium-enrichment centrifuges at Natanz in 2010, says Scott Kemp, an associate research scholar at Princeton University's Woodrow Wilson School for Public and International Affairs in New Jersey. "They overcompensated, they began enriching even faster, and within six weeks, they were caught up," he says. Soon, he adds, the facility had moved even closer to being able to produce enough highly enriched uranium for a nuclear weapon.

Kemp suggests that killing scientists who are allegedly associated with the programme could have the same effect, by bolstering Iran's resolve to develop its nuclear capability. "These things have unintended and unforeseeable consequences," he notes. ■

PHARMACEUTICALS

Russian drug law hinders clinical trials

Legislation to increase availability of new medicines has delayed approvals.

BY ALLA KATSNELSON

A law aimed at improving Russians' access to innovative medicines has backfired badly.

'On the circulation of medicines', a law that came into effect in September 2010, was intended to streamline the bureaucratic system for testing and registering new drugs in Russia. There were high hopes that it would also boost the nascent domestic pharmaceutical industry and attract foreign companies to run trials there. The law set a maximum cost and time limit for drug registrations, for example; mandated compensation for those injured during trials; and declared that drugs manufactured in Russia solely for export did not need to undergo registration.

But the law has actually hampered clinical trials and limited the number of drug approvals over the past year, according to a report from the Association of Clinical Trials Organizations (ACTO) in Moscow, which represents companies conducting clinical trials in Russia.

One problem is that medicines must now undergo local trials before they can be sold in Russia, regardless of whether they have already undergone trials or been approved elsewhere. The goal is to ensure that drugs are backed by high-quality trials that prove their effectiveness in the indigenous population. A few Asian countries have similar rules, but whereas some Asians are thought to metabolize certain drugs in a different way from Europeans, that is not the case for Russians, says Leonid Kokovin, a clinical-trials manager based in Moscow. "We are statistically like European and American subjects, so the argument for testing on our population isn't understandable," he says.

The law also raises ethical concerns because it subjects people to clinical trials that are not expected to yield extra knowledge, Kokovin notes. He suggests that medicines backed by suitable foreign trials should be approved by the Russian authorities, as long as they are tracked in post-marketing studies. "That would be much simpler, and wouldn't put such a barrier on the market," he says.

ACTO's report describes one (anonymized) European company's struggle to register a drug for preventing preterm delivery, a condition that affects about 5% of pregnancies. The medicine is already approved for this use in many Western countries, and for other uses in Russia. The cost of running a Russian trial was estimated to be €1 million (US$1.3 million), and patients in the control group would have to be asked to avoid the drug, raising their risk of preterm delivery. Furthermore, clinical trials on pregnant women are rare in Russia, so there are few accredited facilities that can screen participants.

Despite the difficulties, the company is going ahead with the trial. However, many others have been discouraged by a lack of detail about how to comply with the law. "No one can say — including, probably the regulatory agency — what kinds of trial specifically should be run," says Dmitry Margolin, co-chairman of ACTO's regulatory committee. "No one can give the guarantee that if you do these trials then the data will allow you to register the medicine." Margolin says he is aware of "several tens" of instances in which manufacturers had planned to seek registration fora medicine, but had been deterred by the requirement for fresh trials.

Russian pharmacies can't sell some internationally approved drugs, owing to a strict law.

S. KRASILNIKOV/ITAR-TASS

According to information from ACTO and the Ministry of Health and Social Development, which oversees the registration of medicines, just three repeat trials are under way for innovative medicines already approved outside Russia — two for hepatitis C and one for neuralgia — with two more set to start soon. "If the situation continues," says Margolin, "then a large number of new products will not be registered and will not reach Russian pharmacies."

On the basis of meeting summaries posted on the ministry website, ACTO also estimates that the ministry's Ethics Council rejects more than 30% of applications to run trials, mostly as a result of administrative or clerical errors in the applications. To avoid undue influence, Russia's new medicines law forbids the Ethics Council from directly contacting companies, so it cannot request that small errors are corrected. By contrast, in Germany, where the rejection rate is 1%, approvals are routinely conditional on small changes in the application. "It's one of the main flaws of the Russian regulatory system under the new law," says Sergei Lomakin, a senior associate at the Moscow office of the international law firm Baker & McKenzie. "The system leads to a bigger number of rejections than one that allows scientific dialogue" between companies and regulators, he says.

Another provision of the law — that medicines developed abroad cannot have their first clinical tests in Russia — is hampering foreign companies that are trying to develop products specifically for the Russian market. Ironically, many such companies have agreements with government-backed investors, such as the nanotechnology venture fund Rusnano, to develop new medicines in the country.

"These changes in the guidelines happened at the moment when the country had invested so much in innovative drug development that there were quite a few drugs which were around the corner from starting phase I trials," says Andrei Gudkov, chief scientific officer of Cleveland BioLabs in Buffalo, New York, which partners with Rusnano and is struggling to launch a Russian early-stage trial for one of its anticancer compounds. The Ministry of Health and Social Development did not to respond to *Nature*'s questions on the matter.

ACTO and others say that scrapping the requirement for local registration trials would be an important first step towards fixing the situation. Until that can be done, says Margolin, "companies for the most part are sitting and waiting". ■

J. POLLARD/GEMINI OBSERVATORY

Gemini North fires a laser at the sky (shown in time-lapse) to correct for atmospheric distortions.

ASTRONOMY

Gemini's twin telescopes reboot

Adaptive optics put the observatory at the cutting edge.

BY ERIC HAND

The Gemini Observatory's star has been slow to rise. The publicly funded observatory boasts two of the world's largest telescopes, and offers a rare opportunity to watch both the northern and southern skies from sites in Hawaii and in Chile. Yet it has lagged in terms of scientific productivity since it opened in 2000, and millions of dollars have been spent on overambitious, highly specialized instruments that were ultimately cancelled. "It felt like a child who had gone astray," says Frederic Chaffee, who witnessed Gemini's struggles while director of the rival Keck Observatory on Hawaii's Mauna Kea. "It was frustrating to watch."

But less than a year after Chaffee was drafted out of retirement to become Gemini's interim director when his predecessor stepped down, the once faltering facility may finally be living up to its promise. Last week, at a meeting of the American Astronomical Society in Austin, Texas, Chaffee showed off results from several new instruments. These should benefit a range of observers at the twin 8.1-metre telescopes.

NATURE.COM
For a slideshow of amazing Gemini images, see:
go.nature.com/h3kfvm

"In the last year, Gemini has made huge progress in turning around the observatory," says David Silva, director of the US National Optical Astronomy Observatory in Tucson, Arizona.

The turnaround is good news for astronomers of Gemini's seven member nations: the United States, the United Kingdom, Canada, Chile, Australia, Brazil and Argentina. Many are already planning to exploit the new Gemini multi-conjugate adaptive optics system (GeMS), an instrument that saw first light at Gemini South in December.

Adaptive optics are used at large ground-based telescopes to correct for image distortions caused by the motion of Earth's atmosphere. They monitor a bright guide star or backscattering from a laser beam fired by the observatory to measure the disturbances, then rapidly alter the shape of a mirror to correct them. But they sharpen only a small portion of the telescope's field of view. Astronomers have known for more than a decade how to make corrections over a larger area — by performing three-dimensional tomography of the sky with multiple lasers (R. Ragazzoni *et al. Nature* **403,** 54–56; 2000). But achieving this feat has been another matter.

With its 50-watt laser split into five beams, Gemini has managed it first. The area over which it has achieved sharpness is more than 10 times that of other systems. "This is a big step," says Olivier Guyon, an adaptive optics expert at Japan's 8.2-metre Subaru Telescope in Hawaii. "It's going to be followed by everyone." Guyon adds that the system will help astronomers to study the skies more efficiently, by bringing large objects, such as nearby galaxies, into focus all at once, and by allowing more distant objects to be studied in a single snapshot.

Gemini South also unveiled a long-awaited spectrograph in December. This should allow users to calculate distances to some of the Universe's most remote objects, such as quasars, the luminous cores of distant galaxies containing giant black holes. Yet another instrument, an imager that could spot planets orbiting distant stars, is due by the end of the year. Gemini's notoriously clunky software interface for scheduling observations has also been overhauled.

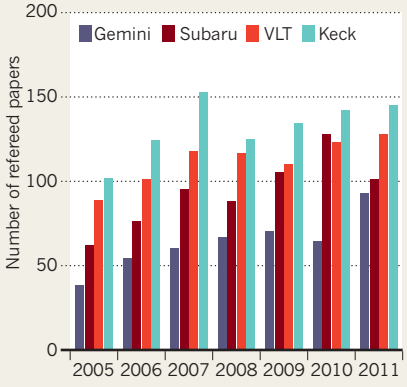
Finding ways to connect with — or at least not alienate — Gemini's far-flung users is important, says Chaffee, because astronomers rarely fly out to the telescopes to conduct their observations. He is establishing a users committee that he hopes will represent observers' concerns, and a science and technology committee, which met last November, that will aim to help Gemini avoid repeating the costly and impractical instrument choices made in the past.

But Chaffee believes that instruments such as GeMS are already starting to remedy years of poor scientific performance. Virginia Trimble, an astronomer at the University of California, Irvine, says that, on a per-telescope basis, both publication and citation counts for Gemini have fallen behind those of its rivals (see 'Paper chase'), and the demand for telescope time has fallen. However, Gemini deputy director Nancy Levenson says that some of this decline in demand reflects waning interest by astronomers in the United Kingdom, which will end its membership of Gemini at the end of 2013.

Gemini's international agreement will be revisited in 2015. By then, Chaffee will have long since returned to retirement. His one-year contract is up in May, and the Association of Universities for Research in Astronomy, which manages Gemini, is due to name a new director within weeks. "I'm trying to leave the place in as good a shape as I possibly can," says Chaffee. ■

PAPER CHASE

On a per-telescope basis, the scientific productivity of the Gemini Observatory has lagged behind that of other large observatories of equivalent size.



SOURCE: D. CRABTREE

Like it? Pay for it

With conventional sources of money drying up, some scientists are turning to crowd-funding.

BY JIM GILES

In October 2010, Cesar Harada found himself in New Orleans with little money and a big idea. Harada, an engineer, had been working on oil-spill mitigation at the Massachusetts Institute of Technology in Cambridge. But he quit the lab in frustration at what he saw as a slow pace of work and a focus on expensive solutions. He travelled south to join the clean-up operation for the Deepwater Horizon oil spill in the Gulf of Mexico. Once there, his mind turned to a futuristic solution: a low-cost clean-up robot that local people could build and deploy themselves. Yet his two criteria for the project — a quick build and open-source intellectual property — all but ruled out academic or industrial funding.

Harada turned to Kickstarter, a website used by authors, film-makers and artists in search of project funding. He uploaded a pitch, set a goal of raising US$27,500 and listed a series of small rewards for donors. Then he started to network furiously. Money came in from friends and engineering colleagues. A few companies heard about his idea; they pitched in several thousand dollars each. Word reached people he had never met, and they contributed too. When Harada's funding appeal closed in April 2011, he had raised almost $34,000 — enough to assemble a team of engineers and build a prototype of the clean-up robot.

PUBLIC INTEREST

If Harada's experience sounds like a one-off, think again. Crowd-funding — raising money for research directly from the public — looks set to become increasingly common. Established platforms such as Kickstarter are wooing scientists. And similar websites dedicated to connecting scientists with potential funders are being built, or have already launched. The public seems to be responding. Last year, for example, a group of scientists wanting to map water quality along the Mississippi River raised $64,000 in a trial project on an online crowd-funding platform called the Open Source Science Project (OSSP).

At a time when universities and research funding agencies are facing budget cuts, the strategy is attracting attention — as are other ways to raise philanthropic support (see page 254). "It's timely because of what's happening with traditional funding sources," says Daniel Gutierrez, co-founder of FundaGeek, a crowd-funding platform for technology projects that launched last month and is based in Yucca Valley, California.

For crowd-funding to make a real difference, advocates will have to prove that the process — which sometimes sidesteps conventional peer review — channels money to good projects, not just marketable ones. But if they succeed, there may be an unexpected bonus: it might help to forge a direct connection between researchers and lay people, boosting public engagement with science. "This is one of the most appealing aspects of crowd-funding," says Jennifer Calkins, an ecologist at Evergreen State College in Olympia, Washington, who has raised money for fieldwork on Kickstarter. "We can

2&3 ILLUSTRATION

involve society in the creative journey that we make as scientists."

Online crowd-funding has already proved its worth in other fields. Kiva, a website through which individuals loan small amounts to entrepreneurs in the developing world, is one notable success: more than 600,000 lenders have channelled almost $275 million through the site since 2005. US President Barack Obama's 2008 election campaign raised a record-breaking $780 million, much of it from small online donations. And donors have pledged more than $100 million to 13,000 Kickstarter projects. By drastically simplifying the process of connecting donor with cause, the Internet has unleashed a new enthusiasm for giving.

Scientists have come a little late to the crowd-funding party, because they have conventionally had other funding streams. Jai Ranganathan, an ecologist at the University of California, Santa Barbara (UCSB), is one of several researchers trying to make up for lost time. Last November, he helped to launch the #SciFund Challenge, an exercise in which close to 50 research groups had six weeks to raise money through proposals on a crowd-funding platform called RocketHub, which mostly serves artists and entrepreneurs. The challenge raised a total of $76,000. Brian Meece, RocketHub's chief executive, based in New York, says that research projects are a "new and exciting" use for his platform, and that he will retain the science section now that the challenge is over.

CASH FOR QUESTIONS

Other crowd-funding enthusiasts are developing donor sites dedicated exclusively to research projects. Sixteen projects are currently vying for funds on SciFlies, a site launched last November by David Fries, a marine engineer at the University of South Florida in St Petersburg. This year, the OSSP hopes to follow up on its success with the Mississippi study by launching a fund-raising appeal for around ten research proposals, says Priyan Weerappuli, a neuroscientist at the University of Michigan in Ann Arbor and the founder of the project.

Each site operates in a slightly different way, but there are common themes. Researchers start by describing and pricing a project, which they submit to the site for approval. If accepted, the pitch is placed online and donors have a few weeks or months to read the proposal and make a donation. Some sites operate on a non-profit basis and channel all proceeds to researchers; others are commercial concerns and take a cut of the money raised.

But although cash-starved scientists are lining up to list their projects, some are also expressing concerns. Take the issue of peer review. SciFlies and the OSSP post projects only after passing them through an expert review process, but Kickstarter's only requirement is that projects have "a creative purpose" — as defined by the site's owners. Projects in the #SciFund Challenge did not undergo formal peer review: Ranganathan and co-founder Jarrett Byrnes, a fellow UCSB ecologist, checked only for obvious fraud. "I don't care if people have badly thought-out projects," says Ranganathan.

That may sound like a recipe for shoddy science, but crowd-funding advocates say that the process has an inbuilt peer-review system, driven by the donors. Most donors will hear of a project through their social networks. They might be former colleagues of the project owner, or members of the public interested in an ecological study site. So project owners put their reputation among their peers and supporters on the line every time they post a proposal.

"There's a strong incentive to be honest," says Kickstarter co-founder Yancey Strickler. "Social forces carry a lot of weight." The system also puts a premium on inventive, well-thought-out proposals. A poorly conceived pitch that attracts no funds will do nothing for a scientist's career; nor will one that never delivers on its promises. "It may not be formal peer review, but crowd-funding has validation based on common trust," says Meece. "It's a pretty heavy filter." Even Sally Rockey, deputy director for extramural research at the US National Institutes of Health in Bethesda, Maryland, sees benefits in an alternative evaluation system if it helps organizations to achieve their research goals. Peer review "is not the only model", she says.

Some sites are trying to enhance this informal review process. FundaGeek has a discussion forum, the 'Geek Lounge', where potential donors are encouraged to debate the merits of a proposal. Last August, the equivalent forum on Kickstarter helped to halt one questionable project. The proposal, for a product called the Tech-Sync Power System, aimed to develop a smartphone app that controls home lighting. It attracted more than $27,000 in pledges, but Kickstarter users with electronics knowledge started to question the viability of the system. The project owner, who could not be reached for comment, eventually deleted his proposal as the criticism mounted, and none of the donors lost their money.

THE HARD SELL

Another objection to crowd-funding may be harder to shake. To sell a project, researchers need an attention-grabbing story (see 'How to woo the crowd'). That is easy to construct if your subject of study is, say, saving pandas or curing cancer. It is less so for researchers working on polymers. So will crowd-funding prove profitable only for 'sexy' science?

Fries concedes that crowd-funding inherently favours certain types of project, particularly those in applied research. He is an optimist, saying that if the approach takes off, conventional funding agencies will simply have to compensate by upping their support for basic science.

Ranganathan, an enthusiastic communicator who runs his own podcast, bristles at the suggestion that crowd-funding will create a two-tier system. "It's all about telling a compelling story about the research," he says. "Panda researchers start ahead, but I 100% believe anyone can do it." A polymer chemist might, for example, focus on new materials that could come out of his or her work.

The pressure to communicate the potential fruits of a research project should not be seen as a burden, adds Ranganathan. Most crowd-funding sites expect project leaders to offer donors something in exchange for their contribution, such as regular updates on the progress of the research. For those who make larger donations there might be visits to a lab or field site. In the case of the Mississippi water-quality study, donors in the region were encouraged to help with collecting water samples from the river. This process should help to forge stronger bonds between researchers and the public.

Whether all this works in wider practice remains to be seen, but many welcome the experiment. "Science thrives on diversity," says Jack Stilgoe, who studies science and society at the University of Exeter, UK. "We shouldn't be afraid of innovations in how it is funded. We should be more afraid when research money is all getting spent in the same way on the same sorts of things." ■ SEE EDITORIAL P.238 AND COMMENT P.260

Jim Giles *is a freelance writer based in San Francisco, California.*

MASS APPEAL

How to woo the crowd

The owners of crowd-funding sites give their tips on pitching winning scientific proposals.

- Create a compelling story about your research. Who will it benefit? And how? Then tell that story to camera — many sites allow project owners to upload short videos as part of the pitch.
- Devise clever rewards for donors. Think about giving away T-shirts decorated with project logos or, for big donors, a chance to visit your lab. Most sites require project owners to offer some reward, but bear in mind the time and expense required to produce and distribute whatever you offer.
- Use your social network, online and offline. Tell friends about the project, and ask them to tell their friends. Tweet it, blog it, publicize it on Facebook.
- Study previous successful pitches. Talk to the researchers behind them. Learn what works and incorporate it into your pitch. **J.G.**

NATURE.COM
Tell us your crowd-funding tips by commenting at:
go.nature.com/a3bi7o

Sponsor my science

Philanthropists will sometimes give large sums of money to support science — but researchers have to learn how to sell themselves first.

BY HEIDI LEDFORD

Asking someone to give you a million dollars is not easy. In 1995, Bruce Walker was seeking philanthropic support to expand his HIV research programme at Massachusetts General Hospital in Boston. He had identified a possible donor: a venture capitalist and the brother of one of his patients. He had teamed up with a personal coach in fund-raising and rehearsed his pitch over and over. But during lunch with the prospective donor, he still nearly choked. "I couldn't quite get it out to say, 'Would you give us a million dollars?'," Walker recounts.

Eventually he spat out his request — and was astonished when the venture capitalist agreed to the entire sum. "At that point, I fell off my chair," Walker says. He then raised another US$100 million from philanthropists Terry and Susan Ragon to launch the Ragon Institute of MGH, MIT and Harvard in Charlestown, Massachusetts, in 2009. He and his colleagues are now planning a formal seminar series on raising philanthropic support. His success at fund-raising, he says, "is not because there's something special about me. It's because I got a little bit of training and I put in a lot of effort."

Walker is not the only one with a determined approach to philanthropy. As public funding for research dwindles, scientists are increasingly seeking private benefactors (see page 252). But they have a lot to learn if they are to win trust and money: how to schmooze contacts, promote their science and deliver results to deadline — all without over-promising on the work (see 'How to woo philanthropists'). Much of this does not come naturally to scientists. "You have to sell yourself," says Cheryl McEwen, who, with her husband Rob, donated US$10 million to found the McEwen Centre for Regenerative Medicine in Toronto, in 2003. "But if you can build a case that we can understand, we'll be there for you."

➲ NATURE.COM
Read a Comment on the dangers of philanthropy:
go.nature.com/te4dz8

To help build that case, research institutions — particularly those in the United States — are becoming savvier about how to approach philanthropists. In April 2010, Steve Rum, the vice-president for development and alumni relations at Johns Hopkins School of Medicine in Baltimore, Maryland, started training faculty at the school in fund-raising techniques. Elsewhere, scientists are enrolling in classes provided by external institutes. Researchers are happy to have the help, and not just because it can bring in new income. It can also allow them to pursue projects that government funding committees would find too risky to support. "When you come to the government agencies with their committees and extreme accountability and taxpayer dollars, everything tends to come out a bit plain vanilla," says Tim Hunt, a Nobel-prizewinning biochemist at Cancer Research UK in London, who credits philanthropic support with helping him to continue his basic cell-biology studies even when applied research was more in vogue. "It's valuable to have funding that's a little bit out of the mainstream."

2&3 ILLUSTRATION

PERSONAL APPROACH

Philanthropy is supporting a growing slice of science: donations from US foundations to science, technology and medical research have grown from $793 million in 1999 to $1.7 billion in 2010, according to figures from the Foundation Center, an organization based in New York that analyses information about philanthropy. But although many scientists have experience in sending grant applications to foundations, few try to win multimillion-dollar donations from rich and generous individuals.

Convincing a philanthropist to fund a programme means cultivating a personal relationship with them. "It's an old adage, but it's

completely true: people give money to people," says Thomas Pierson, chief executive of the SETI Institute, an astrobiology research centre in Mountain View, California. After SETI was cut loose from NASA in 1993, it relied on Bernard Oliver, a well-connected member of its board and a former head of research and development at the computer firm Hewlett Packard in Palo Alto, California, to take its case to local entrepreneurs. The result: support from a who's who of technology luminaries, including Microsoft co-founder Paul Allen.

Most scientists will need to look outside their usual network of colleagues and supporters for prospective donors, says Robert Klein, a real-estate magnate who helped to raise $34.5 million in the 2004 campaign for California's $3-billion stem-cell initiative. Most of that, he says, came from wealthy families with a vested interest — usually a sick relative — in seeing stem-cell technologies succeed. To reach these families, Klein recommends first building contacts with patients and those who support them. "Have a patient advocate call ahead, and sit next to you while you make the pitch," says Klein. "Better yet, if you can get them to do it, have the advocate make the pitch for you."

Walker urges researchers to be on continual alert for new contacts. "When you travel, always try to get upgraded to business class," he says. "Then talk to the people around you about what you do."

This type of schmoozing takes time and a certain personality. Rum says that Morton Goldberg, an ophthalmologist at Johns Hopkins, is a fund-raising "superstar" — helping to build an endowment that now funds more than 30 professorships. The secret to his success, says Rum, is his willingness to forge lasting relationships with philanthropists. Goldberg says he considers many of his donors to be close friends. He has travelled with them, spending weekends on their boats and in their vacation homes. "It has been an extremely rewarding experience, far beyond any connection with money," he says.

Many physicians are reluctant to solicit their former patients, and with good reason, says bioethicist Sheldon Krimsky of Tufts University in Boston. "It's not OK for a doctor who is actively treating a patient to do it." Even after a patient has left the hospital, he or she may return, says Krimsky, and a physician who has received that patient's support may now be expected to provide special treatment.

Another pitfall of philanthropy, Krimsky says, is donors who attach strings to their gifts. When the Charles G. Koch Charitable Foundation in Arlington, Virginia, agreed to fund a faculty position in economics at Florida State University in Tallahassee, it demanded the right to determine the criteria used to pick a professor, and to veto candidates it did not like (S. Krimsky *Nature* **474,** 129; 2011). "It was so egregious," says Krimsky. The university accepted the terms in 2008, but has argued that input from the foundation during the hiring process has not compromised its academic integrity.

> *"Try to get upgraded to business class."*

Krimsky worries that tough economic times can make researchers and their institutions more willing to accept intrusions into academic freedom in exchange for funding. Scientists agree, but several who have received philanthropic support point out that they appreciated the sound financial and strategic advice the donors offered.

For many scientists, the most difficult step is pitching their work as if they were an entrepreneur in need of investment, says Garen Staglin, who, with his wife, Shari, donated some of the family's earnings from their vineyard in Napa Valley, California, to mental-health research after their son was diagnosed with schizophrenia. "You have to go into sell mode," he says. "You have to be able to say 'there are no guarantees here but if this works, this is the kind of stuff we think is possible in our lifetime'." But that type of sell tends to be easier for applied projects, such as potentially life-saving medical studies, than for basic research.

Selling research can also require a willingness to set deadlines. "In difficult economic times, the philanthropies become very goal oriented," Klein says. It's a practice that many scientists resist, because science rarely goes according to plan. When it doesn't, fund-raising specialists say that it's best to acknowledge the failure and stress the importance of the lessons learned. "You will still give them the confidence that there are incremental improvements," says Klein.

PHILANTHROPY SCHOOL

Research institutions often rely on professional fund-raisers to approach philanthropists, but Johns Hopkins isn't alone in training faculty members. Advancement Resources in Cedar Rapids, Iowa, has seen a fivefold increase in demand for its fund-raising training over the past four years, says chief executive Joe Golding. Institutions pay $17,500 for a four-hour workshop of lectures and role-playing exercises.

Golding says that basic researchers would do well to target entrepreneurs. "They are, by definition, risk takers," he says. "And they value leverage." Stress that a $30,000 investment now could lead to a million-dollar government grant later, he advises, and emphasize the impact a project would have on a topic that is close to the entrepreneur's heart. Golding, Rum and other experts also tell their trainees to listen for clues to how much a donor might be willing to commit. "If I ask for $100,000 and they say 'yes' right away, then I didn't ask for enough," says one fund-raiser at a large research university who asked not to be named. "It's a common mistake."

Rum even conducted a trial to find the best way to teach faculty members about how to approach former patients. He divided faculty members into three groups. One received only weekly fund-raising lessons by e-mail. The second also attended a lecture on the topic. The third received individualized coaching, including how to prepare an 'elevator pitch' — a brief description of research so compelling that it could capture someone's attention in the time it takes to travel between floors. Three months after coaching, the first two groups had brought in no money. The third had collected five gifts totalling $219,550 (S. Rum and S. M. Wright *Acad. Med.* **87,** 55–59; 2012).

Walker says that this type of personal coaching — not to mention the agony of asking for money — has been more than worthwhile. "I think we as a scientific community don't do a great job in articulating the transformational power of philanthropy," he says. "But that funding has allowed me to do things that I never could have considered if I'd been trying to do it through traditional sources."

Too few researchers take advantage of philanthropy, says Goldberg. "The most common mistake is not to try." ■ SEE EDITORIAL P.238 AND COMMENT P.260

Heidi Ledford *is a reporter for* Nature *based in Cambridge, Massachusetts.*

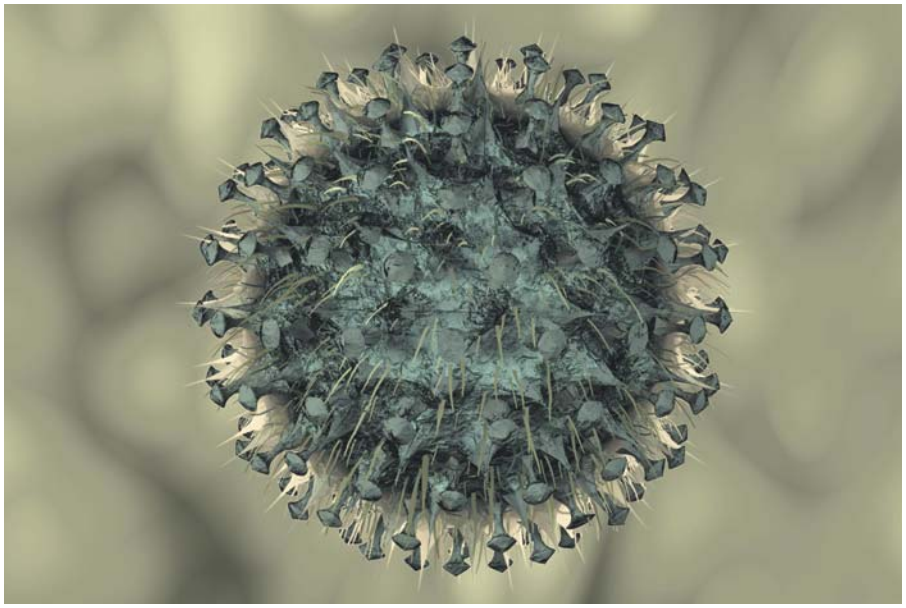
FUND-RAISING 101

How to woo philanthropists

Advice from successful fund-raisers on how to sell science.

- Network outside your usual circles: look for contacts — and contacts of contacts — who can strike an emotional chord with philanthropists.
- Learn how to frame your research and its importance in a few sentences. Practise this 'elevator pitch' in front of the mirror, and with anyone who will listen.
- Aim high: look for hints about how much a donor might be willing to commit.
- Establish deadlines: break projects into small chunks and assign completion dates for each step. Explain the vagaries of scientific research to your donor and don't be afraid to adjust timelines.
- Reward your investor: send regular progress updates, acknowledge their sponsorship in publications and presentations and encourage contributors to come to the lab for visits and tours. **H.L.**

COMMENT



OXFORD SCIENTIFIC/PHOTOLIBRARY/GETTY

The fight over flu

A proposal to restrict the planned publication of research on a potentially deadly avian influenza virus is causing a furore. Ten experts suggest ways to proceed.

RON FOUCHIER & AB OSTERHAUS
Globalize the discussion

Erasmus MC, Rotterdam, the Netherlands

So far, most of the human deaths from the deadly H5N1 strain of bird flu have occurred in Asia and the Middle East. Many labs worldwide — including ours — are trying to understand what makes the virus so virulent, and how to stop it. H5N1 research is thus a global issue, yet the entire research community seems to be following the advice of one country.

We are not questioning the unprecedented recommendations last month from the US National Science Advisory Board for Biosecurity (NSABB) to remove key details from the methods and results sections of published papers, including our own, submitted to *Science* (see *Nature* **481,** 9–10; 2012). But we do question whether it is appropriate to have one country dominate a discussion that has an impact on scientists and public-health officials worldwide. This discussion should include the perspective of people in regions where H5N1 has infected humans. Will the NSABB also advise on which international researchers and officials have the right to see the full papers, to help implement urgently needed surveillance and other intervention strategies?

It is not clear whether an international discussion would lead to different recommendations. There is no global equivalent of the NSABB, but many European experts that we have seen quoted in the press believe that the research should be published in full. We don't know the worldwide opinion until a group of experts from all parts of the globe is formed. An issue this big should not be decided by one country, but by all of us.

JOHN STEINBRUNER
A system for redacted papers

Director, Center for International and Security Studies at Maryland, University of Maryland, College Park

If the two papers (submitted to *Nature* and *Science*) describing a transmissible form of the H5N1 virus are the first to be published with key details missing, they probably won't be the last. We need to establish both a short-term and a long-term solution for how the scientific community should handle such a publication. Who decides who should have access to the full details? Who monitors the community so that the details don't get passed around outside the group of experts cleared to receive them?

I believe that the entire process must be regulated by a global health body, ideally the World Health Organization (WHO). Already, a WHO committee oversees all research involving the smallpox virus. A similar, more developed system could work for H5N1 and other deadly pathogens. An international group of experts would approve research involving those agents, decide who will have access to the details of papers that come out of that work, and hold those vetted individuals accountable for what they do with the information. For instance, such a system might allow permitted experts to view papers only electronically, so nothing is on paper. A database could record the privileged few people who have seen the full paper, and what they do with the information. In all likelihood, legal safeguards would be needed to protect the rights of people who receive the sensitive information. Such a process would not offer complete protection against misuse, but it would show the scientific community that the committee is watching what they do.

Because dangerous pathogens are a global issue, any procedure would need buy-in ▶

from all countries, who would have to give the committee binding jurisdiction over research involving extremely dangerous agents. This will take some time. In the short term, the WHO or some other global health organization should immediately establish an ad hoc committee to review who should receive access to the full H5N1 papers, and ensure that the details do not circulate widely.

Most importantly, these discussions should not be controlled by officials focused on national security. H5N1 is primarily a matter of public health. If there is a threat of bioterrorism, let it be judged by a global health organization, which can set rules that do not deprive scientists of information that could save millions of people in the case of a natural pandemic. If national-security organizations become involved, they will vet scientists on the basis of citizenship, and will be inclined to discriminate against those countries in which terrorists have found refuge. But some of those countries are among the few that have experienced human deaths from H5N1, and are most likely to witness the origins of a natural pandemic. It is crucial that scientists and other experts are judged on their qualifications, not on their nationality. If the world is to accept the process, national-security officials cannot be allowed to dominate the discussion.

N. SHRESTHA/EPA/CORBIS

South Asian countries such as Nepal are among the most likely to be hit by avian flu epidemics.

KWOK-YUNG YUEN
The Hong Kong perspective

Chair of Infectious Disease, Department of Microbiology, University of Hong Kong

As a scientist working in Hong Kong — the site of the first human epidemic of infection by the highly fatal H5N1 virus — I appreciate the public-health significance of knowing which mutations confer airborne transmissibility in an animal model. The new, much-debated research provides this information. Finding similar genomic signatures in animal or human viruses collected from the WHO Global Influenza Surveillance Network may alert public-health workers to an impending epidemic of unthinkable magnitude or severity. But I also appreciate the possibility that such mutants could cause a global disaster if accidentally or deliberately produced and released into animal and human populations. Consequently, I support the recommendation from the NSABB to remove key details from the papers describing this work.

Biological warfare is familiar to people living in this part of the world. During the Sino-Japanese War in the 1930s and 1940s, scientists and physicians of the Japanese army's Unit 731 infected civilians and prisoners of war in Manchuria, China, with *Yersinia pestis*, the bacterium behind the Black Death, which was first identified in plague patients in Hong Kong in 1894. When the unit found that the bacteria caused severe organ damage when serially inoculated into prisoners of war, the unit began spreading plague-contaminated fleas in China, causing outbreaks.

Censoring scientific data for publication will not stop rogue individuals or nations from developing a deadly and highly transmissible form of H5N1, but it would at least buy some time to find and stockpile the appropriate antivirals, immunomodulators and vaccines to protect against most variants of H5N1. Even if the publications omit the methods for making such a deadly virus, the genomic signatures associated with airborne transmissibility should be known to the directors of all public-health laboratories in the WHO surveillance network, after they sign an agreement of confidentiality.

D. A. HENDERSON
The ultimate biological threat

Center for Biosecurity, University of Pittsburgh Medical Center, Pennsylvania

The H5N1 influenza strain poses a potential biological hazard far more serious than any we have ever known. It is a virus that is capable of killing half its victims, a proportion greater than that for any other epidemic disease. Were that coupled with the transmissibility of a pandemic flu virus, it would have characteristics of an ultimate biological weapon unknown even in science fiction (see *Nature* **480,** 421–422; 2011). We should not publish a blueprint for constructing such an organism.

LYNN KLOTZ & ED SYLVESTER
Worry about lab infections

Center for Arms Control and Non-Proliferation, Washington DC; Walter Cronkite School of Journalism at Arizona State University

Asian bird flu is just one of the extremely dangerous pathogens researched in laboratories throughout the world. Along with the two labs that created a potentially contagious form of H5N1, at least 40 others worldwide investigate deadly, highly contagious pathogens not currently present in human populations, such as the SARS virus and the recently resurrected 1918 pandemic flu virus. Public-health experts worry about natural pandemics, and governments worry about the risk that these pathogens pose to national security — but the probability of accidental release is likely to be much higher.

We have analysed the likelihood of escape from 42 labs, using 1% as the estimated probability of an escape from a single lab in a single year. This approximates the historical probability, obtained by dividing the documented number of escapes of these pathogens (3, each involving the SARS virus) by our estimate of the total number of lab-years of research on these pathogens since 2003 (more than 300 lab years). Lab infections can easily spread: in 2004, after the only natural SARS outbreak in humans was contained, two graduate students lab-infected with SARS in Beijing infected seven others, causing one death.

Our analysis shows that the probability of an escape from at least one of 42 labs in a single year is 34%; within less than 4 years,

the chance of escape reaches 80%. If more labs begin studying the (potentially) human contagious form of H5N1, that interval will decrease even further.

In comparison, the average time between the three natural influenza pandemics in the last century was about 30 years (1918, 1957 and 1968). We are creating a risk that is much greater than that posed by nature. Laboratories need stronger precautions, such as mandatory quarantine of lab staff working on live dangerous pathogens until they can be certified clear of infection. Regulators should not be sitting idly by, while the threat of a man-made pandemic looms.

JEFFERY K. TAUBENBERGER
Study how viruses swap hosts

Chief, Viral Pathogenesis and Evolution Section, Laboratory of Infectious Diseases, US National Institute of Allergy and Infectious Diseases

How influenza viruses adapt to humans and cause disease has been a significant interest of my laboratory, including sequencing the genome of the 1918 H1N1 influenza, the virus that caused the worst pandemic on record. This work has led to research on how influenza viruses switch hosts and cause severe disease, and is serving as a basis for new vaccines and therapies. As a result of these studies, scientists and public-health officials are better prepared to prevent or mitigate a similar future pandemic. Studies examining adaptation and transmission of avian H5N1 influenza virus are similarly important.

For reasons not yet clear, some influenza viruses adapted to poultry can acquire sets of mutations that kill almost every bird that they infect, but rarely infect humans. Since 1997, H5N1 viruses have devastated poultry flocks in southeast Asia and other parts of the world, and also caused disease in a limited number of the presumably tens of thousands of humans who have had direct contact with infected birds. Although human infections are uncommon, the high fatality rate in those people infected by H5N1 — more than 50% — raises serious concerns.

A growing body of evidence suggests that influenza host-switching processes are complex, and may be unique to each virus. Changes in the 1918 H1N1 virus crucial for adapting to humans, for example, were not found in the 2009 pandemic H1N1 virus. It is essential to continue research into this with H5N1 and other emerging pathogens, and to investigate how they adapt and cause disease. The work could provide better surveillance tools to detect viruses that are in the process of adapting to humans. Scientists must be able to continue to work with these viruses safely and appropriately, in multiple teams utilizing the talents of many researchers, and under the numerous layers of protection and oversight that safeguard scientists and the public.

Of course, scientists will have to continuously re-examine and discuss the balance between the benefits of conducting and publishing pathogen research, and the risk of misuse of such knowledge. We should continue the dialogue as we continue the research.

➲ NATURE.COM
See *Nature*'s web focus on the 1918 influenza pandemic: go.nature.com/9rcpjv

RICHARD H. EBRIGHT
Mitigate the risks of release

Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey

Engineered derivatives of highly pathogenic avian influenza virus capable of aerosol transmission in non-rodent mammals pose potential threats to human health and to food production. The major concern is accidental release, for example through infection of a lab worker who then infects others. Deliberate release by a disturbed or disgruntled lab worker, bioterrorism and biowarfare are also concerns.

The following steps should be taken immediately. To address accidental release, assign the viruses as pathogens that require the highest biosafety level ('level 4', not 'level 3+' as has been the case to date). To address deliberate release, assign the viruses as pathogens that require the highest level of biosecurity ('Tier 1', like smallpox virus and anthrax bacterium, under the revised US select-agent rule that enters into force this year).

To minimize risks from future research, additional steps should be taken. We need to implement a system of mandatory prior review of research directed at increasing a pathogen's virulence, transmissibility or ability to evade countermeasures. We also need to re-evaluate, and preferably terminate, biodefence expenditures on research directed at creating and assessing new biothreats, as opposed to addressing existing biothreats. Creating and assessing new threats rarely increases security. Doing so in biology — where the number of potential threats is nearly infinite, and where the asymmetry between the ease of creating threats and the difficulty of addressing threats is nearly absolute — is especially counterproductive.

DAVID L. HEYMANN
We will always need vaccines

Head, Centre on Global Health Security, Chatham House, London

The creation *in vitro* of highly transmissible forms of H5N1 is a fresh reminder of how the endgame in infectious-disease control has changed.

For years, scientists and officials have believed that the ultimate application of disease control was eradication. In a world free from a particular pathogen, the disease no longer occurs, and the expense — and small risk — of vaccination is avoided. As a result, there would be no need to develop and maintain vaccine stockpiles. This idea began to change in 2002, when the poliovirus was synthesized *de novo*. The change has been profound — the endgame must now take into consideration the possibility that consolidation and safe storage, or destruction, of remaining polioviruses is not sufficient to eliminate the risk that it will be reintroduced. With the possibility of synthesizing a virus, it is likely that we will always need to have vaccines readily available.

But some say that maintaining samples of deadly pathogens is a necessary precaution. Take the discussions about the samples of smallpox virus (*Variola*), maintained at two WHO-sanctioned laboratories since the virus's eradication. Some now argue that we need these samples to study the virus, and to develop new vaccines and therapies in case of a bioterrorist attack. Indeed, the WHO Advisory Group of Independent Experts that reviews the smallpox research programme noted this year that DNA sequencing, cloning and gene synthesis could now allow *de novo* synthesis of the entire *Variola* virus genome and creation of a live virus, using publicly available sequence information, at a cost of about US$200,000 or less.

It is tempting to argue that, for the sake of public health, research that constructs highly pathogenic forms of viruses, such as H5N1, should be squashed. But even if we eliminated H5N1 and destroyed all lab samples, we would still need to consider keeping and producing a vaccine, because the virus could potentially be recreated. And stopping this type of research would open a dangerous door. Although adding and deleting genes can create super-strains that put the entire world at risk, such research also helps to develop public-health tools such as vaccines and diagnostic tests. Preventing this research would also prevent us from using all possible scientific options to prepare for naturally occurring — or deliberately caused — outbreaks. ■

2&3 ILLUSTRATION

The price of charity

Philanthropists should pay their fair share of research costs, says **Patrick Aebischer**.

With public funding for universities diminishing, private and philanthropic sources are increasingly being pursued to support academic research. Although welcome, this money comes with a catch: charities and foundations rarely pay the full costs of running a lab — building construction, maintenance, utility bills or salaries, for instance — and tend to be selective in the projects they sponsor. Universities therefore need to do more to pass on the true costs of research to donors. They must also have a balance of income sources to avoid bias in their research directions.

Private foundations have been significant sponsors of science since the nineteenth century. Until the US National Institutes of Health started funding extramural research in the mid-1940s, for example, more than one-quarter of US medical research was paid for by philanthropic organizations. As public funding grew, the share contributed by foundations and charities declined to less than 4% by 2007 (ref. 1).

Since then, support from non-profit organizations — especially those associated with rich individuals such as Bill Gates — has risen worldwide. In Europe, philanthropic sources now supply 6.5% of competitive research funding on average — 3–4% in most European countries and almost 10% in the United Kingdom[2]. At my own institution, the Swiss Federal Institute of Technology in Lausanne, private sponsorship has tripled, from 3% to 9% of research income over the past 10 years.

Biomedicine is benefiting most. And nowhere more so than in the United Kingdom: led by the Wellcome Trust, charities accounted for 40% of British biomedical research funding in 2009 (ref. 3).

LEGACY OF URGENCY

Private institutions, however, have a disproportionately large influence on the research landscape relative to their financial contribution. Because they can make funding decisions quickly, their money is welcomed by principal investigators looking to fill funding gaps. But directed awards can divert an entire group's research towards one end.

Keen to ensure that results are delivered and milestones met, charities' legacy of urgency has taken them away from backing long-term research projects and towards satisfying narrow goals that fit their mission. Since the boom in 'venture philanthropy' in the 1990s, the personal agendas of powerful entrepreneurs have accelerated research into AIDS; infectious diseases such as malaria and tuberculosis; cancer; and neurodegenerative diseases such as Parkinson disease.

Charities often send the message that not one cent invested is spent on anything other than finding the cures and so minimize their contributions towards a university's overhead costs. That means that institutions with many privately funded projects are effectively 'punished' for their success. To meet the higher research-infrastructure costs, universities may drain resources from education, or diminish 'expensive' disciplines such as physics, chemistry or engineering, in which philanthropic support is scarce.

Private bodies should not hijack university resources. They should contribute a fair share of the expense of a sustained research enterprise. To make it easier for them to do so, universities should better identify the full cost of research activities and pass it on. And because most charities operate internationally, these overheads should be harmonized worldwide.

FULL COSTS

Such an accounting effort is ongoing in some countries, especially the United States. But the concept of overheads is almost unknown in many European countries, where universities are largely state supported. Some European universities are addressing this issue, in response to requirements by the European granting system to supply full research costs.

Estimates of overhead costs vary by field and by country. In the United States, they typically range from 40% to 70% of grant income, depending on whether the principal investigator's salary is included. But awards are usually lower in Europe. For example, in 2008, the Swiss Parliament agreed to compensate universities with an overhead of approximately 10% of each successful project submitted to the Swiss National Science Foundation. Grants from the Seventh Framework Program and the European Research Council support a maximum overhead of 20%. This is inadequate and should rise.

Universities can do much to manage their varied funding streams. They should develop, champion and apply transparent full-cost accounting mechanisms. Leaders engaging in fund-raising should promote the inclusion of overhead costs. And academies must avoid propping up underfunded research activities with educational resources.

Ultimately, universities should not rely on short-term funding for basics such as infrastructure and faculty salaries. They need to retain a healthy level of public funding if they are to survive long term. Universities have a mission of education and of long-term, fundamental research; charities aim to solve real-world problems now. Fruitful partnership should imperil neither but achieve the noble goals of both. ■

Patrick Aebischer *is president of the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
e-mail: patrick.aebischer@epfl.ch

1. Dorsey, E. R. *et al. J. Am. Med. Assoc.* **303,** 137–143 (2010).
2. Estermann, T. & Pruvot, E. B. *Financially Sustainable Universities II: European Universities Diversifying Income Streams* (European University Association, 2011).
3. Martin, L. *All Together Now: Improving Cross-sector Collaboration in the UK Biomedical Industry* (NESTA, 2011).

D. LEFRANC/GAMMA-RAPHO/GETTY

Despite threats such as the 2001 US anthrax attacks, vaccine innovation has slowed in the past 40 years.

PUBLIC HEALTH

Biodefence built on teamwork

The golden era of US vaccine research holds beneficial lessons for today, finds **John Grabenstein**.

Biomedical researchers who are familiar with the US National Institutes of Health as a funding body and agenda-setter might be surprised by *Long Shot*. Science historian Kendall Hoyt shows that there was once a different way to drive medical progress. In the 1940s and 1950s, focused expertise and goal-oriented collaboration between the military, academia and industry yielded key vaccine innovations that bolstered national security, yet had benefits for civilians. That era of medical discovery stands in marked contrast to today's patent constraints, regulatory hurdles and disjointed efforts.

The book is timely, with the United States and other countries continuing to search for pharmacological defences against unforeseen biological threats such as the 2001 US anthrax attacks, which killed 5 people and infected 17. The current US biodefence effort has been expensive — less so when compared with the cost of certain military aircraft — yet it has not focused adequately on developing products that are usable. Recalling a time when introducing vaccines seemed easier could provide salutary lessons for today.

NATURE.COM
For more on civilian and military disease collaborations:
go.nature.com/thfn8s

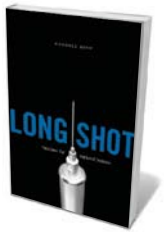
Hoyt begins her deft treatment by describing the threats from pathogens and how she measures vaccine innovation, then reviews vaccine research during and after the Second World War. She also discusses recent factors that have disrupted vaccine-producing government–industrial networks and led to contemporary frustrations.

In the 1940s, vital research aiming to create practical products for protecting soldiers and nations overwhelmed theoretical studies. Most developments were not a response to enemy bioweapons, but rather to viral and bacterial pathogens in countries where troops were based, or to diseases linked to crowding or poor hygiene. For example, influenza A and B vaccine was developed in response to the 1918–19 pandemic that killed tens of millions of people; pneumococcal polysaccharide vaccine came about because of the crowded conditions in barracks that often led to pneumonia; and typhus and tetanus vaccines were produced to combat two common diseases faced by troops throughout history.

As Hoyt shows, the urgent need to find vaccine solutions drove highly effective collaborations between various military, academic and industrial research bodies. But that urgency led to stumbles too. When production of yellow-fever vaccine was scaled up to enable vaccination of US Army troops in the 1940s, the vaccine was stabilized with human serum that was unknowingly contaminated with hepatitis B virus, leading to almost 50,000 hospital cases. Formulation changes later allowed the vaccine to be used to prevent yellow fever across broad swathes of the globe.

The book's narrative is enriched by a description of how the US Office of Scientific Research and Development, the Office of the US Army Surgeon General and the War Research Service conducted vaccine-research planning and operations during the war. With peace in 1945, wartime scientific relationships continued to bear fruit, such as vaccines against adenovirus and meningococcal infections. She lauds collaborations throughout the 1960s between the Walter Reed Army Institute of Research in Silver Spring, Maryland, and vaccine firm Merck, Sharp and Dohme in West Point, Pennsylvania. She describes them as science integrators that brought together experts from disciplines including medicine, immunology and virology, all focused on common goals. Much of their success, she says, is due to that focus and to teamwork that avoided over-specialization.

To measure vaccine innovation between 1900 and 1999, Hoyt reconstructs historical records of annual US vaccine introductions. This was a challenge: responsibility for vaccine registration (or licensing) was transferred from bureau to bureau. Between the establishment of the US Hygienic Laboratory in 1902 to that of the Center for Biologics Evaluation and Research in 1987, there were seven institutional transitions, and multiple variations in record-keeping.



Long Shot: Vaccines for National Defense
KENDALL HOYT
Harvard University Press: 2012. 320 pp. $29.95, £22.95

According to Hoyt, vaccine innovation activity peaked in the 1940s at 50 innovative licences per decade, remained at

35–40 licences per decade through the 1960s, and declined steadily to 12 per decade in the 1990s. The numbers reflect all vaccines developed for endemic and exotic pathogens. She convincingly contests similar analyses.

Today's biological revolution cuts both ways, offering great advances in medicine while providing new means of attack for terrorist groups. It is thus short-sighted that US biodefence funding is largely overseen by congressional committees that are oriented towards health rather than national security.

If the 1940–69 model was a worthy paradigm, why did we migrate away from it? Hoyt attributes the change to several factors. Government contracting and licensing became mired in bureaucracy, and concerns over intellectual property curtailed collaborations. Oversight of vaccine development across a number of bodies became inadequate. The political constituency for biodefence issues weakened and funding for products suffered. Arguably, the lack of US government prioritization of vaccines was the most damaging factor. The result is that now, the time it takes to develop a vaccine is increasing. A better balance of basic and applied research could restore a product-oriented focus.

Government-sponsored research of bioweapon countermeasures has generated knowledge and publications, but faltered in delivering practical results. Biodefences cannot be put in place solely by accepting or dismissing research hypotheses; they require safe and effective pharmaceutical products. We need the biological equivalent of the defences that now protect Hawaii's Pearl Harbor.

So what should society and governments do next? At this point, the book's contributions fade. Hoyt dismisses most vaccine stockpiles as unresponsive to today's threats, yet doesn't indicate any need for quick-acting antibody formulations or therapeutics. Instead, she advocates an emphasis on multipurpose technologies and 'platforms', many of which are hypothetical. How can we measure success using a platform approach? How quickly could platforms be transformed into defences against bioweapons?

Perhaps Hoyt's idea is the right interim solution for today's nadir in national will. Perhaps later, if and when it returns, the well-coordinated, product-centric development of vaccines can resume. History offers us lessons in how to do so. ■

John Grabenstein *is a senior medical director at Merck Vaccines, West Point, Pennsylvania, USA.*
e-mail: john_grabenstein@merck.com

ILLUSTRATION BY ALESSANDRO GOTTARDO

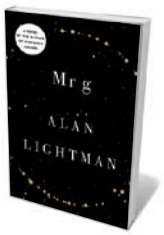
FICTION

Cosmic creation

Pedro Ferreira explores Alan Lightman's latest novel — a magical-reality take on the origins of the Universe.

When a physics heavyweight is mentioned in the same breath as Salman Rushdie and Italo Calvino, it is tough for a reviewer. Few venture into air that rarefied and make it out alive. But when the book is *Mr g*, a creation myth by physicist Alan Lightman, it is worth the risk.

In *Mr g*, Lightman has taken the core of what we know about the origins of the Universe from physics, chemistry and biology and wrapped a few characters around it. The protagonist is the narrator: god, dubbed Mr g. Mr g lives in a timeless Void with his Aunt Penelope and Uncle Deva, playing out skits that could have been lifted from a Woody Allen film — but with the humour on mute. One day, Mr g wakes up from a nap, decides to create a Universe called "Aalam-104729", and from then on marvels at his creation as it evolves and becomes more complex — from the beginning of space and time, to the emergence of fundamental laws, particles, forces, stars, galaxies, planets and, ultimately, life itself.



Mr g: A Novel About the Creation
ALAN LIGHTMAN
Pantheon/Corsair: 2012. 224 pp. $24.95/£9.99

As sentient beings finally emerge out of the cosmic mess, Mr g is torn: should he intervene or let them go their own way? Throughout, he is taunted by the creepy Belhor, a devilish character (a fine role for Al Pacino if this were ever made into a film) and Belhor's annoying daemons, the Baphomet siblings. Belhor pushes Mr g to allow his creations to do their own thing, and watches with glee as evil and unhappiness begin to emerge — leaving Mr g to observe as, for instance,

an impoverished young woman anguishes over stealing meat to feed her starving siblings. Lightman uses the exchanges between Mr g and Belhor to riff on good and evil, free will and relative morality.

Lightman's grasp of the science, in all its gory detail, is unerring. His ability to interweave the fantastical with the factual is impressive — not surprising, given his background. In the 1970s, Lightman established himself as an astrophysical relativist, writing several important papers and books. I still use them. In 1993, he published a magical collection of short stories, *Einstein's Dreams* (Pantheon), in which he took the strange concepts and consequences of Einstein's principle of relativity and wove them into vignettes. I was shocked at how well it worked — and I was not alone. The book has been translated into 30 languages and has led to several stage productions around the world.

Lightman had done something that I had thought impossible: he had brought in the hard science and softened it up. *Einstein's Dreams* is sensual; it breathes. *Mr g* is different. Much more abstract and almost pedagogical, it is a detailed description of the birth and evolution of the Universe that reads like a Rushdiesque fable about an invented place and time. The facts are faultless. Yet the book can occasionally be unintelligible.

For example, Lightman defines a tick of a clock in terms of a particular frequency of the hydrogen atom. Beautifully precise and clear. But as a result, any other timescales that he mentions must be written in scientific notation. And, of course, other big numbers — such as the number of neurons in a brain or of Universes in what he calls the Void — can also be accurately presented only in scientific notation. You begin to wonder about the readership. Are these nuggets meant to be seen as just icons, like hieroglyphs? Or is the generalist supposed to know what they mean?

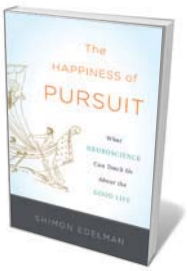
It has to be said too that Lightman's work could have worked well as a short story. It loses steam early on and sometimes feels like interspersed vignettes on science and morality with a dose of 'magic' thrown in. Finely crafted it may be, but it can be hard going. Throughout, I had a pretty good idea of what was going to happen and, given the nature of the story, the characters didn't need to be developed. There was nothing to wait for.

Yet ultimately, this is a marvellous counterpoint to all of the other nonsense out there on creation. Lightman writes exquisitely, so this fable on the origin of space, time, matter and life is a wordfest that is securely pinned to the rational — making him a 'magic realist' of a refreshingly different stripe. ■

Pedro Ferreira *is professor of astrophysics at the University of Oxford, UK, and a member of the Oxford Martin School.*
e-mail: p.ferreira1@physics.ox.ac.uk

Books in brief

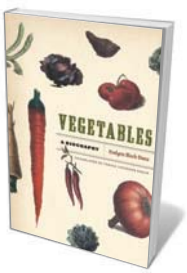
The Green Paradox: A Supply-Side Approach to Global Warming
Hans-Werner Sinn MIT Press *288 pp. $29.95 (2012)*
Frustrated by "counterproductive" climate policies, economist Hans-Werner Sinn proposes a radical alternative. Policy-makers, he notes, have ignored the oil sheikhs and coal barons who supply the fossil-fuel market. Yet it is they who call the shots, as shown by the 'green paradox' — announcements of future reductions in carbon consumption that drive carbon-resource controllers to bump up production. Sinn's antidote to the ideology that plagues policy-making is a "Super-Kyoto" system: unified countries, coordinated caps and trade, and taxation designed to curb the 'extraction habit'.



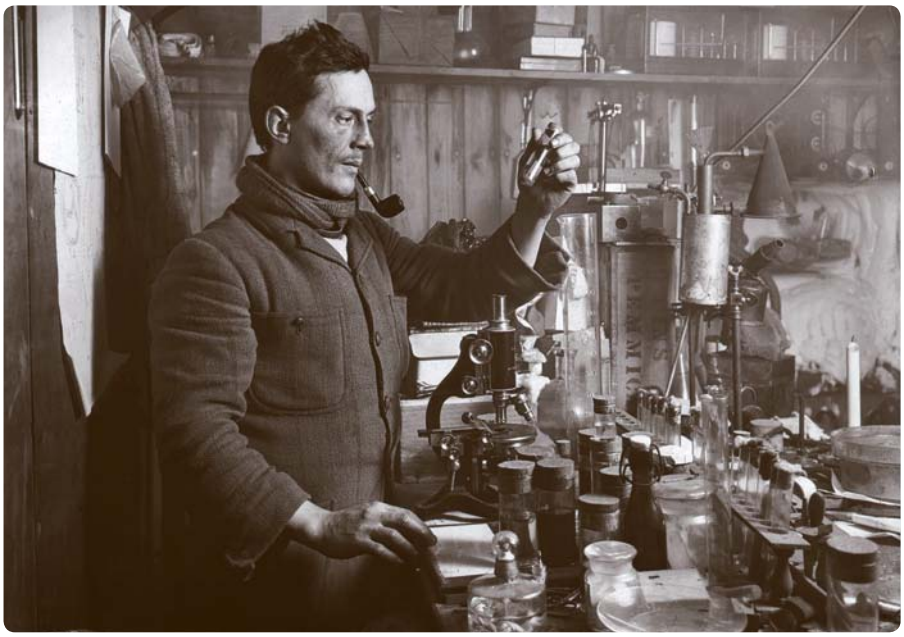
The Happiness of Pursuit: What Neuroscience Can Teach Us About the Good Life
Shimon Edelman Basic Books *256 pp. $25.99 (2012)*
The mind, says cognitive psychologist Shimon Edelman, is a literal "meat computer". Our experience of the world is a series of computations carried out by neural wetware. But where does that leave philosophical conundrums such as joy? Taking passages by luminaries including Homer, William Shakespeare and Jorge Luis Borges as touchstones, Edelman powers along on his "quest for an algorithmic understanding of happiness", revealing that it is this computational journey itself that constitutes the good life.

How Not to Be Eaten: The Insects Fight Back
Gilbert Waldbauer University of California Press *240 pp. $27.95 (2012)*
Insects and their predators use a vast, often bizarre array of strategies to eat or to avoid being eaten. Entomologist Gilbert Waldbauer tours tactics on both sides, from sticky lures to kamikaze speed. His fascinating cast ranges from goatsuckers (birds in the order Caprimulgiformes) that can catch massive moths in their bristled gapes, to the brilliantly hued lubber grasshopper (*Romalea guttata*), which, when threatened, both vomits a noxious substance and hisses as stinking froth erupts from its thorax.

Knowing Nature: Art and Science in Philadelphia, 1740–1840
Edited by Amy R. W. Meyers and Lisa L. Ford Yale University Press *432 pp. $65 (2012)*
When writer Thomas Paine saw Philadelphia in 1774, he noted that nearly every citizen had "some scientific interest or business" — a state that persisted to the mid-nineteenth century. The city boasted artist-naturalists such as John James Audubon, and the study of nature filtered through all strata of society. These 14 illustrated essays, edited by art historians Amy Meyers and Lisa Ford, trace a century of influential scientific endeavour, from botanic gardens and cabinets of curiosity to pioneering colour-plate techniques.



Vegetables: A Biography
Evelyne Bloch-Dano (translated by Teresa Lavender Fagan) University of Chicago Press *128 pp. $20 (2012)*
Next time you chop a carrot, spare a thought for the centuries of breeding, farming technology and trade that brought it to your kitchen. In brief 'biographies' of 11 vegetables, from pumpkins and artichokes to chard, writer Evelyne Bloch-Dano serves up a feast of associated genetics, agriculture, history and culture. From the role of chillies in world trade to the strange link between peas and the court of Versailles in France, there is much to savour.



Biologist Edward Atkinson in the lab at Cape Evans hut, base camp of the Antarctic Terra Nova expedition.

CANTERBURY MUS., NEW ZEALAND

ANTARCTICA

Scientists to the end

Colin Martin celebrates a London exhibition revealing the research legacy of Robert Scott's final journey.

It is one of the most poignant moments in polar history. One hundred years ago this week, British Antarctic Expedition leader Robert Falcon Scott and his four companions reached the spot where, 33 days previously, Roald Amundsen and his party had planted a Norwegian flag. All five members of Scott's party died while attempting to return to their base camp at Cape Evans.

That tragic outcome has tended to overshadow the expedition's scientific achievements. In *Scott's Last Expedition*, London's Natural History Museum aims to redress the balance. As curator Elin Simonsson explains, "We want people to know that Scott's British Antarctic Expedition was not simply a quest to reach the South Pole, but an important scientific expedition that carried out work across many fields."

Those fields included glaciology (then in its infancy), biology, magnetism, meteorology, chemistry and physics. The scientific team that participated in Scott's Terra Nova expedition of 1910–13 numbered 12 — the largest sent to the Antarctic at that time — including zoologist Edward Wilson, geologist T. Griffith Taylor and meteorologist George Simpson.

When the ship *Terra Nova* returned to England, it was laden with specimens of 2,109 animals, plants and fossils — with more than 401 new to science. Geological specimens (1,919 of which are in the museum's permanent collection) helped to construct knowledge of the continent. The meteorological data — the longest unbroken record of weather collected in the early twentieth century — provide baselines that are now used in assessing climate change.

The shore party of 25 men lived and worked in a wooden hut measuring 112 square metres (about half the area of a singles tennis court). Although the hut was equipped with scientific instruments, the scientists' real lab was the Antarctic. The exhibition includes a full-size representation of the hut (the original is preserved at Cape Evans), with actual artefacts from the expedition on display, including scientific apparatus and notebooks as well as a range of biological and geological specimens.

These include the shells of three eggs from the emperor penguin (*Aptenodytes forsteri*; pictured), which contained embryos at different stages of development. Wilson, his assistant Apsley Cherry-Garrard and marine lieutenant Henry Bowers collected them under unimaginably harsh conditions, trudging to the penguin breeding colony at Cape Crozier — a trip immortalized by Cherry-Garrard in his 1922 book *The Worst Journey in the World*. The team's painstaking preservation and study of the embryos ultimately came to little, as theories about bird evolution from reptiles had changed by the time their report was published in 1934.

However, the fossils of the extinct plant *Glossopteris indica*, found by Scott and his final companions, provided important evidence that the climate of Antarctica once supported vegetation and was a part of the supercontinent Gondwana.

The expedition is also documented by photographs, many by photographer Herbert Ponting. These include an iconic image of Scott writing in his diary, and parasitic-infection specialist Edward Atkinson working at the bench. Similarly informative are Scott's own 'lost' photographs. Only a handful of the 120 images he took under Ponting's tutelage had been available until last year, when Wilson's great-nephew, polar historian David M. Wilson, published *The Lost Photographs of Captain Scott* (Little, Brown; 2011). The snowy panoramas and shots of expedition life help to reveal the daily realities of his last endeavour, while polar scientists still use the images for comparison with contemporary photographs to assess climate impacts.

A concurrent exhibition, *These Rough Notes*, is showing at the Scott Polar Research Institute in Cambridge, UK. It takes its title from a journal entry written by Scott shortly before his death — "These rough notes and our dead bodies must tell the tale…" — and displays personal papers of Scott and other expedition members. They record the heartbreaking poignancy of the polar party's death. But as the exhibition at the Natural History Museum shows, we need to reinforce the dual importance of this extraordinary expedition. "We want the scientific work to make the bagging of the Pole merely an item in the results," wrote Wilson at the expedition's outset. He didn't, of course, foresee that his own fate, and that of Scott and other colleagues, would long eclipse an enduring scientific legacy that led to the publication of 81 papers. ■

M. WATSON/ARDEA.COM

Scott's Last Expedition
Natural History Museum, London.
20 January to 2 September 2012.

These Rough Notes
Scott Polar Research Institute, Cambridge, UK.
Until 5 May 2012 .

➲ **NATURE.COM**
For more on Scott's legacy:
go.nature.com/9fbuuj

Colin Martin *is a writer based in London.*
e-mail: cmpubrel@aol.com

Correspondence

Rare day to highlight rare diseases

In the last leap year in 2008, 29 February was chosen to mark Rare Disease Day by EURORDIS, the European Organisation for Rare Diseases. This year, this rare day will again serve to raise awareness of the plight of patients with rare disorders.

Rare diseases fail to attract commercial research efforts because of the small number of people affected (fewer than 5 in 10,000). Examples include the nerve disorder Charcot–Marie–Tooth disease and Proteus syndrome, which causes excess overgrowth of tissues.

As a result, only a handful of scientists, often working in isolation, dedicate their research to a particular rare disease. But such select studies can deliver groundbreaking insights into more general disease processes. Exposure of molecular mechanisms underlying rare diseases therefore has the potential to help many other patients.

Also on 29 February, the first international congress dedicated to rare diseases will be held in Basel, Switzerland (www.react-congress.org). It will highlight the recently established International Rare Diseases Research Consortium (IRDiRC), which plans to coordinate international policy-making with results from national research projects. These initiatives should eventually translate scientific advances into benefits for patients.

Susan M. Gasser *Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. susan.gasser@fmi.ch*
James R. Lupski *Baylor College of Medicine and Texas Children's Hospital, Houston, Texas, USA.*
Yann Le Cam *European Organisation for Rare Diseases, Paris, France.*
Olivier Menzel *BLACKSWAN Foundation, Swiss Foundation for Research on Orphan Diseases, Porza, Switzerland.*

Asian medicine: small species at risk

The demands of traditional Asian medicine (TAM) don't just pose a threat to the survival of tigers and rhinos (*Nature* **480,** S101–S103; 2011). Numerous smaller species are also at risk, as a result of being traded in large volumes.

For example, millions of dried seahorses (*Hippocampus* spp.) and Tokay geckos (*Gekko gecko*) are sold annually for use against impotence and circulatory problems in TAM. Gecko sales have been further fuelled by rumours that larger individuals can cure HIV. Trade in dried orchid parts for eye ailments and in ginseng rhizomes is consuming plants in their millions. The decline in Asia's turtle population is being aggravated by the marketing of TAMs that contain turtle plastrons (the flat underbelly of the shell) to treat skin diseases.

The Convention on International Trade in Endangered Species seeks to protect most of these globally threatened species. The high demand for scaly anteaters, or pangolins (*Manis* spp.), for example, has led to an international trade ban. However, the scales of tens of thousands of pangolins are still confiscated annually in Asia.

For some species, captive breeding and plant propagation have been successful, but the demand for wild-sourced ingredients means that such techniques cannot wholly mitigate the impact of trade.

The killing of endangered tigers and rhinos for TAM will not go unnoticed. But the insidious destruction of smaller species must also be halted by enforcing stricter regulations.

Vincent Nijman, K. Anne-Isole Nekaris *Oxford Brookes University, Oxford, UK. vnijman@brookes.ac.uk*
David P. Bickford *National University of Singapore, Singapore.*

Call to split fisheries at home and abroad

I agree that "no one should doubt that our seas need protection" (*Nature* **480,** 151; 2011) and that establishing marine protected areas alone will not do the job (*Nature* **480,** 14–15; 2011). A combination of measures is needed, including the elimination of overfishing subsidies. These incentives were introduced when fisheries seemed inexhaustible, but they inflate profitability and drive fishing beyond economic or sustainable levels.

The global community mandated the World Trade Organization (WTO) to discipline overfishing subsidies more than ten years ago, but the issues were still unresolved at their meeting last month. One reason is that WTO negotiators are trying to broker an all-inclusive deal that encompasses domestic and international, small- and large-scale fisheries. But this approach is hindered by national interests.

The answer is to split the world's fisheries into domestic and international ones. Domestic fisheries would operate within a country's economic exclusion zone and target fish stocks that spend all their lives there. This split is necessary because the incentives to eliminate overfishing subsidies differ according to whether a fishery is domestic or international.

For a domestic fishery, the heavy lifting should be on the home front; for an international fishery, global coordination would be needed, because unilateral action by one country will not eliminate overfishing. Categorizing fisheries in this way would make it easier to identify leverage points for eliminating overfishing subsidies.

U. Rashid Sumaila *University of British Columbia, Vancouver, British Columbia, Canada. r.sumaila@fisheries.ubc.ca*

Investors unfazed by drug-patent expiry

Your data on changing stock prices for five pharmaceutical companies from 1997 to 2010 (*Nature* **480,** 16–17; 2011) erroneously indicate a 39% drop in aggregated share value, when in fact it would have risen by 82% (see Correction, *Nature* **480,** 425; 2011). The error was due mainly to the selection of unadjusted, rather than adjusted, stock-market closing prices.

The corrected data set indicates that the aggregated share prices rose roughly in line with the Dow Jones Industry Average. Considering that these companies have the biggest patent cliffs in the industry, this result is very positive; moreover, it argues against your implication that the impending expiry of their patents is affecting investor confidence in pharmaceutical stocks.

Investors accept that the strength of a company's patent portfolio is only one of several factors. Hence, despite the impending expiry of patents for blockbuster products, including Pfizer's Lipitor, the performance of the five worst-hit companies was comparable with that of other sectors for the period 1997–2010.

Other factors such as decreasing revenues can also affect investor confidence. Your simplistic view that the patent cliff is the only contributor does an injustice to the efforts of pharmaceutical companies and governments, and unnecessarily perpetuates a false uncertainty.

David Brindley *University College London, UK; Harvard Stem Cell Institute and Harvard Business School, Massachusetts, USA. david_brindley@harvard.edu*
Brock Reeve *The Harvard Stem Cell Institute, Cambridge, Massachusetts, USA.*
Chris Mason *University College London, UK.*

Paul Mead Doty

(1920–2011)

Chemist and arms-control advocate who unravelled the structures of biomolecules.

Paul Mead Doty, who died of heart failure at home on 5 December 2011 at the age of 91, leaves three legacies: his revelation of the structures of numerous proteins and nucleic acids; his contribution to our very survival through his tireless activity on behalf of arms control and disarmament; and the many successful scientists and policy figures whom he mentored.

Doty was born on 1 June 1920 in Charleston, West Virginia, and was schooled in Pennsylvania. His exceptional talent was recognized at high school, where a chemistry teacher asked him to take over the class. Doty earned a Bachelor's degree in chemistry at Pennsylvania State University (then Pennsylvania State College) in University Park in 1941 and a PhD in chemistry at Columbia University in New York in 1944. He was renowned for being late. On the first day of classes at Columbia, as the students were seated in alphabetical order, he missed his place. So Doty sat at the end, and thereby made the acquaintance of Bruno Zimm, a friendship that lasted a lifetime and would lead to seminal scientific collaborations.

During his doctorate, Doty measured the electron affinity of bromine. He moved into the new field of polymer science and, with Bruno Zimm, pioneered the theory and use of light scattering to characterize polymer molecules in solution. A postdoctoral position at the Polytechnic Institute of Brooklyn, New York, brought Doty into contact with founders of polymer science, such as Herman Mark, and alerted him to the unusual chemistry and physics of macromolecules. During a second postdoc, at the University of Cambridge, UK, Doty learned, mainly from Max Perutz, about the X-ray diffraction of protein crystals, which allowed their structures to be determined.

BIOLOGICAL MACROMOLECULES

After briefly joining the faculty at Notre Dame University, Indiana, Doty accepted an assistant professorship at Harvard University in Cambridge, Massachusetts, in 1948. He remained at Harvard for the rest of his life. During his tenure there, he founded the Department of Biochemistry and Molecular Biology and the Center for Science and International Affairs.

Doty's research focus shifted at Harvard to large biological molecules. Realizing that proteins could not be random polymers, because they could be crystallized and had unique structures, he established a programme to study synthetic polypeptides. These chains of amino acids allow structural transitions, such as from helix to random coil, to be followed by a technique called optical rotatory dispersion. By tracing

S. WILKE/HKS BELFER CENTRE

structural transitions induced by changes in temperature, he was able to estimate the fraction of each protein that was helical. These results laid the foundations for modern studies of protein folding.

He then, in the 1950s, turned to nucleic acids, a step made possible by the availability of synthetic polynucleotides. Out of this work came the discovery that DNA structure could be altered using heat, that the melting temperature of DNA from various bacteria differed according to the fraction of G and C nucleotides in it, and that RNA contains regions of secondary structure. These observations are the basis of much modern biotechnology, including DNA sequencing and the polymerase chain reaction.

Doty's strength lay in identifying important scientific issues with conceptual and practical implications that could be resolved by devising simple model systems. He was adept at finding the right person for a particular job, hiring Jacques Fresco for the polynucleotide work, for example, and Julius Marmur for studies using DNA. He was discriminating in how he worked with his students, balancing a laissez-faire attitude with specific guidance according to each student's needs. As PhD students, we both benefited from this style of mentorship.

PUBLIC AFFAIRS

His most important choice of associate was a graduate student, Helga Boedtker, whose thesis on the structure of collagen in solution became a landmark. They married, and Helga became the mainstay of the laboratory for more than 40 years until her death in 2001. When Doty was absent from the lab during his trips abroad to promote arms control, a quiet word to Helga brought urgent matters to his attention.

Very early in his career, Doty became active in science and public affairs, with a focus on US–Soviet relations. He chaired the Federation of American Scientists and the US Pugwash Committee, worked on disarmament plans under President Dwight Eisenhower's science adviser James Killian, and was a member of President John F. Kennedy's Science Advisory Committee. Doty co-chaired the Soviet–US disarmament committee (now the Committee on International Security and Arms Control of the National Academy of Sciences), helped to arrange the first visits to Moscow of US secretaries of state Henry Kissinger and Cyrus Vance and played a key part in developing the basis for the US–Soviet Anti-Ballistic Missile Treaty.

As noted by molecular biologist Matthew Meselson, Doty "surely helped to avoid catastrophe". ■

Stuart A. Rice *is in the Department of Chemistry and the James Franck Institute, University of Chicago, Illinois 60637, USA.*
Robert Haselkorn *is in the departments of chemistry, and of molecular genetics and cell biology, University of Chicago, Illinois 60637, USA.*
e-mail: rh01@uchicago.edu

materials opens the door to the hierarchical assembly of functional materials that could react sensitively to changes in ambient conditions or to mechanical stimuli. One way to extend the functionality of their system could be to make colloidal particles that are not only chiral, but also have surfaces patterned with microscopic domains that attract or repel each other. Such patterned colloidal particles have recently been shown to self-assemble into useful functional structures such as 'kagome' lattices[6].

Chiral interactions between particles occur in all kinds of materials, from liquid crystals[2,3] to cytoskeletal filaments in cells (for which the hierarchical assembly of bundle-like structures is dependent on, and sensitive to, the helical twist of the filaments[7,8]). Most of the chiral materials studied so far are passive or in thermal equilibrium. This means that their structural assembly is governed only by diffusion and by local interactions between the constituent particles. But most naturally occurring materials are far from passive. Instead, they constantly consume energy so that their particles self-organize into higher-order structures — the cytoskeletons of cells are prime examples of this[9,10].

The next step, therefore, is to apply the principles identified by Gibaud and colleagues[1] to active systems. Imagine replacing the viruses used in this study with self-propelling chiral bacteria — would they still pack together in a two-dimensional membrane, and, if so, would the membrane move, or rotate, across mesoscopic or macroscopic distances? Could actively beating ribbons form, or would the noise stemming from active movement of the bacteria prevent the formation of such structures? All we can say for certain is that chiral interactions will add a new twist not only to the self-assembly of colloids, but also to the self-organization of active materials. ■

Volker Schaller *and* **Andreas R. Bausch** *are at the Lehrstuhl für Biophysik-E27, Technische Universität München, Garching 85748, Germany.*
e-mail: abausch@mytum.de

1. Gibaud, T. *et al. Nature* **481,** 348–351 (2012).
2. Coles, H. J. & Pivnenko, M. N. *Nature* **436,** 997–1000 (2005).
3. Hough, L. E. *et al. Science* **325,** 456–460 (2009).
4. Aarts, D. G. A. L., Schmidt, M. & Lekkerkerker, H. N. W. *Science* **304,** 847–850 (2004).
5. Barry, E. & Dogic, Z. *Proc. Natl Acad. Sci. USA* **107,** 10348–10353 (2010).
6. Chen, Q., Bae, S. C. & Granick, S. *Nature* **469,** 381–384 (2011).
7. Claessens, M. M. A. E., Semmrich, C., Ramos, L. & Bausch, A. R. *Proc. Natl Acad. Sci. USA* **105,** 8819–8822 (2008).
8. Grason, G. M. & Bruinsma, R. F. *Phys. Rev. Lett.* **99,** 098101 (2007).
9. Kasza, K. E. *et al. Curr. Opin. Cell Biol.* **19,** 101–107 (2007).
10. Köhler, S., Schaller, V. & Bausch, A. R. *Nature Mater.* **10,** 462–468 (2011).

GENOMICS

The path to retinoblastoma

Genomic analyses of tumours of the childhood cancer retinoblastoma reveal a low mutation rate, challenging the view that genomic instability is crucial for its progression. The work also identifies a new therapeutic target. SEE ARTICLE P.329

JULIEN SAGE & MICHAEL L. CLEARY

Retinoblastoma is a rare tumour that affects retinal cells in the eyes of children. Analyses of familial and sporadic cases of this cancer, backed by studies in genetically engineered mice, have shown that loss of function of the tumour-suppressor protein RB1 (also known as RB) is required for the development of most, if not all, tumours of this type. However, it is not clear how RB1-deficient retinal cells progress to malignant tumour cells[1]. In addition, emerging evidence that loss of RB1 function can induce genomic instability[2] has raised the tantalizing possibility that RB1-deficient retinal cells might be predisposed to accumulating many additional mutations, further complicating the identification of mutations that contribute to the development and maintenance of retinoblastoma. On page 329 of this issue, however, Zhang *et al.*[3] demonstrate that retinoblastoma genomes have very few recurrent mutations in genes other than *RB1*. Instead, the expression of cancer-related genes is affected by epigenetic modifications on chromosomes, which do not affect DNA sequence but are inherited after cell division.

To identify mutations that could cooperate with loss of RB1 function in tumour development, Zhang and colleagues[3] sequenced and compared the genomes of normal tissue and retinoblastoma tumours from four patients. The researchers found that *RB1* was the only known cancer-related gene consistently mutated, and that the retinoblastomas had 15-fold fewer total mutations than other types of solid tumour whose genomes have been sequenced[4].

Next, the authors searched for epigenetic alterations and for abnormal gene expression in retinoblastoma cells. They identified the gene that encodes the protein kinase enzyme SYK as a potential oncogene whose expression is consistently higher in retinoblastoma cells than in normal immature retinal cells. Moreover, the activity of SYK was essential for the growth of retinoblastoma cells. The authors also show that certain small molecules that selectively inhibit SYK activity induce the death of retinoblastoma cells in a mouse model of the disease.

These findings[3] indicate that SYK may be a promising target for treating patients with retinoblastoma. *SYK* was not identified in previous searches for genes with a role in this cancer because it is not mutated or structurally rearranged in retinoblastoma. Therefore, Zhang and colleagues' study emphasizes the importance of high-throughput approaches that integrate genome sequencing with gene-expression analysis and epigenomics to identify cancer genes.

Low mutation frequency has also been observed[5] in medulloblastoma (a type of brain tumour that affects children), suggesting that it could be a general feature of childhood cancers. A possible explanation for this difference between the cancers of children and adults is that childhood tumours arise in cells that are naturally undergoing rapid developmental growth, with fewer brakes on their proliferation than cells in adults. An alternative explanation is that, in children, these cells are negotiating crucial developmental checkpoints that are susceptible to corruption, leading to incomplete or abnormal maturation. In both cases, only a few mutations would be needed to trigger the cellular changes associated with cancer.

Furthermore, epigenetic changes in children and excessive mutations in adults may have similar roles in cancer development. Another childhood cancer, Wilms' tumour, also has a relatively stable genome and displays an increased variation in DNA-methylation patterns compared with normal cells[6]. RB1-deficient retinal cells may be particularly susceptible to this tumour-formation mechanism, because RB1 interacts with the machinery that controls the epigenetic status of cells, including enzymes that remodel chromatin (DNA–protein complexes) and other enzymes that add methyl groups to DNA[1]. Thus, childhood cancers such as retinoblastoma may carry epigenetic abnormalities that change gene expression and are stably propagated through subsequent cell divisions, helping to maintain tumour-associated features.

If this model is correct, it is possible that RB1-deficient tumours in adults — such as small-cell lung cancer and some breast,

prostate or blood cancers — also display epigenetic changes that are crucial for tumour development and maintenance. Future experiments, including integrated genomic analyses, may provide the answer.

The underlying mechanism of alterations in gene expression by epigenetic means in retinoblastoma is still unclear. But Zhang and co-workers' observation[3] that 13% of these tumours show recurrent mutations in the *BCOR* gene offers a possible explanation. The BCOR protein is highly expressed in the fetal retina and is essential for eye development, as evidenced by a congenital eye disorder, syndromic microphthalmia, which results from a heritable mutation in *BCOR*[7]. BCOR associates with proteins that repress gene expression epigenetically[8]. This raises the possibility that loss of BCOR function, due to acquired mutations in its gene, may contribute to an altered epigenetic landscape in RB1-deficient retinal cells. Consequently, inappropriate expression of crucial genes may impair cell maturation and so facilitate the progression of retinoblastoma. It will be of interest to determine whether BCOR is part of the repression machinery that silences SYK expression in the normal retina, and whether acquired *BCOR* mutations provide at least one route to altered expression of SYK in retinoblastoma.

Nevertheless, the current work — using a comprehensive, integrated genomics approach — is notable not just for demonstrating that epigenetic alterations have a predominant role in the progression of retinoblastoma. The new possibilities it raises for therapies in this childhood malignancy, and possibly in other types of tumour, are equally noteworthy. ■

Julien Sage *is in the Departments of Pediatrics and of Genetics and* **Michael L. Cleary** *is in the Departments of Pathology and of Pediatrics, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California 94305, USA.*
e-mails: julsage@stanford.edu; michael.cleary@stanford.edu

1. Chinnam, M. & Goodrich, D. W. *Curr. Top. Dev. Biol.* **94,** 129–169 (2011).
2. Manning, A. L. & Dyson, N. J. *Trends Cell Biol.* **21,** 433–441 (2011).
3. Zhang, J. *et al. Nature* **481,** 329–334 (2012).
4. Stratton, M. R. *Science* **331,** 1553–1558 (2011).
5. Parsons, D. W. *et al. Science* **331,** 435–439 (2011).
6. Hansen, K. D. *et al. Nature Genet.* **43,** 768–775 (2011).
7. Ng, D. *et al. Nature Genet.* **36,** 411–416 (2004).
8. Fan, Z. *et al. Nature Cell Biol.* **11,** 1002–1009 (2009).

EVOLUTIONARY BIOLOGY

A ratchet for protein complexity

Molecular machines containing related protein subunits are common in cells. Reconstruction of ancient proteins suggests that this type of complexity can evolve in the absence of any initial selective advantage. SEE LETTER P.360

W. FORD DOOLITTLE

Organisms and cells are bewilderingly complicated, and the molecular machines that perform many basic cellular functions are often giant, multisubunit, multifunctional protein complexes with tangled evolutionary histories. It is generally assumed that such complexes arose by the stepwise accretion of individual proteins, each addition representing a selective advantage by adding to or refining the machine's performance. But on page 360 of this issue, Thornton and colleagues[1] argue against this standard explanation in one particular instance — that of a ring-shaped protein complex in fungi. The authors show how evolutionary processes entailing loss of function rather than gain might initially drive a system towards complexity, independently of selection*.

Many cellular molecular machines contain several proteins that self-assemble into a multi-subunit ring. In simple cases, rings are homo-oligomeric; that is, all the subunits are identical and thus probably the products of a single gene. In more complex (hetero-oligomeric) examples, the protein molecules in the ring are different, but may be related. Often, hetero-oligomeric rings seem to have arisen from homo-oligomeric complexes after a gene encoding a single subunit became duplicated, producing two or more gene copies (called paralogues), with each copy subsequently evolving to encode a slightly different protein.

The question that Thornton and colleagues address is, why does this happen? Do the structurally distinguishable subunits and/or their specific pattern of assembly confer improved or additional functions on the protein complex, with selection for their enhanced performance being the evolutionary driving force? Alternatively, might neutral processes be responsible, at least initially?

To understand how the latter possibility might come about, imagine that a gene encoding, for example, eight subunits of a homo-oligomeric ring undergoes duplication (Fig. 1). The two paralogous gene copies that this duplication produces — and the protein subunits they encode (call them A and B) — will of necessity begin to diverge through accumulation of neutral mutations. These mutations might not at first affect any of the subunits' functions, so that functional octameric rings could continue to form by a random sampling

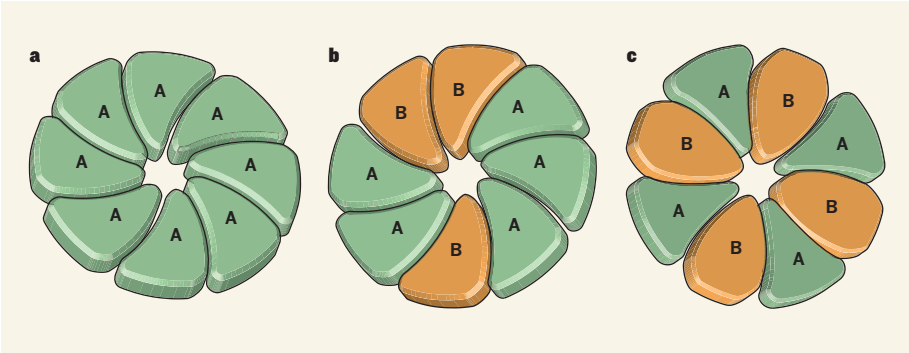


Figure 1 | Evolution of complexity in a protein ring. **a**, Homo-oligomeric protein complexes contain identical subunits (A) that are typically encoded by a single gene. **b**, If the gene is duplicated, the two gene copies can diverge through the accumulation of neutral mutations, generating structurally distinguishable but functionally unaltered subunits A and B, which can form functional hetero-oligomeric rings by random mixture. **c**, If additional mutations prevent subunits from binding to others of their own type, functional rings could still be formed by alternating subunit types. As further mutations accumulate, the probability of returning to the initial homo-oligomeric situation becomes very small. Thus, the subunit composition of a protein association may become complex in the absence of initial selective advantages. Thornton and colleagues[1] provide experimental evidence suggesting that this type of process has occurred in the evolution of a six-membered protein ring that forms part of the vacuolar H^+-ATPase enzymes in fungi.

*This article and the paper[1] under discussion were published online on 8 January 2012.

50 Years Ago

At the International Geological Congress in Copenhagen in 1960, and at the Russian Trade Exhibition in London last year, one of the outstanding scientific exhibits was a hand-coloured tectonic map of Europe compiled by Russian draughtsmen from copy submitted by the various national geological surveys. The map, on a scale of 1:2,500,000, has been sponsored by the Sub-Commission on the Tectonic Map of the World of the International Geological Congress … Western geologists who wish to make sure of receiving a copy should place an order, in advance of publication, with their national agents for Russian books.
From *Nature* 20 January 1962

100 Years Ago

The possibility of the discovery of a remedy for cancer has been advanced a stage by the preparation by Prof. Wassermann, of Berlin, of a substance which possesses a curative action experimentally on cancer of mice. Prof. Wassermann reasoned that since the cancer-cells are growing rapidly, their oxygen requirements would be different from, and greater than, those of the cells of the body generally. He sought for some substance which might interfere with the oxygen supply to the cancer-cells, and finally adopted selenium as a means to do this. The next problem was to convey selenium to the cancer-cells by means of the blood stream, and … a compound of selenium with an anilin dye eosin was found to fulfil this condition … After two or three injections of the substance into a mouse the subject of cancerous tumours, the tumours are found to have softened, and after six to eight doses they become cystic, diminish in size, and finally disappear, and no recurrence takes place. The eosin–selenium compound is, however, poisonous, and a certain number of mice succumb under the treatment.
From *Nature* 18 January 1912

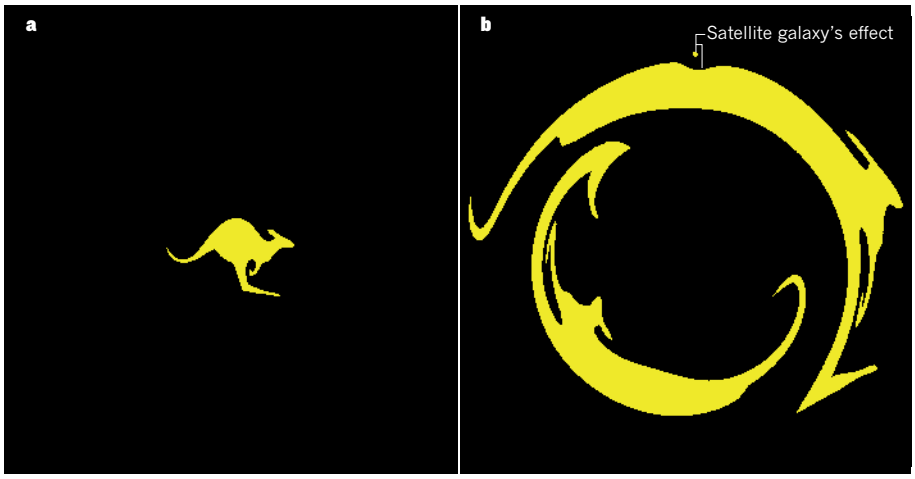


Figure 1 | Seeking satellites. Vegetti *et al.*[4] have discovered a satellite galaxy around a larger galaxy, located at a cosmological distance, by examining how both galaxies act as gravitational lenses and distort light from a background source as the light travels to Earth. **a**, Were there no lenses between an observer on Earth and a source — here illustrated as a kangaroo[10] — the observer would, if instrumentation allowed, see one image of the source. **b**, If there is a massive lens between the observer and the source, the source is magnified and imaged more than once. In this case, as in the authors' study, the source is deformed into a ring that consists mainly of two distorted images of the source. The two images are caused by the larger galaxy, whereas the dent and blob features (labelled) are caused by the satellite galaxy, the mass of which has been exaggerated to demonstrate the effect more clearly.

In the present study, the gravitational lens system, called JVAS B1938+666, consists of a very distant galaxy that bends light from an even more distant background galaxy. Light from the background galaxy is deformed into a ring around the lens galaxy — a prime example of what is known as an Einstein ring. In this case, the ring is formed mainly by two lensed images of the background galaxy[5].

Vegetti and her team[4] obtained a near-infrared image of JVAS B1938+666 using the 10-metre Keck telescope in Hawaii. They used an optics system that corrects for the blurring effect of Earth's atmosphere to improve the image quality. With such data, they could neatly determine the mass distribution of the lens galaxy, as well as the shape and brightness of the background galaxy.

And here comes the connection to satellite galaxies. The sophisticated numerical technique[6] used by the authors allowed them not only to derive a model of the lens galaxy's mass, but also to map any excess lens mass that could not be accounted for by the galaxy. They found an excess mass near the Einstein ring that they attributed to the presence of a satellite galaxy (Fig. 1). This method has been dubbed gravitational imaging[6]. Vegetti *et al.* also used an analytical model to test the detected excess mass and found that a satellite galaxy is indeed required to explain the data.

This satellite galaxy is exciting because it was detected in the excess-mass map despite its low mass. Assuming that the object is in the neighbourhood of the main lens galaxy, it has a mass of some 113 million solar masses within a radius of about 600 parsecs (2,000 light years) — values that put it firmly in the realm of satellite galaxies. The mass is tenfold lower than those of two other satellite galaxies that have been detected[7,8] recently using gravitational lenses (with the possible exception of another low-mass satellite[8], although the detection of this object is not yet confirmed).

Vegetti and colleagues[4] went on to determine the mass function of satellite galaxies — that is, the expected number of satellites for a given mass — at distances beyond the local Universe by combining data from the new object and another satellite galaxy[7]. They found that the resulting mass function is consistent with current galaxy-formation theory. But owing to the small number of known objects, uncertainties in the deduced function are quite large.

A natural question to ask is whether the satellite galaxy can be observed directly rather than by its gravitational effect on the shape of a background object. With current instrumentation, the answer is no. The object is simply too distant to be imaged directly. But the message here[4] is that it is possible to spot these elusive objects around distant lens galaxies without knowing where to look for them.

The satellite should be considered in the context of about 50 satellite galaxies that are known to exist in the Local Group and that have been found by imaging their starlight[3,9]. The number of these nearby satellites has greatly increased over the past decade, and often only a hundred (or fewer) stars are detected in them. From measurements of the velocities of their member stars, the satellites have been shown to be dominated by an unseen mass component called dark matter. Dark matter is the reason that satellites such as the newly discovered one are massive enough to be found using gravitational lenses.

Uhlmann go to some lengths to exclude the involvement of Cdc14 localization within the nucleus, the dependence of dephosphorylation on specific events associated with mitotic exit, and the specificity that results from the activation of CDK by its different activating partners (cyclins).

Uhlmann and colleagues[7] have previously suggested that a CDK-based oscillator acting as a counterbalance to protein phosphatases might be responsible for order not just during mitotic exit, but throughout cell division. However, there is substantial biochemical and genetic evidence for interdependent relationships that impose order throughout the cell cycle[6]. For example, a recent study[8] demonstrated that the sequential action of specific cyclin–CDK complexes is required to eliminate Sic1 — a potent CDK inhibitor that restricts passage from G1 into S phase. The researchers[8] report that the order of action of those cyclin–CDK complexes is due not to differences in their intrinsic activity, but to differences in their specificity for particular phosphorylation sites on Sic1. That is, phosphorylation of a set of sites by one cyclin–CDK complex depends on prior phosphorylation of Sic1 by a different cyclin–CDK complex.

It seems, therefore, that multiple mechanisms are involved in imposing order on cell-cycle events. Nevertheless, the finding[2] that a single cyclin–CDK complex can orchestrate an effective cell-division cycle suggests that, at least in some systems and under certain physiological conditions, the specificity of different cyclin–CDK complexes is dispensable. It remains to be established whether phosphatase-based mechanisms similar to those described by Bouchoux and Uhlmann[3] are essential for ordering cell-cycle events beyond mitotic exit, and whether they are even sufficient to replace cyclin–CDK specificity under particular conditions. It is largely because the other phosphatases that provide the counterpoint to CDK are poorly understood that the choreography of this intricate dance remains incompletely described. ■

Curt Wittenberg *is in the Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92024, USA.*
e-mail: curtw@scripps.edu

1. Morgan, D. O. *Annu. Rev. Cell Dev. Biol.* **13,** 261–291 (1997).
2. Coudreuse, D. & Nurse, P. *Nature* **468,** 1074–1079 (2010).
3. Bouchoux, C. & Uhlmann, F. *Cell* **147,** 803–814 (2011).
4. Queralt, E. & Uhlmann, F. *Curr. Opin. Cell Biol.* **20,** 661–668 (2008).
5. Murray, A. W. & Kirschner, M. W. *Science* **246,** 614–621 (1989).
6. Sullivan, M. & Morgan, D. O. *Nature Rev. Mol. Cell Biol.* **8,** 894–903 (2007).
7. Uhlmann, F., Bouchoux, C. & López-Avilés, S. *Phil. Trans. R. Soc. Lond. B* **366,** 3572–3583 (2011).
8. Kõivomägi, M. *et al. Nature* **480,** 128–131 (2011).

ORGANIC CHEMISTRY

Stabilizers cause instability

Any chemist will tell you that amide compounds are renowned for their stability, yet a family of amides has been discovered that is surprisingly unstable. The cause of this phenomenon is even more surprising.

JONATHAN CLAYDEN

Some molecules are stable under all kinds of conditions; others are so unstable that they'll react with almost anything. Amides, the class of compounds that includes proteins and nylon, are usually stable molecules. Strong acid or alkali is needed to break them down — nylon fabrics are highly resistant to chemical damage, and dietary proteins can be digested only by the acidity of the stomach coupled with a battery of enzymes. Writing in *Angewandte Chemie*, Hutchby *et al.*[1] report a group of amides that turns this typical character on its head. These compounds share structural characteristics with the least reactive amides of all, yet a remarkable mechanistic quirk means they exhibit reactivity that is usually associated with unstable reagents.

In most cases, the reactivity of a molecule can be predicted with a fair degree of certainty by considering the combined effects of bond strength, polarity and steric hindrance (the physical volume occupied by parts of a molecule). For example, a carbon–iodine bond is more reactive than a carbon–fluorine bond because it is much weaker, whereas an oxygen–hydrogen bond is more reactive than a carbon–hydrogen bond because it is more polarized. These effects mean that the most interesting properties of organic molecules tend to be concentrated in reactivity hotspots known as functional groups, each of which contains one or more atoms that either attract or repel electrons more than does carbon or hydrogen. The concept of the functional group, and the fact that functional groups have a typical reactivity profile, is central to the ability of chemists to assess, predict and design the properties of molecules.

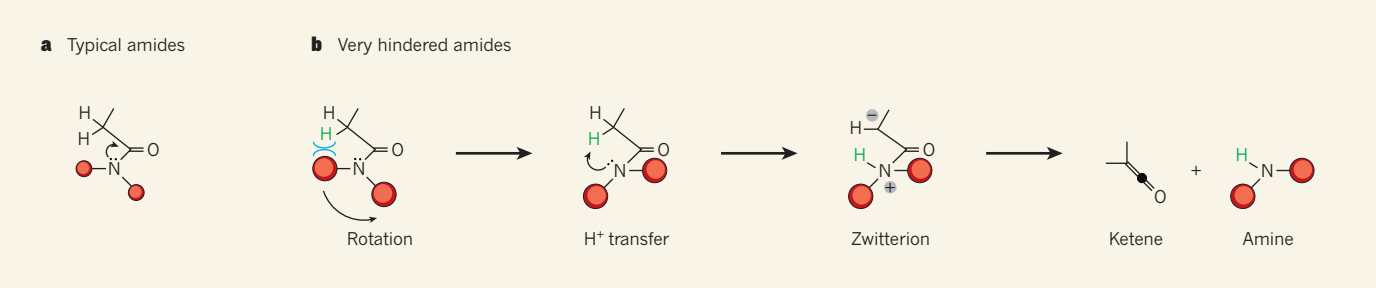


Figure 1 | Amide reactivity. **a**, Most amides are unreactive because the nitrogen atom in the compounds donates a pair of electrons (shown as dots) to the carbonyl (C=O) group, and because the functional groups (red spheres) on the nitrogen atom may block the approach of attacking reagents. The curly arrow indicates electron donation. **b**, Hutchby *et al.*[1] report that very hindered amides, which have outsize groups on the nitrogen atom, buck the trend, because the groups are repelled (blue curved lines) by other groups in the molecule. This repulsion facilitates rotation of the bond between the nitrogen and the carbon of the carbonyl group, and so of the groups on the nitrogen atom, and prevents donation of the nitrogen's electron pair to the carbonyl group. Instead, a hydrogen atom (green) on the carbon adjacent to the carbonyl group is transferred as a H^+ ion to the nitrogen, forming a zwitterion (a molecule that contains both a positive and a negative charge), which falls apart to produce a ketene and an amine. The black dot in the ketene represents a carbon atom.

One particular family of functional groups — those that contain a carbonyl group, C=O — is of particular importance because its reactions are central to most common methods of constructing carbon–carbon bonds. The carbonyl group is electrophilic (it reacts with electron-rich species such as alcohols and amines), but just how electrophilic it is depends enormously on what else is attached to the group's carbon atom. Attach an electron-withdrawing chlorine atom, and a highly reactive acid chloride results; attach two chlorine atoms and you get phosgene, an even more reactive (and toxic) gas. Appending an electron-rich nitrogen atom, on the other hand, satisfies the carbonyl group's desire for electrons because the nitrogen shares its own electron pair with the group (Fig. 1a), a phenomenon known as delocalization. The resulting compound is an unreactive amide.

Amides are unreactive not just because the nitrogen atom satisfies the electron deficiency of the carbonyl group, but also because the nitrogen atoms of amides can carry two substituents, both of which can get in the way of incoming reagents (the phenomenon of steric hindrance; Fig. 1a). Hindered amides, in which the nitrogen atom carries two big groups, are usually among the least reactive carbonyl compounds known. This makes them valuable in organic chemistry — with the reactivity of the carbonyl group shut down, reactions can occur elsewhere without interference from competing side reactions at the carbonyl group. Hindered amides are used in this way in the preparation of reactive, synthetically useful organometallic compounds[2], for example, and can be even less reactive towards electron-rich reagents than is the proverbially stable benzene ring[3,4].

The group of compounds now reported by Hutchby *et al.*[1] overturns these expectations. The authors prepared highly encumbered amides in which the nitrogen atom carries two very large groups, and observed that, far from being the least reactive of their class, the compounds were more reactive than their less encumbered relatives. A possible explanation for this is an effect that takes over from steric hindrance when the groups carried by an amide's nitrogen atom are abnormally large: the groups effectively switch off the carbonyl-deactivating effect of the nitrogen's pair of electrons (Fig. 1b).

Such 'non-delocalized' amides, in which the nitrogen's electrons are forced into a position from which they cannot interact with the carbonyl group, have been made before[5,6], and are indeed more reactive. But Hutchby *et al.* propose[1] that the 'off switch' is more subtle in their case. They found that these amides have structural characteristics very similar to unhindered amides, but that the repulsion between the nitrogen's large groups and other parts of the molecule makes it much more favourable for the groups to move away from their neighbouring atoms. This allows the bond between the carbonyl group and the nitrogen atom to rotate (Fig. 1b), in turn allowing the nitrogen atom to take back its previously delocalized, deactivating electrons. Those electrons therefore become available to grab hold of a hydrogen ion (H^+) from the other side of the carbonyl group, setting in train a series of events — again accelerated by the size of the groups on the nitrogen — in which the molecule is forced to fragment into a compound called a ketene, which is much more reactive than an amide. In the authors' experiments, the ketene subsequently reacted with the surrounding solvent (methanol), forming ester compounds as the final products.

Not all highly encumbered amides will behave as Hutchby and colleagues' compounds do, only those that have a methylene group (CH_2) on the other side of the carbonyl group from the nitrogen atom. And for the process to work really well, an electron-withdrawing functional group must be attached to the methylene group. Nevertheless, it is remarkable that the least reactive of carbonyl compounds can so easily be bumped up the reactivity scale, and even more so that the very change responsible — increased molecular bulk — usually decreases reactivity. This discovery exposes a chink in the armour of amides that could be used to circumvent many current and future challenges in synthetic chemistry. For example, amides could be used as a Trojan horse to incorporate hidden reactivity into a molecule, reactivity that is later revealed using Hutchby and colleagues' chemistry when other modifications to the molecule have been made. ■

Jonathan Clayden *is in the School of Chemistry, University of Manchester, Manchester M13 9PL, UK.*
e-mail: clayden@man.ac.uk

1. Hutchby, M. *et al. Angew. Chem. Int. Edn* **51,** 548–551 (2012).
2. Snieckus, V. *Chem. Rev.* **90,** 879–933 (1990).
3. Clayden, J., Foricher, Y. J. Y. & Lam, H. K. *Chem. Commun.* 2138–2139 (2002).
4. Clayden, J., Foricher, Y. J. Y. & Lam, H. K. *Eur. J. Org. Chem.* 3558–3565 (2002).
5. Clayden, J. & Moran, W. J. *Angew. Chem. Int. Edn* **45,** 7118–7120 (2006).
6. Szostak, M. & Aubé, J. *Org. Biomol. Chem.* **9,** 27–35 (2011).

CANCER

Reprogramming clinical outcome

On binding to its target hormone, the oestrogen-receptor protein modulates the expression of many genes. Changes in the receptor's interaction with DNA have now been linked to clinical outcome in patients with breast cancer. SEE LETTER P.389

GENEVIÈVE DEBLOIS & VINCENT GIGUÈRE

In about two-thirds of breast cancers, tumour growth depends on oestrogen. This hormone binds to and activates its nuclear receptor, which then binds to specific sites on the genome and modifies gene expression. Hormonal therapies that inhibit oestrogen synthesis or that block the oestrogen receptor with drugs such as tamoxifen can restrain the proliferation of cancer cells in oestrogen-dependent breast tumours. Consequently, the presence of this receptor and the expression of its target genes in primary tumours are generally predictors of a good prognosis, as the tumour will be amenable to these treatments. But despite the sustained presence of the receptor, these tumours can eventually develop resistance to hormonal therapies and spread to distant sites, leading to a poor outcome. On page 389 of this issue, Ross-Innes *et al.* (Carroll and colleagues)[1] present evidence that reprogramming of the interaction of the oestrogen receptor with DNA in breast tumours significantly contributes to this change in clinical outcome.

Functional genomics — genome-wide studies of dynamic interactions between genes and proteins — has greatly enhanced our understanding of the action of the oestrogen receptor (ERα) in breast cancer[2]. Such analyses have identified a total of some 80,000 ERα-binding sites in the genomes of breast-cancer cells. However, there is only a relatively modest overlap between the sets of binding sites described by different studies. This indicates that, with the exception of a small set of core binding sites, plasticity is an attribute of ERα recruitment to chromatin (the complex of DNA and proteins). So, could clinicians use the ERα-binding profile of a breast tumour to predict the cancer's fate?

To address this question, Ross-Innes *et al.*[1] mapped the ERα binding sites in the genomes of tumour samples by sequencing DNA segments to which ERα was bound. The results clearly establish that different ERα-binding profiles in primary breast tumours can be

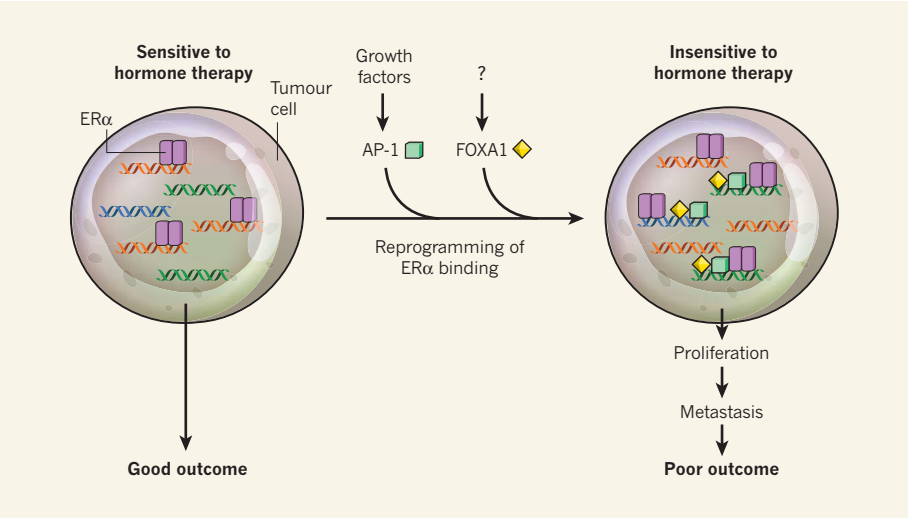


Figure 1 | Reprogramming of oestrogen-receptor activity in breast cancer. Tumour cells that express the oestrogen receptor ERα are sensitive to hormone-based therapy, leading to a good therapeutic outcome. Carroll and colleagues[1] show that ERα binding to DNA is dynamically regulated, and that different ERα-binding profiles are associated with different clinical outcomes. They propose that the transcription factor FOXA1 can mediate the reprogramming of ERα activity by facilitating the recruitment of ERα to different locations on DNA, leading to the development of resistance to hormonal therapy and a bad therapeutic outcome. The biological cues acting on FOXA1 to induce ERα reprogramming are unknown. Another transcription factor, AP-1, might also reprogram ERα activity in breast tumours. Signalling by growth factors has been linked to resistance to hormonal therapy[7] and is known to activate AP-1. Ross-Innes *et al.*[1] observe a comparable presence of sites for AP-1 and FOXA1 binding to DNA in ERα-bound DNA segments identified in tumours with poor outcome.

linked to distinct clinical outcomes. Thus, profiling of ERα-binding sites might contribute to a better molecular classification of breast tumours.

In the current study, the authors pre-classified breast tumours from 15 patients into two groups — good and poor prognosis — depending on the presence of certain proteins that are used as standard markers for responsiveness to hormonal therapy. On its own, this approach can be deceptive, because some tumours that express markers predicting high responsiveness show only a modest response to tamoxifen and therefore have a poor outcome. So, in their pre-classification, Ross-Innes *et al.*[1] also took into consideration information from long-term follow-up of the patients. However, to establish precise binding profiles that, by themselves, could be used to predict outcome without prior knowledge of other indicators, many more tumours from larger groups of diverse patients would have to be profiled.

What mechanism underlies the differences among breast tumours in ERα recruitment to genomic sites? Previous functional genomic studies[3,4] have shown that the interaction of ERα with chromatin largely depends on the presence of the protein FOXA1. This belongs to a class of transcription factors that facilitate protein binding to chromatin[5]. Recently, Carroll and colleagues[6] argued that FOXA1 is essential for ERα binding to chromatin, and for the receptor's subsequent transcriptional activity, not only in the presence of oestrogen but also in cells exposed to tamoxifen. Whether FOXA1 plays a part in the response of ERα-positive breast tumours to hormonal therapies in patients remained unclear, however.

In the absence of co-profiles of FOXA1 and ERα binding in the same breast-tumour samples, Ross-Innes *et al.*[1] addressed this question by searching within ERα-binding regions for known DNA sequence motifs recognized by different transcription factors. Unexpectedly, they found that the FOXA1-binding motif occurs on ERα-bound DNA segments obtained from tumours associated with a poor outcome but not on segments from tumours associated with a good outcome.

To validate this observation, the authors used cell-based models of breast cancer and again found preferential enrichment of the FOXA1-binding motif in tamoxifen-resistant cells and in cells that had been treated with a cocktail of agents promoting growth and proliferation. They conclude that in the tumours associated with poor outcome, FOXA1 could be triggering a rapid and dynamic reprogramming of ERα recruitment to chromatin at different sites from those in tumours associated with a good outcome (Fig. 1).

Whether FOXA1-dependent programming of ERα binding to DNA occurs during the course of the disease or is intrinsic to the primary tumour is yet to be determined. In either case, it would be of interest to identify the physiological cues and mechanisms dictating FOXA1 recruitment to chromatin. Identifying FOXA1-binding sites in breast-tumour samples would be an essential first step in assessing these issues.

On the basis of their previous work[6], Carroll and colleagues logically favour the hypothesis that FOXA1 has a key role in mediating reprogramming of ERα activity. It is likely, however, that additional mechanisms are involved, including the participation of other proteins. In particular, AP-1, a transcription factor that stimulates cell growth and proliferation, might also contribute to ERα reprogramming (Fig. 1). Indeed, the authors[1] find an enrichment of the binding motif for AP-1 in ERα-bound DNA segments from tumours associated with poor outcome. This observation nicely parallels previous results[7] showing enrichment of AP-1 motifs in ERα-bound segments upon treatment of breast cancer cells with epidermal growth factor, a condition that mimics the increase in growth-factor signalling in hormone-resistant tumours. The involvement of AP-1 in this process is also supported by its function in facilitating chromatin availability for the binding of the glucocorticoid receptor — another nuclear hormone receptor — to chromatin[8].

Over the past few years, functional genomics has enhanced our knowledge of hormone action. Ross-Innes and co-workers' technical prowess[1] should motivate similar studies of other nuclear receptors and their interacting protein partners involved in the initiation and progression of all subtypes of breast cancer[9–12]. Profiling of transcription factors in cancer-cell lines has proved extremely useful in deciphering the mechanisms underlying hormone-driven gene expression. But using primary tumours and metastatic tissues, as Ross-Innes *et al.* have done, will be required to reach relevant conclusions about the roles of each of these factors in cancer. ■

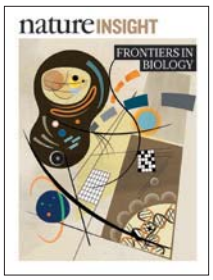
Geneviève Deblois *and* **Vincent Giguère**
are at the Goodman Cancer Research Centre and Departments of Biochemistry, Medicine and Oncology, McGill University, Montreal, Quebec H3A 1A3, Canada.
e-mail: vincent.giguere@mcgill.ca

1. Ross-Innes, C. S. *et al. Nature* **481,** 389–393 (2012).
2. Lupien, M. & Brown, M. *Endocr.-Relat. Cancer* **16,** 381–389 (2009).
3. Carroll, J. S. *et al. Cell* **122,** 33–43 (2005).
4. Laganière, J. *et al. Proc. Natl Acad. Sci. USA* **102,** 11651–11656 (2005).
5. Zaret, K. S. & Carroll, J. S. *Genes Dev.* **25,** 2227–2241 (2011).
6. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. *Nature Genet.* **43,** 27–33 (2011).
7. Lupien, M. *et al. Genes Dev.* **24,** 2219–2227 (2010).
8. Biddie, S. C. *et al. Mol. Cell* **43,** 145–155 (2011).
9. Robinson, J. L. *et al. EMBO J.* **30,** 3019–3027 (2011).
10. Ni, M. *et al. Cancer Cell* **20,** 119–131 (2011).
11. Deblois, G. *et al. Cancer Res.* **69,** 6149–6157 (2009).
12. Xu, J., Wu, R. C. & O'Malley, B. W. *Nature Rev. Cancer* **9,** 615–630 (2009).

nature INSIGHT

FRONTIERS IN BIOLOGY

19 January 2012 / Vol 481 / Issue No 7381



Cover illustration
by Nik Spencer

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

nature publishing group

The Nature Insight 'Frontiers in Biology' aims to cover timely and important developments in biology, ranging from the subcellular to the organismal level, and including molecular mechanisms and biomedicine. In this Insight the reviews discuss the role of innate immune signalling in tissue homeostasis and the response to infection, how understanding the DNA damage response has guided the development of inhibitors and helped to establish new principles for treating cancer, the potential therapeutic promise of patient-derived pluripotent stem cells, the forces that govern tumour evolution and the impact of mouse genetics on the study of bone physiology.

Richard Flavell and colleagues provide a timely overview of the activation and function of different inflammasomes. These are protein complexes that sense microbes and signs of tissue damage by responding to exogenous and endogenous signals. They can be beneficial and detrimental in a diverse set of inflammatory and metabolic diseases.

Christopher Lord and Alan Ashworth discuss how understanding the different response pathways to DNA damage has led to a multipronged approach to cancer therapy. Inherent mutations of the cancer cell and repair-enzyme inhibitors are used to synergistically target multiple repair pathways to enhance the cell killing effect.

Daisy Robinton and George Daley highlight the recent progress in the derivation of induced pluripotent stem cells, outline the functional assessments of pluripotency and discuss whether induced pluripotent stem cells equal embryonic stem cells, and thus may be a suitable alternative for research and therapy.

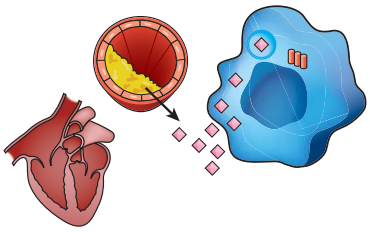
Mel Greaves and Carlo Maley consider our current knowledge of clonal evolution in cancer. Tumours and metastases evolve through the successive accumulation of genetic and epigenetic alterations and clonal selection within the adaptive environment. This leads to phenotypic and functional heterogeneity of cancer cells within a tumour. Although therapy can eliminate most cancer cells, it can drive the selection of resistant clones.

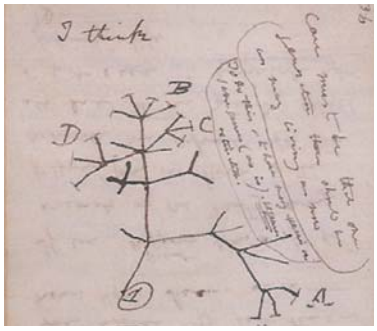
Finally, Gérard Karsenty and Mathieu Ferron discuss how recent evidence has changed our view of the skeleton and revealed an intimate connection between bone and the regulation of metabolism.

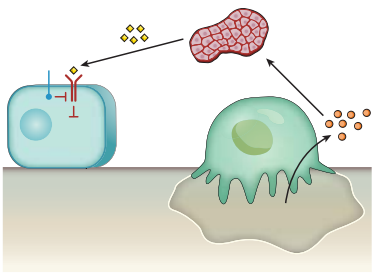
Alex Eccleston, Angela Eggleston, Marie-Thérèse Heemels, Barbara Marte & Ursula Weiss
Senior Editors

CONTENTS

REVIEWS







REVIEW

doi:10.1038/nature10759

Inflammasomes in health and disease

Till Strowig[1]*, Jorge Henao-Mejia[1]*, Eran Elinav[1]* & Richard Flavell[1,2]

Inflammasomes are a group of protein complexes built around several proteins, including NLRP3, NLRC4, AIM2 and NLRP6. Recognition of a diverse range of microbial, stress and damage signals by inflammasomes results in direct activation of caspase-1, which subsequently induces secretion of potent pro-inflammatory cytokines and a form of cell death called pyroptosis. Inflammasome-mediated processes are important during microbial infections and also in regulating both metabolic processes and mucosal immune responses. We review the functions of the different inflammasome complexes and discuss how aberrations in them are implicated in the pathogenesis of human diseases.

Inflammation is an acute response to infection and tissue damage to limit harm to the body[1]. However, dysregulated and chronic inflammation may result in secondary damage and immune pathology to the host. Inflammation is initiated on the sensing of signs of acute damage or disturbances of the steady state. Several recognition systems have evolved together to distinguish between homeostasis and threats to the host. Some of these receptors recognize distinct conserved pathogen-associated molecular patterns (PAMPs) that are predominantly found in microbes and hence allow the exquisite sensing of pathogens in tissues that are normally devoid of these structures[2]. Multiple additional receptors recognize host-derived signals, called damage-associated molecular patterns (DAMPs), that are released as a result of perturbations of tissue homeostasis caused by microbial or non-microbial insults, allowing a general sensing of stressed tissue[3].

Inflammasomes are a group of protein complexes that recognize a diverse set of inflammation-inducing stimuli that include PAMPs and DAMPs and that control the production of important pro-inflammatory cytokines such as interleukin-1β (IL-1β) and IL-18 (refs 4, 5). Furthermore, they have been found to regulate other important aspects of inflammation and tissue repair such as pyroptosis, a form of cell death. Recent studies have characterized distinct molecular activation mechanisms for several sensor proteins and have identified a multitude of ligands of both endogenous and exogenous origins. Moreover, the diverse functions of these complexes in antimicrobial responses, as well as in multifaceted diseases such as metabolic syndrome and inflammatory bowel disease (IBD), have started to be revealed. Importantly, mutations in components of inflammasome complexes have been associated with a propensity for the development of several immune-mediated diseases in humans. We review several of these conditions, discuss the different models that have been proposed for inflammasome involvement in normal and aberrant immune response, and highlight the challenges and future directions for this field.

Inflammasomes are intracellular innate immune sensors

Inflammasomes control the activity of caspase-1

IL-1β is one of the quintessential pro-inflammatory cytokines that broadly affects inflammatory processes[6]. Tight control of its production is therefore required at the transcriptional and post-translational levels. IL-1β is synthesized as a pro-protein without a typical signal sequence that would allow its secretion, and instead its activation and cellular release are controlled by the cysteine protease caspase-1 (ref. 7). Similarly, caspase-1 is responsible for the processing and secretion of IL-18, as well as the secretion of other proteins such as IL-1α and fibroblast growth factor-2 through an unconventional protein secretion pathway[8]. Moreover, caspase-1 is required for pyroptosis, a form of cell death frequently observed during microbial infections that combines characteristics of apoptosis (DNA fragmentation) and necrosis (inflammation and cytokine release)[9]. Like other caspases, caspase-1 is synthesized as an inactive zymogen (pro-caspase-1) and becomes proteolytically active only after controlled dimerization in inflammasomes that are built around one of several different molecules[4,5,10–18] (Fig. 1). Whereas the leucine-rich repeat (LRR) domain is thought to be involved in autoinhibition that is disabled on direct or indirect sensing of the activating signal, the nucleotide-binding domain (NBD) is involved in the regulation of homo-oligomerization or hetero-oligomerization, which is required for inflammasome assembly. On receiving an activating signal, inflammasome sensors recruit pro-caspase-1 (which has a caspase activation and recruitment domain (CARD)) either directly through homotypic binding of CARD or indirectly through a pyrin domain (PYD) by means of the adaptor apoptosis-associated speck-like protein containing a CARD (ASC), which contains a PYD and a CARD.

Diverse signals induce inflammasome formation

Inflammasome assembly is unique in its induction by a variety of both exogenous and endogenous signals. The range of activation signals sensed by each protein is distinct, but may include overlapping signals (Box 1). Whereas the AIM2 and NLRC4 inflammasomes are activated only by specific PAMPs, double-stranded DNA (dsDNA) and specific bacterial proteins, respectively[15–17], NLRP3 is activated by a large variety of signals, including PAMPs, DAMPs and bacterial toxins[19–22]. The structural diversity of ligands activating the NLRP3 inflammasome is in striking contrast to other innate pattern recognition receptors such as Toll-like receptors, which usually recognize more confined structural motifs[23]. Several models have therefore been proposed to explain how all of these signals could activate the NLRP3 inflammasome (Box 1). These non-exclusive mechanisms include both direct and indirect signal recognition mediated by additional accessory proteins[21,24,25]. We discuss the nature of some of these NLRP3-activating signals in the context of the pathophysiology of several diseases, but they have been more extensively reviewed elsewhere[4,5].

[1]Department of Immunobiology, Yale University, Connecticut 06520, USA. [2]Howard Hughes Medical Institute, Maryland 20815-6789, USA. *These authors contributed equally to this work.

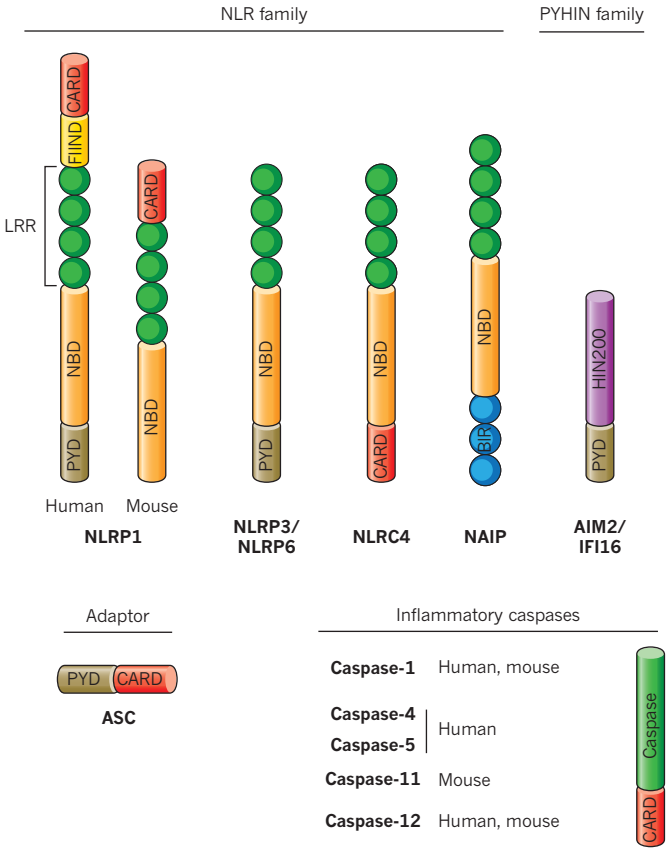


Figure 1 | Domain organization of inflammasome proteins. The identified core components belong to two families, the NOD-like receptor (NLR) family and the PYHIN (pyrin and HIN200 (haematopoietic interferon-inducible nuclear antigens with 200 amino-acid repeats) domain-containing protein) family. The NLR family members include NLRP1, NLRP2 (not shown), NLRP3, NLRP6, NLRC4 and, potentially, NLRP12. They all contain a nucleotide-binding domain (NBD), carboxy-terminal leucine-rich repeat (LRR), and can contain either a PYD or a caspase activation and recruitment domain (CARD) or both. The PYHIN family members AIM2 and IFI16 are characterized by having, in addition to a PYD, a HIN200 domain, which is involved in ligand binding. BIR, baculoviral inhibition of apoptosis repeat domain; FIIND, domain with function to find.

Regulation of inflammasome activity

Inflammasome activity needs to be tightly regulated by the host to avoid the excess production of cytokines or overt cell death. Regulation occurs at transcriptional and post-transcriptional levels. First, the expression of inflammasome sensors, in particular NLRP3, is relatively low in many cell types and requires a priming signal to be induced[26]. This signal is frequently referred to as signal I and can be provided by microbial ligands, cytokines or reactive oxygen species (ROS). In addition, alternative splicing of inflammasome components generates protein variants with different activities. Indeed, splice variants of ASC have been identified with distinct abilities to serve as an inflammasome adaptor, with one variant even having inhibitory activity[27]. Furthermore, the host expresses proteins that regulate inflammasome activity mainly by sequestering inflammasome components through homotypic interaction of CARDs or PYDs or through directly inhibiting caspase-1 function[28,29]. Similarly, subcellular location and trafficking of inflammasome components seems to be important for the regulation of inflammasome activity because ASC is found mainly in the nucleus in resting cells[30]. Another level of regulation may occur through the assembly of complexes with different components, suggested by the observation that cytokine processing downstream of NLRC4 activation is ASC dependent, whereas pyroptosis is independent of ASC[31]. In addition, it has been suggested that differential processing of caspase-1 might contribute to this finding because catalytically active caspase-1 that has been rendered unable to perform autocleavage is impaired in its ability to cleave cytokines, but it is still able to initiate pyroptosis on NLRC4-mediated recognition of *Salmonella* infection[32]. Regulation of inflammasome activity is also achieved through crosstalk with cellular stress-associated processes such as autophagy. Induction of autophagy leads to the degradation of cellular substrates such as protein aggregates and organelles in autolysosomes for the recycling of metabolites. Strikingly, cells deficient in autophagy have a decreased threshold for inflammasome activation[33]. This has been suggested to be a result of the impaired clearance of defective mitochondria resulting in elevated levels of ROS, hinting at an involvement of NLRP3 as sensor[34,35]. Another aspect of the regulation of inflammasome activity is its downregulation either through secreted factors or cell–cell interactions. Examples of these signals are type I interferons or interactions between $CD4^+$ T cells and macrophages or dendritic cells, respectively, leading to transcriptional and post-transcriptional downregulation of inflammasome activity[36,37]. All of these process probably cooperate in the temporal and spatial organization of inflammasome-mediated processes. In summary, the host has evolved distinct mechanisms to regulate inflammasome activation to prevent the dire consequences of inflammasome overactivation.

Regulation of inflammation by inflammasomes

Inflammasome-regulated processes depend on the simultaneous expression of the multiple inflammasome protein components in the same cell type of inflamed tissues. ASC and caspase-1 are found in many tissues and cell types, whereas the inflammasome sensors feature a more distinct expression pattern, suggesting tissue-specific mechanisms for sensing the microenvironment. In the following sections, we discuss how activation of inflammasomes in different lineages of cells regulates physiological reactions in the context of health and disease.

Inflammasomes and the antimicrobial response

In vivo, inflammasomes have been shown to participate in the antimicrobial innate immune response[38]. The most widely studied inflammasome in this regard is the NLRP3 inflammasome, shown to be involved in antibacterial, viral, fungal and parasitic immune responses. Despite the evidence linking the NLRP3 inflammasome to the immune response to infection, only in a minority of cases has inflammasome activation by direct recognition of the pathogen been documented; many studies have indicated inflammasome activation through induction of signals related to cellular stress and damage[4] (see Box 1). The influenza A virus is an example of an indirect viral NLRP3 inflammasome activator. On infection, recognition of viral RNA by means of Toll-like receptor 7 (TLR7) induces transcription of the NLRP3 inflammasome components[39]. Subsequently, the activity of the viral ion-channel protein M2 induces pH neutralization of the *trans*-Golgi network, leading to potassium efflux and ROS formation, which in turn induce NLRP3 inflammasome assembly. Recently, messenger RNA of microbial origin was shown to activate the NLRP3 inflammasome in a TRIF (TIR-domain-containing adaptor-inducing interferon-β)-dependent manner, providing a mechanism for NLRP3 recognition of infection and the resultant differentiation by the host between viable and non-viable bacteria, leading to the induction of a potent immune response only on exposure to live microbes[40]. Mouse NLRP1 inflammasome activation is central to the initiation of the antimicrobial response to *Bacillus anthracis* infection through caspase-1-induced pyroptosis of infected macrophages, which permits self-limitation of infection and initiation of an antimicrobial neutrophilic reaction[41]. Human NLRP1 forms an ASC-dependent inflammasome, whereas mouse NLRP1b may activate caspase-1 in an ASC-independent manner[42]. NLRP1, like NOD2 and NLRP3, can sense muramyl dipeptide (MDP), a building block of the bacterial cell wall, suggesting possible cooperative roles of the different pattern recognition receptors in the sensing of and

response to bacteria[43]. NLRC4 inflammasome activation is driven by type III and type IV secretion systems (T3SS and T4SS) of bacteria such as *Salmonella*, *Pseudomonas*, *Legionella* and *Yersinia*, which allow the cytoplasmic entry of the NLRC4 ligand flagellin, leading to activation of the NLRC4 pathway[31,44]. In addition, flagellin-independent activation of the NLRC4 inflammasome involves recognition of the T3SS rod protein PrgJ[45]. Interestingly, NLRC4 activation during *Legionella* infection is dependent on another NLR protein, NAIP5, but only partial dependence on NAIP5 was demonstrated for NLRC4 activation during *Pseudomonas* and *Salmonella* infection[45–47]. This apparent discrepancy was recently resolved by demonstrating that oligomerization of different NAIP family proteins with NLRC4 confers ligand specificity, so that a NAIP2–NLRC4 complex binds PrgJ and other structurally related rod proteins, whereas a NAIP5–NLRC4 complex binds flagellin[48]. Different pathogen receptors therefore assist NLRC4 in the recognition of microbial ligands, broadening the diversity of sensed structures. The recently discovered protein AIM2 recognizes bacterial and viral dsDNA, resulting in an antimicrobial response to intracellular pathogens, such as *Francisella tularensis*, *Listeria monocytogenes* and some DNA viruses[49,50].

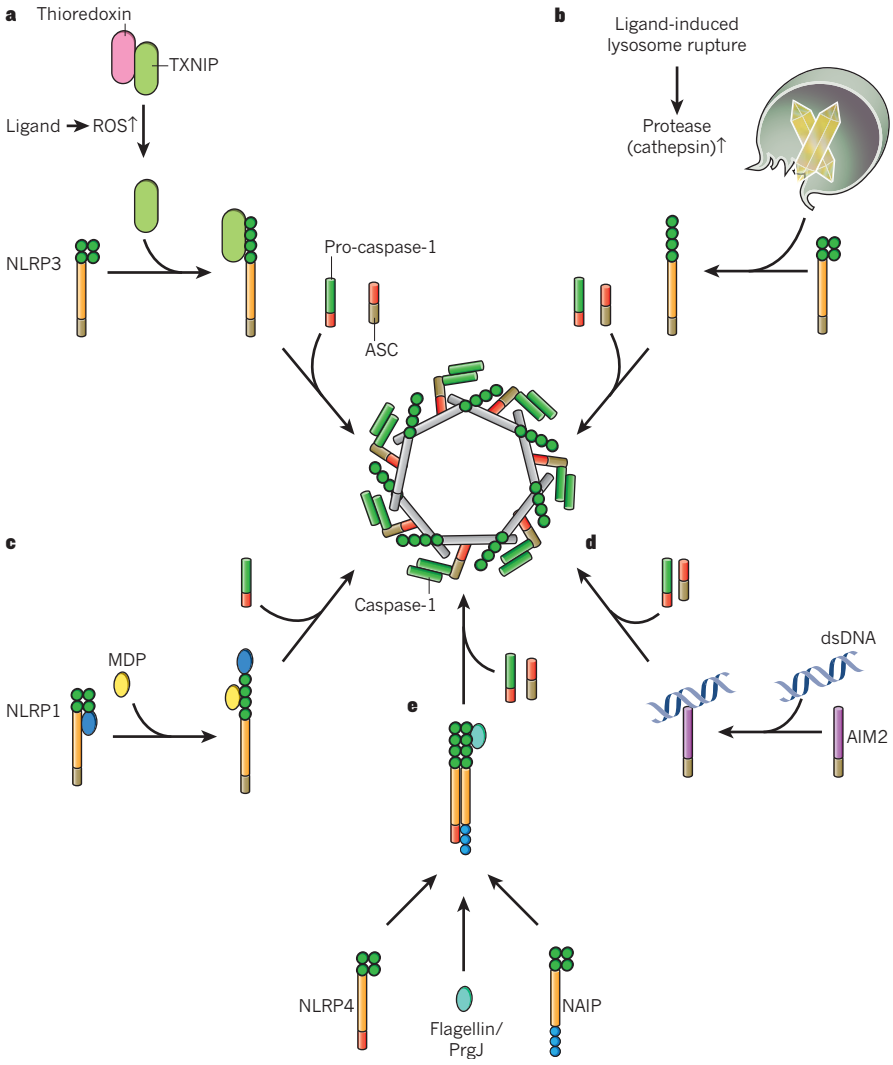
During microbial infection, distinct effector mechanisms of inflammasomes seem to be important for host defence. Although inflammasome-induced IL-1β and IL-18 are essential for the clearance of influenza virus and *Shigella*, respectively, pyroptosis has been suggested to be essential for caspase-1-mediated effects *in vivo* during infection by *Salmonella enterica* Typhimurium, *Legionella pneumophila* or *Burkholderia thailandensis*[51]. Altogether, it seems that the cooperative activity of distinct inflammasomes is pivotal in the host's ability to raise a protective immune response. This cooperation is illustrated during infection by *S. Typhimurium*, in which deficiencies in either NLRP3 or NLRC4 *in vivo* do not lead to increased bacterial infection, whereas mice deficient in both NLRC4 and NLRP3 have a similarly increased susceptibility to infection to that of caspase-1-deficient mice[52] (Fig. 2).

To avoid inducing inflammasome activation, many pathogens encode inhibitors of inflammasome function[53]. Frequently, these inhibitors, such as NS1 of influenza virus or Crm1 of cowpox virus, act on central inflammasome components and directly inhibit caspase-1 activity[54,55]. Other inhibitors, such as ORF63 of Kaposi's sarcoma-associated herpesvirus, bind NLR sensors and prevent the formation of NLRP1 and NLRP3 inflammasomes[56]. In general, the inhibition of inflammasome activity by the immune-evasive mechanisms of many microbes highlights the importance of these complexes in

BOX 1

Models for inflammasome activation

Inflammasomes are assembled after sensing a structurally diverse repertoire of PAMPs and DAMPs. Several models have been proposed to explain how these signals are sensed, including models based on recognition of general cellular stress (Fig. **a** and **b**) or on direct and indirect recognition of activation signals (Fig. **c**–**e**). NLRP3 senses the reactive oxygen species (ROS), which is produced in the cell (potentially by mitochondria) directly or indirectly by activators of the NLRP3 inflammasome. Increased amounts of ROS are sensed by a complex of thioredoxin and thioredoxin-interacting protein (TXNIP), leading to the dissociation of this complex. Subsequent binding of TXNIP to NLRP3 leads to the activation of NLRP3, the recruitment of ASC and pro-caspase-1, and formation of the active inflammasome complex (**a**). NLRP3 is activated after lysosome destabilization. The phagocytosis of specific crystalline and particulate structures can lead to lysosome destabilization and the release of lysosomal content, including proteases. These proteases could lead to proteolytic inactivation of a negative regulator or to proteolytic activation of a positive regulator of NLRP3, resulting in inflammasome assembly (**b**). NLRP1 and AIM2 sense the ligand directly. The direct binding of specific ligands (muramyl dipeptide (MDP) and double-stranded DNA (dsDNA)) can lead to conformational changes in NLRP1 and AIM2, resulting in inflammasome activation. Inflammasome formation in NLRP1 is independent of ASC (**c, d**). NAIP proteins sense bacterial proteins resulting in the recruitment of NLRC4 and assembly of the NLRC4 inflammasome (**e**).



antimicrobial immunity. Over the next few years, it will be important to further characterize how pathogens are sensed by inflammasomes and how inflammasomes interact with additional innate immune pathways to regulate antimicrobial immune responses *in vivo*.

Inflammasome activation by particulate compounds

In addition to infectious agents, there is evidence in mice and in humans that aberrant inflammasome activation by non-infectious agents may be linked to the pathogenesis of diseases characterized by sterile inflammation. For instance, crystal deposition diseases are a group of disorders in which inflammatory damage is elicited by exposure to exogenous or endogenous crystalline molecules. Examples of exogenous activators are silica and asbestos, whose endocytosis by pulmonary macrophages results in NLRP3 inflammasome activation involving ROS and lysosome destabilization, leading in turn to silicosis and asbestosis, respectively[24,57]. Similarly, aberrant formation of crystals from endogenous molecules such as monosodium urate (MSU) and calcium phosphate may lead to NLRP3 inflammasome activation in macrophages. Indeed, particular accumulation of MSU had long been noted during gout, but only recently has it been demonstrated that NLRP3-deficient mice feature defective MSU-induced neutrophil infiltration and inflammation[22] (Fig. 2). On the basis of these findings, clinical trials have been started to block IL-1β in gout; promising early results suggest that aberrant inflammasome activation drives inflammation in human crystal deposition disease. Furthermore, in patients with osteoarthritis, a common degenerative articular disorder, uric acid levels in synovial fluids are positively correlated with both IL-1β and IL-18 levels, as well as disease severity[58]. Calcium phosphate crystals were recently shown to activate NLRP3 (ref. 59). Hydroxyapatite crystals, a component of bone, are frequently found in osteoarthritis synovial fluid, activate IL-1β production by means of the NLRP3 inflammasome, and mediate inflammation and joint disease[59]. Future studies and clinical trials will be needed to explore whether targeting inflammasomes and their substrates are effective treatment options for these diseases.

An interesting example of inflammation that can be induced by an exogenous crystalline compound is the activation of the NLRP3 inflammasome by alum, an aluminium salt and the most commonly used adjuvant in human vaccines. Exposure of macrophages to alum *in vitro* leads to the NLRP3-dependent activation of caspase-1, and several studies have demonstrated that NLRP3 inflammasome-deficient mice have defects in alum-induced adaptive immune responses *in vivo*[60–62]. Several other studies have debated the role of the NLRP3 inflammasome in the adjuvant effect of alum, some suggesting that vaccination-induced adaptive immune responses may be ASC dependent but independent of caspase-1 (and hence independent of inflammasomes)[63–66]. A recent study suggested an alternative NLRP3-independent mechanism for alum's adjuvant effects, namely attachment to the plasma membrane of dendritic cells, leading to abortive phagocytosis and imprinting dendritic cells to induce a humoral immune response[67]. The precise mechanisms related to vaccination adjuvant effects and the possible contribution by inflammasomes remain to be clarified in future studies.

Inflammasomes and metabolic syndrome

The incidence of obesity worldwide has increased markedly in recent decades. Obesity is associated with multiorgan (namely pancreatic, adipose, hepatic, cardiac and muscle tissue) chronic metabolic and inflammatory alterations that together are termed 'metabolic syndrome'. These alterations include impaired insulin sensitivity, pancreatic β-cell dysfunction, non-alcoholic fatty liver disease and atherosclerosis. Unequivocal experimental and clinical evidence causally link IL-1β and IL-18 to the development of these metabolic pathologies and their complications[68,69]. However, the underlying molecular mechanisms that result in tissue-specific inflammasome activation in the context of obesity have only just begun to be explained.

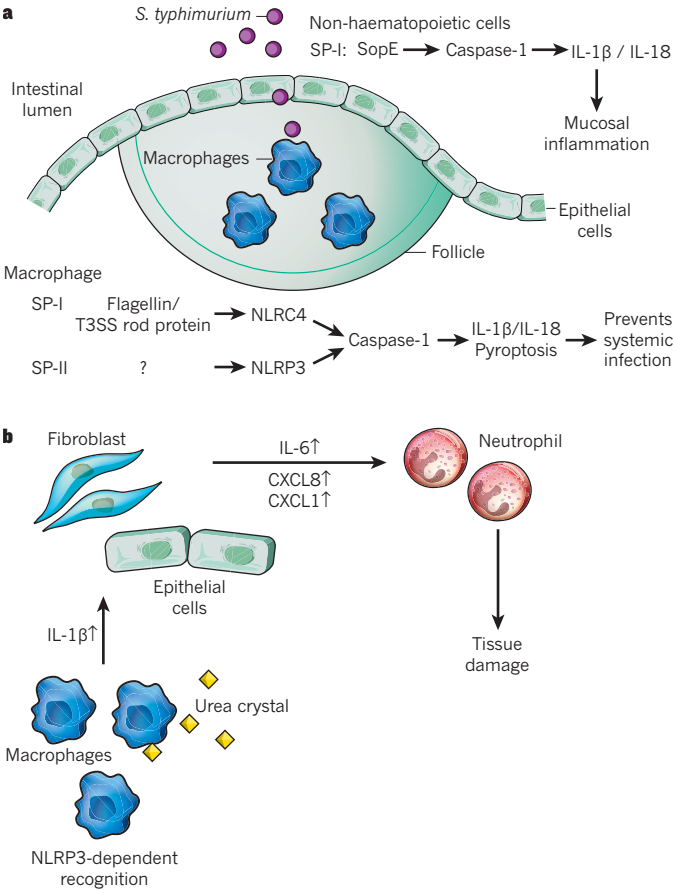


Figure 2 | Inflammasome activity regulates inflammation during microbial infection and autoinflammatory diseases. **a**, *S. typhimurium* enters the host by crossing through the intestinal epithelial barrier. M cells, a specialized epithelial cell type layering Peyer's patches, are particularly involved in the transcytosis of *Salmonella* and the infection of macrophages in Peyer's patches. The inflammasome and caspase-1 are involved in several cell types and at several steps of the infection. Injection of the bacterial effector protein SopE into epithelial cells induces the activation of caspase-1 independently of NLRP3 and NLRC4 through a process involving the GTPase Rac1. The resultant mucosal inflammation is dependent on IL-1β and IL-18 produced by non-haematopoietic cells. On infection of macrophages, the bacterial proteins flagellin and PrgJ (part of the T3SS) are sensed by means of NLRC4. This results in the activation of caspase-1, leading to both IL-1β/IL-18 processing and pyroptosis, which limit systemic infection. NLRP3 contributes to these processes by recognition of an unknown signal. **b**, Phagocytosis of monosodium urate (MSU) crystals by macrophages induces the NLRP3-dependent caspase-1 activation and release of IL-1β, which stimulates non-haematopoietic cells to produce IL-6 and chemokines (CXCL1 and CXCL8), attracting neutrophils. Activated neutrophils then cause tissue damage. Therapeutic blockade of IL-1β in humans ameliorates inflammatory bouts in gout.

Inflammasome activation regulates insulin signalling

NLRP3 inflammasome components and caspase-1 activation are increased in the adipose tissue and liver of obese mice and humans; moreover, their level of expression is directly correlated with the severity of type 2 diabetes mellitus (T2DM) in obese individuals[70]. During nutritional surplus, in addition to adipocyte hypertrophy owing to increased lipid storage, adipose tissue is infiltrated by classically activated, M1, macrophages that secrete pro-inflammatory cytokines[71]. NLRP3, ASC and caspase-1 are preferentially expressed in adipose-tissue-infiltrating macrophages, in which the saturated fatty acid palmitate and lipotoxic ceramides trigger NLRP3 inflammasome activation through a mechanism that involves defective autophagy and the accumulation of mitochondrial ROS[70,72] (Fig. 3). Enhanced

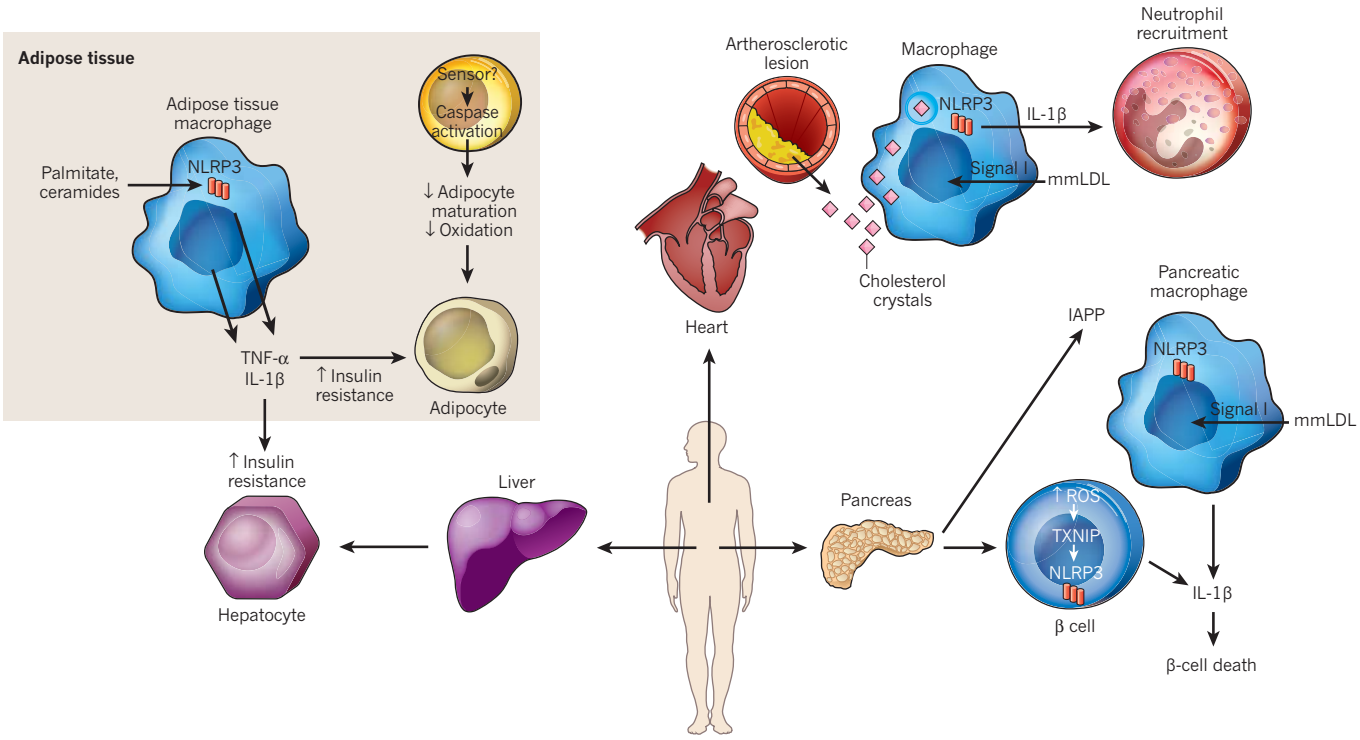


Figure 3 | The role of inflammasomes in metabolic syndrome. During obesity, the NLRP3 inflammasome is activated by obesity-associated DAMPs in multiple tissues and cell types; the resultant pro-inflammatory-induced state often leads to a deterioration in metabolic functions. In adipose tissue, palmitate and ceramides activate the NLRP3 inflammasome in infiltrating macrophages, which leads to an enhancement of insulin resistance. In addition, caspase-1 activation through an unknown sensor protein regulates adipocyte differentiation and fatty acid oxidation. In the pancreas, IAPP and increased mitochondrial ROS production activate the NLRP3 inflammasome in mmLDL-primed macrophages and β cells, respectively. The increased levels of IL-1β in pancreatic islets result in increased β-cell death and decreased insulin production. Minute cholesterol crystals in early atherosclerotic lesions activate the NLRP3 inflammasome in mmLDL-primed macrophages, promoting inflammatory-cell infiltration and increased atherosclerosis progression.

NLRP3-mediated caspase-1 activation in adipose-tissue-infiltrating macrophages of obese mice results in at least two deleterious consequences in insulin-sensitive tissues: first, IL-1β inhibits insulin signalling by direct serine phosphorylation of insulin receptor substrate 1 and induces the expression of tumour necrosis-factor-α, a well-characterized insulin-resistance-promoting cytokine; and second, IL-1β and IL-18 induce type 1 CD4 T helper cells in adipose tissue[70].

Inflammasomes also seem to have adipocyte-intrinsic functions. Caspase-1 is upregulated during adipocyte differentiation, in which it promotes insulin resistance through autocrine IL-1β-mediated effects; however, perhaps more interestingly, caspase-1-deficient precursors differentiate more efficiently into mature adipocytes and have an increased oxidation rate[73] (Fig. 3), which suggests that adipocyte-specific proteins may serve as caspase-1 substrates, leading to functional changes in adipose tissue of obese animals. Hence, *Nlrp3*$^{-/-}$, *Asc*$^{-/-}$ and *Casp1*$^{-/-}$ mice have been reported to have decreased weight gain and fat mass, as well as decreased insulin resistance. Interestingly, the loss of function of NLRP3 decreases but does not eliminate caspase-1 activation in adipose tissue or liver, suggesting that additional inflammasome sensors might contribute to the pathophysiology of obesity; furthermore, the nature of the stimuli that trigger caspase-1 activation during the accumulation of fat in adipocytes remains to be determined.

Inflammasome activation impairs β-cell function

Chronic hyperglycaemia as a result of peripheral insulin resistance is compensated for by increased insulin output by pancreatic β cells. Local inflammatory processes coupled with the toxic effects of glucose lead to accelerated mass loss of β cells and decreased insulin secretion over time, which prompts the progression from obesity and insulin resistance to overt T2DM. IL-1β, preferentially expressed by pancreatic infiltrating macrophages and to a smaller extent by β cells, has been implicated as a critical driver of β-cell death in conditions of chronic exposure to elevated concentrations of glucose[74]. Additional support for a pathological role of inflammasomes in T2DM comes from a recent report showing that glyburide, an insulin-secretion-promoting drug, suppresses NLRP3-mediated IL-1β release[75], and more importantly from human clinical trials of IL-1 receptor antagonist (IL-1RA) to treat T2DM that demonstrated improved glycaemic control and β-cell function[76].

In pancreatic islet failure during T2DM, NLRP3 inflammasomes may be activated by different mechanisms. First, hyperglycaemia triggers mitochondrial ROS production in β cells by increasing the activity of the electron transport chain, which induces thioredoxin-interacting protein (TXNIP) to dissociate from thioredoxin, leading to the activation of NLRP3 (ref. 25) (Fig. 3 and Box 1). Second, the severity of T2DM is closely correlated with the levels of deposition of amyloid polypeptide (IAPP, also known as amylin) in pancreatic islets[77]. Remarkably, IAPP has been shown to specifically activate the NLRP3 inflammasome in mouse pancreatic macrophages previously primed with the abundant minimally oxidized low-density lipoprotein (mmLDL) through a pathway that involves disruption of the phagolysosomal pathway[78] (Fig. 3). Further evidence for the role of IAPP and NLRP3 activation in pancreatic islet deterioration is provided by the observation that macrophages express more IL-1β in pancreatic islets of human *IAPP*-transgenic mice[78]. Overall, secreted IL-1β can then signal in an autocrine or paracrine manner to induce β-cell death or dysfunction and to promote the expression of chemotactic factors that further worsen immune-cell infiltration. It is important to mention that TXNIP is not required for NLRP3 activation in macrophages, strengthening the notion that the cellular milieu is a critical determinant of the way in which the NLRP3 inflammasome senses damage signals in the

context of metabolic abnormalities[78]. However, despite great advances in recent years, several questions remain unresolved, including the mechanisms of pro-IL-1β induction in β cells, the exact contribution of inflammasome activation in infiltrating macrophages and β cells to disease progression, and the potential roles of additional inflammasome sensors in pancreatic inflammatory processes.

Inflammasome activation and atherosclerosis

The inflammatory nature of atherosclerosis is well established, but the biological agents that trigger artery wall inflammation remain largely unknown. Cholesterol crystal deposition in arterial vessels has long been a pathognomonic feature of atherosclerosis; moreover, recent evidence suggests that they are present at early stages of atherosclerotic lesions, coinciding with the first appearance of inflammatory cells[79]. Similarly to IAPP, cholesterol crystals activate the NLRP3 inflammasome through phagolysosome destabilization in mmLDL-primed mouse and human macrophages[79] (Fig. 3). LDL-receptor-deficient mice (prone to atherosclerosis) reconstituted with bone marrow deficient in NLRP3, ASC or IL-1α/β are markedly resistant to the development of atherosclerosis, suggesting that NLRP3 inflammasome activation and IL-1 secretion from the haematopoietic compartment are key events in the early stages of disease[79]. Another model for atherosclerosis is the *Apoe*$^{-/-}$ mouse, which develops severe hypercholesterolaemia and spontaneous atherosclerosis when being fed with a high-fat diet[80]. Earlier reports have demonstrated that IL-1-receptor-deficient *Apoe*$^{-/-}$ mice and IL-1RA-treated *Apoe*$^{-/-}$ mice showed decreased atherosclerosis, suggesting a role for the inflammasome in atherogenesis in this mouse model[81,82]. Interestingly, using mice deficient in both ApoE and different components of the NLRP3 inflammasome, Menu *et al.*[83] showed that atherosclerosis progresses independently of the NLRP3 inflammasome in this context. The most reasonable explanations for the discrepancies between the two studies lie in the differences between atherosclerosis models (*Apoe*$^{-/-}$ versus *Ldlr*$^{-/-}$) and in a putative role for IL-1α in the *Apoe*$^{-/-}$ mouse model. However, it also raises the interesting possibility that environmental factors could account for the contradictory phenotypes. Clearly, further research is warranted regarding the implication of the inflammasomes in the pathogenesis of atherosclerosis and their role in non-haematopoietic cells during atherosclerotic plaque formation.

Inflammasomes and the mucosal immune response

In the gastrointestinal mucosa, the host is separated from an immense microbial ecosystem by only a single layer of epithelial cells[84]. To avoid ongoing inflammatory reactions to commensal microbes and food antigens, while preserving the ability to react to pathogenic insults, mammals have evolved a complex mucosal immune system composed of epithelial and stromal cells acting together with subsets of cells of haematopoietic origin. These cells interact closely with each other and with the surrounding microbial milieu. When such interactions are perturbed, autoinflammation may develop, potentially leading to IBD. The ability of inflammasomes to recognize exogenous and endogenous signals has led to several studies characterizing their role during chemical-induced intestinal autoinflammation, a model for human IBD. Interestingly, conflicting results have been reported by several groups (Fig. 4). Using a common acute and chronic epithelial injury colitis model based on the administration of dextran sulphate to mice, several groups reported decreased disease severity in mice deficient in caspase-1 or NLRP3, which correlated with lower IL-1β production during disease[85,86]. Using the same model, other groups found that mice deficient in NLRP3, ASC and caspase-1 show an exacerbated disease severity[87–90]. In these reports, a role was suggested for the NLRP3 inflammasome in the promotion of tissue regeneration in response to injury. Although variations in these results could be explained by differences in experimental designs, a recent paper[91] has offered an alternative explanation. This study reported that an NLRP6 inflammasome participates in the steady-state regulation of the commensal microflora. Deficiency in this inflammasome is associated with an alteration in the microflora communities and the emergence of normally suppressed bacteria that have pro-inflammatory activity. It was suggested that regulation of the microflora by epithelial inflammasomes is mediated, at least in part, through the induction of basal secretion of IL-18 by epithelial cells. Discrepancies in the earlier studies may therefore be explained by differences in microflora communities in different animal facilities, coupled with the inability of NLRP6-inflammasome-deficient mice (*Nlrp6*$^{-/-}$, *Asc*$^{-/-}$, *Casp1*$^{-/-}$) to regulate the microflora in a given facility.

Phenotypic alterations may be induced by differences in animal housing; wild-type mice housed with NLRP6-inflammasome-deficient mice acquire their colitogenic microflora and develop phenotypes that may be profoundly different from that of wild-type mice housed alone. The factors inducing the formation of the NLRP6 inflammasome and the precise effector mechanisms for regulation of the microflora remain to be studied. In addition to NLRP6, other inflammasomes that are expressed mainly within the haematopoietic compartment, such as the NLRP3 inflammasome, may also function by regulating the microflora, as well as the autoinflammatory process itself. In the mucosal environment, inflammasomes therefore function in the sensing of pathogens and the commensal flora by non-haematopoietic cells, such as the epithelial layers, but also by haematopoietic cells. They cooperate to maintain tolerance towards commensal microbes and to initiate a potent immune response towards pathogens (Fig. 4). Distinct inflammasomes expressed in different cell lineages may orchestrate these seemingly opposite functions during acute mucosal inflammation.

It has been suggested that inflammasomes participate in inflammation-induced tumorigenesis, a common complication of chronic intestinal autoinflammation (such as IBD), yet mechanisms accounting for this regulatory effect remain elusive[87,89]. Two recent reports suggest that NLRP6 deficiency may be associated with enhanced inflammation-induced tumorigenesis through impaired mucosal self-renewal and proliferation mediated by alterations in the intestinal stem-cell niche, or alternatively by an impaired haematopoietic immune response[92,93]. It remains to be determined whether the effect of NLRP6 on tumorigenesis is a direct consequence of NLRP6 deficiency in one or more intestinal cell types or whether it is indirectly induced by the documented alterations in the regulation of microflora composition of these mice.

Alternatively, another report suggests that epithelial NLRC4 inflammasome activity may also participate in the prevention of intestinal inflammation-induced tumorigenesis, because NLRC4-deficient mice were more prone to the development of colonic neoplasms after the induction of chronic inflammation[94]. This effect was independent of inflammation or the composition of the intestinal microflora and was suggested to involve enhanced proliferation of epithelial cells and impaired apoptosis of colonic epithelial cells. The precise molecular mechanism linking the NLRC4 inflammasome dysfunction to increased tumorigenesis remains to be explored.

Future perspectives

The discovery of inflammasomes as protein platforms that control the processing of IL-1β and IL-18 constitutes a milestone in the innate immunology field. However, many questions remain unanswered, including those regarding the biochemical and genetic regulation of these protein complexes, as well as their roles in complex diseases. These, in our opinion, may be the focus of inflammasome research in coming years.

The mechanism through which sensor proteins recognize activating signals remains a critical unresolved issue in the field. Although *in vitro* reconstitution assays with purified components have been performed with NLRP1 and AIM2 and have shown direct binding of the respective ligands to both sensors, intense research has not been successful in identifying such direct interactions with any of the other inflammasome sensors[15–17,43]. Hence, it is plausible that as yet unidentified proteins may facilitate the recognition of structurally diverse activating

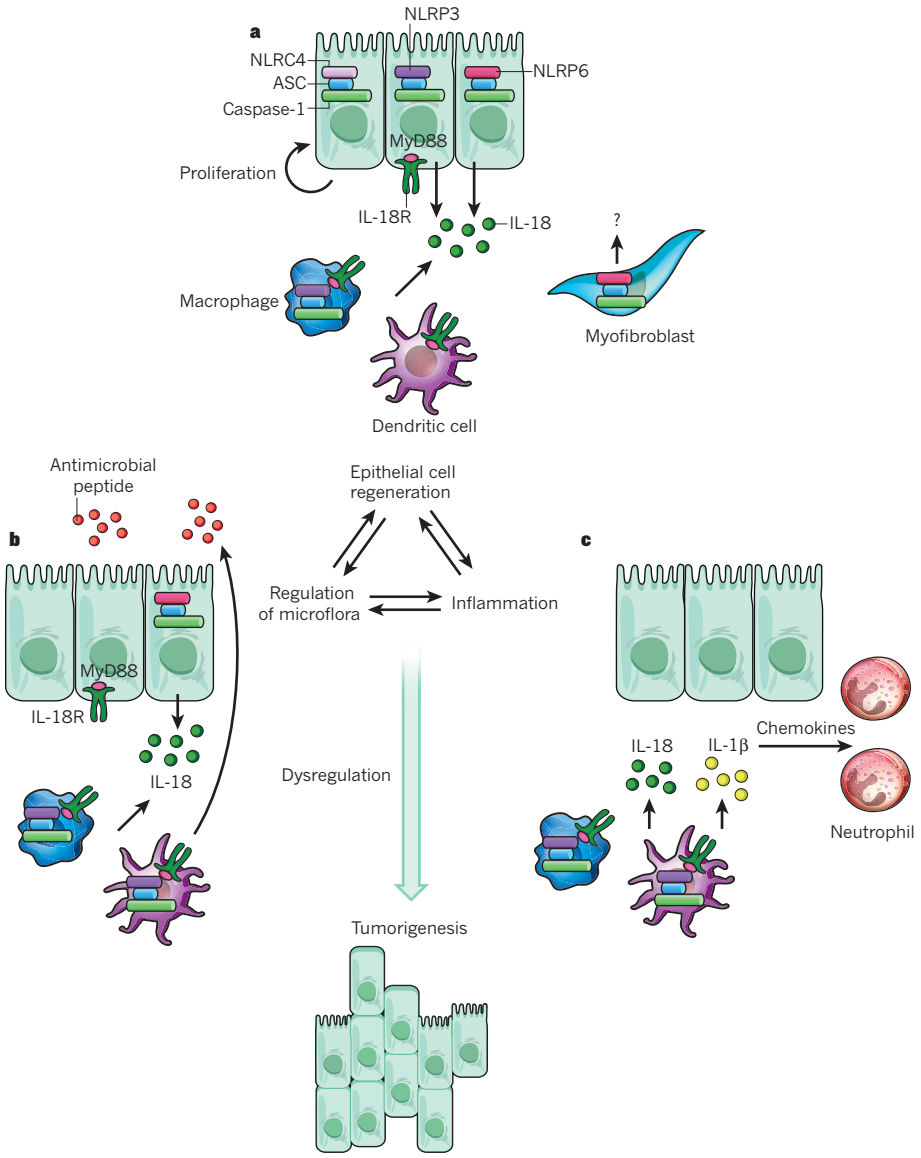


Figure 4 | Inflammasomes regulate multiple aspects of tissue homeostasis and immune response in the mucosa. NLRP3, NLRP6 and NLRC4 inflammasomes are formed in many lineages of haematopoietic and non-haematopoietic cells present in the intestine. The activity of these inflammasomes regulates epithelial cell regeneration, the microflora and mucosal inflammation. All of these processes are interconnected, and the dysregulation of any of them may increase the risk of tumorigenesis. **a**, Epithelial cell regeneration: recognition of as yet unknown signals leads to the processing of pro-IL-18 and potentially unidentified effectors through NLRP3, NLRP6 and NLRC4. This allows the regulation of epithelial regeneration on injury, through effects on epithelial cells, lamina propria immune cells, myofibroblasts and intestinal epithelial stem cells. **b**, Regulation of microflora: during the steady state, the production of IL-18 through NLRP3 and NLRP6 in the colon participates in the regulation of a 'healthy' microflora, potentially through regulation of antimicrobial peptide secretion by epithelial cells and immune cell subsets. Disruption of the NLRP6 inflammasome results in aberrant regulation of the intestinal microflora, leading to the emergence of pro-colitogenic bacteria. **c**, Inflammation: after breaching of the intestinal epithelial barrier, the NLRP3 inflammasome is involved in the regulation of intestinal inflammation in response to microbial invasion. Both IL-1β and IL-18 are suggested to act as effector molecules, leading to the recruitment of multiple cell subsets, including neutrophils, that promote local inflammation.

signals and ligands by NLRP3. Such 'bridging' mechanisms have been suggested to link cellular stress (namely potassium efflux, generation of ROS and lysosome destabilization) and microbial proteins to NLRP3 (refs 21, 24, 25). However, the exact identity of these bridging molecules remains to be fully characterized. Indeed, for the NLRC4 inflammasomes, NAIP proteins were identified as additional components that are responsible for the specificity of microbial ligands recognized by the complexes[48]. Another open issue is the interesting observation that several phenotypes related to a deficiency of common inflammasome effectors and adaptors cannot be attributed to the well-known NLR and PYHIN molecular sensors and their ligands, indicating the existence of additional sensors and ligands to be identified. An example of a recently identified sensor is IFI16, a sensor of viral DNA[18].

The early discovery that mutations in inflammasome components in humans are the cause of several inherited autoinflammatory diseases that can be treated by neutralizing IL-1β illustrates the potential detrimental effects of overactivation of inflammasomes[95]. However, recent evidence from mouse studies on the role of inflammasomes in mucosal immunology indicates a compartmentalized and dual function (detrimental versus beneficial). To dissect the contributions of inflammasomes in different cell types and tissues, novel tools such as conditional gene-knockout mice will have to be developed. This would allow better investigation of areas that bear fundamental importance and present complex challenges to the research community, such as the involvement of inflammasome signalling in circuits regulating metabolism, inflammation and the intestinal microflora in mice and humans[96].

It is worth noting that NLRP6 inflammasome deficiency was recently shown to associate with expansion of an autoinflammation-promoting microbiota including bacterial taxa such as Prevotellaceae and TM7 (ref. 91). The mechanisms by which inflammasomes sense microflora members or communities, regulate tissue repair and regeneration, and orchestrate mucosal immune responses during the steady state and inflammation remain to be fully characterized. Interestingly, both inflammasome deficiency and alterations in the intestinal microflora were previously linked in separate studies to a propensity for the development of elements of the metabolic syndrome, such as obesity and atherosclerosis, in mice and humans[97,98]. Indeed, recent characterization of the microflora community in a cohort of individuals of European origin identified the bacterial family Prevotellaceae as a prominent component of one human microflora cluster that featured a trend towards increased body mass index[99]. Moreover, faecal contents of rural African children featured expanded representation of Prevotellaceae that was absent in their European counterparts, probably reflecting a combination of genetic and environmental factors that differentially regulate the microflora composition[100]. Whether the regulation of the gut flora by inflammasomes affects weight, metabolism and inflammation is expected to become a major interest of the field.

Whereas the first decade of inflammasome research has been characterized mostly by *in vitro* studies and *in vivo* studies with small animal models, the contribution of altered inflammasome function to complex human disease is expected to gradually take centre stage. This, together with a vast effort to discover and develop small-molecule approaches to the specific inhibition of inflammasome activation, may pave the way to therapeutic intervention targeting inflammasome-regulated pathways that are involved in the pathogenesis of human disease. ■

1. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454,** 428–435 (2008).
2. Medzhitov, R. & Janeway, C. A. Jr. Innate immunity: the virtues of a nonclonal system of recognition. *Cell* **91,** 295–298 (1997).
3. Matzinger, P. Tolerance, danger, and the extended family. *Annu. Rev. Immunol.* **12,** 991–1045 (1994).
4. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140,** 821–832 (2010).
5. Davis, B. K., Wen, H. & Ting, J. P. The inflammasome NLRs in immunity, inflammation, and associated diseases. *Annu. Rev. Immunol.* **29,** 707–735 (2011).
6. Dinarello, C. A. IL-1: discoveries, controversies and future directions. *Eur. J. Immunol.* **40,** 599–606 (2010).
7. Cerretti, D. P. *et al.* Molecular cloning of the interleukin-1β converting enzyme. *Science* **256,** 97–100 (1992).
8. Keller, M., Ruegg, A., Werner, S. & Beer, H. D. Active caspase-1 is a regulator of unconventional protein secretion. *Cell* **132,** 818–831 (2008).
9. Bergsbaken, T., Fink, S. L. & Cookson, B. T. Pyroptosis: host cell death and inflammation. *Nature Rev. Microbiol.* **7,** 99–109 (2009).
10. Agostini, L. *et al.* NALP3 forms an IL-1β-processing inflammasome with increased activity in Muckle–Wells autoinflammatory disorder. *Immunity* **20,** 319–325 (2004).
11. Martinon, F., Burns, K. & Tschopp, J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-β. *Mol. Cell* **10,** 417–426 (2002).
12. Grenier, J. M. *et al.* Functional screening of five PYPAF family members identifies PYPAF5 as a novel regulator of NF-κB and caspase-1. *FEBS Lett.* **530,** 73–78 (2002).
13. Wang, L. *et al.* PYPAF7, a novel PYRIN-containing Apaf1-like protein that regulates activation of NF-κB and caspase-1-dependent cytokine processing. *J. Biol. Chem.* **277,** 29874–29880 (2002).
14. Poyet, J. L. *et al.* Identification of Ipaf, a human caspase-1-activating protein related to Apaf-1. *J. Biol. Chem.* **276,** 28309–28313 (2001).
15. Fernandes-Alnemri, T., Yu, J. W., Datta, P., Wu, J. & Alnemri, E. S. AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature* **458,** 509–513 (2009).
16. Hornung, V. *et al.* AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* **458,** 514–518 (2009).
17. Roberts, T. L. *et al.* HIN-200 proteins regulate caspase activation in response to foreign cytoplasmic DNA. *Science* **323,** 1057–1060 (2009).
18. Kerur, N. *et al.* IFI16 acts as a nuclear pathogen sensor to induce the inflammasome in response to Kaposi sarcoma-associated herpesvirus infection. *Cell Host Microbe* **9,** 363–375 (2011).
19. Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440,** 228–232 (2006).
20. Dostert, C. *et al.* Innate immune activation through Nalp3 inflammasome sensing of asbestos and silica. *Science* **320,** 674–677 (2008).
21. Petrilli, V. *et al.* Activation of the NALP3 inflammasome is triggered by low intracellular potassium concentration. *Cell Death Differ.* **14,** 1583–1589 (2007).
22. Martinon, F., Petrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440,** 237–241 (2006).
23. Jin, M. S. & Lee, J. O. Structures of the Toll-like receptor family and its ligand complexes. *Immunity* **29,** 182–191 (2008).
24. Hornung, V. *et al.* Silica crystals and aluminum salts activate the NALP3 inflammasome through phagosomal destabilization. *Nature Immunol.* **9,** 847–856 (2008).
25. Zhou, R., Tardivel, A., Thorens, B., Choi, I. & Tschopp, J. Thioredoxin-interacting protein links oxidative stress to inflammasome activation. *Nature Immunol.* **11,** 136–140 (2010).
26. Hornung, V. & Latz, E. Critical functions of priming and lysosomal damage for NLRP3 activation. *Eur. J. Immunol.* **40,** 620–623 (2010).
27. Bryan, N. B. *et al.* Differential splicing of the apoptosis-associated speck like protein containing a caspase recruitment domain (ASC) regulates inflammasomes. *J. Inflamm.* **7,** 23 (2010).
28. Stehlik, C. & Dorfleutner, A. COPs and POPs: modulators of inflammasome activity. *J. Immunol.* **179,** 7993–7998 (2007).
29. Young, J. L. *et al.* The serpin proteinase inhibitor 9 is an endogenous inhibitor of interleukin 1β-converting enzyme (caspase-1) activity in human vascular smooth muscle cells. *J. Exp. Med.* **191,** 1535–1544 (2000).
30. Bryan, N. B., Dorfleutner, A., Rojanasakul, Y. & Stehlik, C. Activation of inflammasomes requires intracellular redistribution of the apoptotic speck-like protein containing a caspase recruitment domain. *J. Immunol.* **182,** 3173–3182 (2009).
31. Mariathasan, S. *et al.* Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature* **430,** 213–218 (2004).
32. Broz, P., von Moltke, J., Jones, J. W., Vance, R. E. & Monack, D. M. Differential requirement for Caspase-1 autoproteolysis in pathogen-induced cell death and cytokine processing. *Cell Host Microbe* **8,** 471–483 (2010).
33. Saitoh, T. *et al.* Loss of the autophagy protein Atg16L1 enhances endotoxin-induced IL-1β production. *Nature* **456,** 264–268 (2008).
34. Zhou, R., Yazdi, A. S., Menu, P. & Tschopp, J. A role for mitochondria in NLRP3 inflammasome activation. *Nature* **469,** 221–225 (2011).
35. Nakahira, K. *et al.* Autophagy proteins regulate innate immune responses by inhibiting the release of mitochondrial DNA mediated by the NALP3 inflammasome. *Nature Immunol.* **12,** 222–230 (2011).
36. Guarda, G. *et al.* Type I interferon inhibits interleukin-1 production and inflammasome activation. *Immunity* **34,** 213–223 (2011).
37. Guarda, G. *et al.* T cells dampen innate immune responses through inhibition of NLRP1 and NLRP3 inflammasomes. *Nature* **460,** 269–273 (2009).
38. Elinav, E., Strowig, T., Henao-Mejia, J. & Flavell, R. A. Regulation of the antimicrobial response by NLR proteins. *Immunity* **34,** 665–679 (2011).
39. Ichinohe, T., Pang, I. K. & Iwasaki, A. Influenza virus activates inflammasomes via its intracellular M2 ion channel. *Nature Immunol.* **11,** 404–410 (2010).
40. Sander, L. E. *et al.* Detection of prokaryotic mRNA signifies microbial viability and promotes immunity. *Nature* **474,** 385–389 (2011).
41. Boyden, E. D. & Dietrich, W. F. *Nalp1b* controls mouse macrophage susceptibility to anthrax lethal toxin. *Nature Genet.* **38,** 240–244 (2006).
42. Hsu, L. C. *et al.* A NOD2–NALP1 complex mediates caspase-1-dependent IL-1β secretion in response to *Bacillus anthracis* infection and muramyl dipeptide. *Proc. Natl Acad. Sci. USA* **105,** 7803–7808 (2008).
43. Faustin, B. *et al.* Reconstituted NALP1 inflammasome reveals two-step mechanism of caspase-1 activation. *Mol. Cell* **25,** 713–724 (2007).
44. Sutterwala, F. S. *et al.* Immune recognition of *Pseudomonas aeruginosa* mediated by the IPAF/NLRC4 inflammasome. *J. Exp. Med.* **204,** 3235–3245 (2007).
45. Miao, E. A. *et al.* Innate immune detection of the type III secretion apparatus through the NLRC4 inflammasome. *Proc. Natl Acad. Sci. USA* **107,** 3076–3080 (2010).
46. Lightfield, K. L. *et al.* Critical function for Naip5 in inflammasome activation by a conserved carboxy-terminal domain of flagellin. *Nature Immunol.* **9,** 1171–1178 (2008).
47. Lamkanfi, M. *et al.* The Nod-like receptor family member Naip5/Birc1e restricts *Legionella pneumophila* growth independently of caspase-1 activation. *J. Immunol.* **178,** 8022–8027 (2007).
48. Kofoed, E. M. & Vance, R. E. Innate immune recognition of bacterial ligands by NAIPs determines inflammasome specificity. *Nature* **477,** 592–595 (2011).
49. Fernandes-Alnemri, T. *et al.* The AIM2 inflammasome is critical for innate immunity to *Francisella tularensis*. *Nature Immunol.* **11,** 385–393 (2010).
50. Rathinam, V. A. *et al.* The AIM2 inflammasome is essential for host defense against cytosolic bacteria and DNA viruses. *Nature Immunol.* **11,** 395–402 (2010).
51. Miao, E. A. *et al.* Caspase-1-induced pyroptosis is an innate immune effector mechanism against intracellular bacteria. *Nature Immunol.* **11,** 1136–1142 (2010).
52. Broz, P. *et al.* Redundant roles for inflammasome receptors NLRP3 and NLRC4 in host defense against *Salmonella*. *J. Exp. Med.* **207,** 1745–1755 (2010).
53. Taxman, D. J., Huang, M. T. & Ting, J. P. Inflammasome inhibition as a pathogenic stealth mechanism. *Cell Host Microbe* **8,** 7–11 (2010).
54. Stasakova, J. *et al.* Influenza A mutant viruses with altered NS1 protein function provoke caspase-1 activation in primary human macrophages, resulting in fast apoptosis and release of high levels of interleukins 1β and 18. *J. Gen. Virol.* **86,** 185–195 (2005).
55. Ray, C. A. *et al.* Viral inhibition of inflammation: cowpox virus encodes an inhibitor of the interleukin-1β converting enzyme. *Cell* **69,** 597–604 (1992).
56. Gregory, S. M. *et al.* Discovery of a viral NLR homolog that inhibits the inflammasome. *Science* **331,** 330–334 (2011).
57. Cassel, S. L. *et al.* The Nalp3 inflammasome is essential for the development of silicosis. *Proc. Natl Acad. Sci. USA* **105,** 9035–9040 (2008).
58. Denoble, A. E. *et al.* Uric acid is a danger signal of increasing risk for osteoarthritis through inflammasome activation. *Proc. Natl Acad. Sci. USA* **108,** 2088–2093 (2011).
59. Pazar, B. *et al.* Basic calcium phosphate crystals induce monocyte/macrophage IL-1β secretion through the NLRP3 inflammasome *in vitro*. *J. Immunol.* **186,** 2495–2502 (2011).
60. Kool, M. *et al.* Alum adjuvant stimulates inflammatory dendritic cells through activation of the NALP3 inflammasome. *J. Immunol.* **181,** 3755–3759 (2008).
61. Li, H., Willingham, S. B., Ting, J. P. & Re, F. Inflammasome activation by alum and alum's adjuvant effect are mediated by NLRP3. *J. Immunol.* **181,** 17–21 (2008).
62. Eisenbarth, S. C., Colegio, O. R., O'Connor, W., Sutterwala, F. S. & Flavell, R. A. Crucial role for the Nalp3 inflammasome in the immunostimulatory properties of aluminium adjuvants. *Nature* **453,** 1122–1126 (2008).
63. Kuroda, E. *et al.* Silica crystals and aluminum salts regulate the production of prostaglandin in macrophages via NALP3 inflammasome-independent mechanisms. *Immunity* **34,** 514–526 (2011).
64. Ellebedy, A. H. *et al.* Inflammasome-independent role of the apoptosis-associated speck-like protein containing CARD (ASC) in the adjuvant effect of MF59. *Proc. Natl Acad. Sci. USA* **108,** 2927–2932 (2011).
65. McKee, A. S. *et al.* Alum induces innate immune responses through macrophage and mast cell sensors, but these sensors are not required for alum to act as an adjuvant for specific immunity. *J. Immunol.* **183,** 4403–4414 (2009).

66. Franchi, L. & Nunez, G. The Nlrp3 inflammasome is critical for aluminium hydroxide-mediated IL-1β secretion but dispensable for adjuvant activity. *Eur. J. Immunol.* **38,** 2085–2089 (2008).
67. Flach, T. L. *et al.* Alum interaction with dendritic cell membrane lipids is essential for its adjuvanticity. *Nature Med.* **17,** 479–487 (2011).
68. Feve, B. & Bastard, J. P. The role of interleukins in insulin resistance and type 2 diabetes mellitus. *Nature Rev. Endocrinol.* **5,** 305–311 (2009).
69. Netea, M. G. *et al.* Deficiency of interleukin-18 in mice leads to hyperphagia, obesity and insulin resistance. *Nature Med.* **12,** 650–656 (2006).
70. Vandanmagsar, B. *et al.* The NLRP3 inflammasome instigates obesity-induced inflammation and insulin resistance. *Nature Med.* **17,** 179–188 (2011).
71. Lumeng, C. N., Bodzin, J. L. & Saltiel, A. R. Obesity induces a phenotypic switch in adipose tissue macrophage polarization. *J. Clin. Invest.* **117,** 175–184 (2007).
72. Wen, H. *et al.* Fatty acid-induced NLRP3-ASC inflammasome activation interferes with insulin signaling. *Nature Immunol.* **12,** 408–415 (2011).
73. Stienstra, R. *et al.* The inflammasome-mediated caspase-1 activation controls adipocyte differentiation and insulin sensitivity. *Cell Metab.* **12,** 593–605 (2010).
74. Dinarello, C. A., Donath, M. Y. & Mandrup-Poulsen, T. Role of IL-1β in type 2 diabetes. *Curr. Opin. Endocrinol. Diabetes Obes.* **17,** 314–321 (2010).
75. Lamkanfi, M. *et al.* Glyburide inhibits the cryopyrin/Nalp3 inflammasome. *J. Cell Biol.* **187,** 61–70 (2009).
76. Larsen, C. M. *et al.* Interleukin-1-receptor antagonist in type 2 diabetes mellitus. *N. Engl. J. Med.* **356,** 1517–1526 (2007).
77. Seino, S & Study Group of Comprehensive Analysis of Genetic Factors in Diabetes Mellitus. S20G mutation of the amylin gene is associated with type II diabetes in Japanese. Study Group of Comprehensive Analysis of Genetic Factors in Diabetes Mellitus. *Diabetologia* **44,** 906–909 (2001).
78. Masters, S. L. *et al.* Activation of the NLRP3 inflammasome by islet amyloid polypeptide provides a mechanism for enhanced IL-1β in type 2 diabetes. *Nature Immunol.* **11,** 897–904 (2010).
79. Duewell, P. *et al.* NLRP3 inflammasomes are required for atherogenesis and activated by cholesterol crystals. *Nature* **464,** 1357–1361 (2010).
80. Plump, A. S. *et al.* Severe hypercholesterolemia and atherosclerosis in apolipoprotein E-deficient mice created by homologous recombination in ES cells. *Cell* **71,** 343–353 (1992).
81. Chi, H., Messas, E., Levine, R. A., Graves, D. T. & Amar, S. Interleukin-1 receptor signaling mediates atherosclerosis associated with bacterial exposure and/or a high-fat diet in a murine apolipoprotein E heterozygote model: pharmacotherapeutic implications. *Circulation* **110,** 1678–1685 (2004).
82. Isoda, K. & Ohsuzu, F. The effect of interleukin-1 receptor antagonist on arteries and cholesterol metabolism. *J. Atheroscler. Thromb.* **13,** 21–30 (2006).
83. Menu, P. *et al.* Atherosclerosis in ApoE-deficient mice progresses independently of the NLRP3 inflammasome. *Cell Death Dis.* **2,** e137 (2011).
84. Maloy, K. J. & Powrie, F. Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature* **474,** 298–306 (2011).
85. Siegmund, B., Lehr, H. A., Fantuzzi, G. & Dinarello, C. A. IL-1β -converting enzyme (caspase-1) in intestinal inflammation. *Proc. Natl Acad. Sci. USA* **98,** 13249–13254 (2001).
86. Bauer, C. *et al.* Colitis induced in mice with dextran sulfate sodium (DSS) is mediated by the NLRP3 inflammasome. *Gut* **59,** 1192–1199 (2010).
87. Allen, I. C. *et al.* The NLRP3 inflammasome functions as a negative regulator of tumorigenesis during colitis-associated cancer. *J. Exp. Med.* **207,** 1045–1056 (2010).
88. Hirota, S. A. *et al.* NLRP3 inflammasome plays a key role in the regulation of intestinal homeostasis. *Inflamm. Bowel Dis.* **17,** 1359–1372 (2011).
89. Zaki, M. H. *et al.* The NLRP3 inflammasome protects against loss of epithelial integrity and mortality during experimental colitis. *Immunity* **32,** 379–391 (2010).
90. Dupaul-Chicoine, J. *et al.* Control of intestinal homeostasis, colitis, and colitis-associated colorectal cancer by the inflammatory caspases. *Immunity* **32,** 367–378 (2010).
91. Elinav, E. *et al.* NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145,** 745–757 (2011).
92. Normand, S. *et al.* Nod-like receptor pyrin domain-containing protein 6 (NLRP6) controls epithelial self-renewal and colorectal carcinogenesis upon injury. *Proc. Natl Acad. Sci. USA* **108,** 9601–9606 (2011).
93. Chen, G. Y., Liu, M., Wang, F., Bertin, J. & Nunez, G. A functional role for Nlrp6 in intestinal inflammation and tumorigenesis. *J. Immunol.* **186,** 7187–7194 (2011).
94. Hu, B. *et al.* Inflammation-induced tumorigenesis in the colon is regulated by caspase-1 and NLRC4. *Proc. Natl Acad. Sci. USA* **107,** 21635–21640 (2010).
95. Hoffman, H. M. & Wanderer, A. A. Inflammasome and IL-1β-mediated disorders. *Curr. Allergy Asthma Rep.* **10,** 229–235 (2010).
96. Saleh, M. & Trinchieri, G. Innate immune mechanisms of colitis and colitis-associated colorectal cancer. *Nature Rev. Immunol.* **11,** 9–20 (2011).
97. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472,** 57–63 (2011).
98. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
99. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
100. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl Acad. Sci. USA* **107,** 14691–14696 (2010).

Acknowledgements We thank S. Eisenbarth, M. de Zoete, C. Thaiss, P. Licona-Limón, V. Nagy and A. Rongvaux for critically reading the manuscript, and F. Manzo for help with its submission. We apologize to those whose work was not cited owing to space constraints. T.S. and J.H.M. are supported by postdoctoral fellowships from the Leukemia and Lymphoma Society. E.E. is supported by the Cancer Research Institute (2010–2012) and the United States–Israel Educational Foundation (2009), and is the recipient of the Claire and Emmanuel G. Rosenblatt award from the American Physicians for Medicine in Israel Foundation. R.F. is an Investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to R.F. (richard.flavell@yale.edu).

REVIEW

doi:10.1038/nature10760

The DNA damage response and cancer therapy

Christopher J. Lord[1]* & Alan Ashworth[1]*

Genomic instability is one of the most pervasive characteristics of tumour cells and is probably the combined effect of DNA damage, tumour-specific DNA repair defects, and a failure to stop or stall the cell cycle before the damaged DNA is passed on to daughter cells. Although these processes drive genomic instability and ultimately the disease process, they also provide therapeutic opportunities. A better understanding of the cellular response to DNA damage will not only inform our knowledge of cancer development but also help to refine the classification as well as the treatment of the disease.

Towards the end of his life, the German biologist Theodor Boveri (1862–1915) published a prophetic and much-quoted manuscript on the genetic basis of cancer[1,2] that in contemporary terms[2] would be hailed as a great example of 'translation' or 'impact'. In his study of chromosome and cell behaviour in the sea urchin, Boveri suggested that the pathogenesis of cancer might be driven by a "specific and abnormal chromosome constitution"[1,2]. A century later, genomic instability is recognized as a characteristic of most solid tumours and adult-onset leukaemias. In most cancers, the instability is obvious as alterations in chromosome number and structure, a phenotype termed chromosomal instability, and as changes to the structure of DNA, such as nucleotide substitutions, insertions and deletions. When they occur in crucial 'driver' genes (of which there are probably fewer than ten per tumour) these mutations can alter cell behaviour, confer a selective advantage and drive the development of the disease. Importantly, these mutations can also influence how the tumour will respond to therapy. Alongside key driver mutations, emerging data from cancer genome sequencing suggests that a typical tumour may contain many thousands of other genetic changes. These 'passenger' mutations do not contribute directly to the disease, but are probably collateral damage from exposure to various environmental factors or defects in the molecular mechanisms that maintain the integrity of the genome.

DNA damage and the DNA damage response

Given the potentially devastating effects of genomic instability, cells have evolved an intricate series of interlocking mechanisms that maintain genomic integrity[3]. The size of this task is daunting; the integrity of DNA is continually challenged by a variety of agents and processes that either alter the DNA sequence directly or cause mutation when DNA is suboptimally repaired. For example, the ultraviolet component of sunlight can cause up to 1×10^5 DNA lesions per cell per day[4], many of which are pyrimidine dimers. If left unrepaired, dimers that contain cytosine residues are prone to deamination, which can ultimately result in cytosine being replaced with thymine in the DNA sequence. Likewise, ionizing radiation (for example, from sunlight or cosmic radiation) can cause single-strand breaks (SSBs) and double-strand breaks (DSBs) in the DNA double helix backbone. If misrepaired — for example, the inaccurate rejoining of broken DNA ends at DSBs — these breaks can induce mutations and lead to widespread structural rearrangement of the genome. Environmental agents that cause DNA damage and mutation include cigarette smoke (40 years of heavy smoking is estimated to cause as many as 1,000 DNA lesions per cell[5]), industrial chemicals, mustard gases used in warfare and a wide variety of drugs used in chemotherapy. This final group of agents has been the mainstay of cancer therapy for decades and includes drugs that induce covalent crosslinks between DNA bases (for example, cisplatin, carboplatin, oxaliplatin and mitomycin C), drugs that attach alkyl groups to bases (such as methyl methanesulphonate and temozolomide) and agents that cause SSBs or DSBs by trapping topoisomerase I or II enzymes on DNA (such as camptothecin and etoposide)[3]. Endogenous processes also present a challenge to the integrity of the genome. For example, normal cellular metabolism generates reactive oxygen species that can oxidize DNA bases and cause SSBs. Similarly, errors in DNA replication caused by deoxyribonucleoside 5′-triphosphate (dNTP) misincorporation are potentially mutagenic, as is base depurination and deamination, an intrinsic chemical liability of DNA[3].

The variety and frequency of DNA lesions are matched by the complexity of mechanisms that counteract these threats to genomic integrity. Collectively, these mechanisms are known as the DNA damage response (DDR). In general terms, the DDR can be divided into a series of distinct, but functionally interwoven, pathways, which are defined largely by the type of DNA lesion they process (Fig. 1). Most DDR pathways encompass a similar set of tightly coordinated processes: namely the detection of DNA damage, the accumulation of DNA repair factors at the site of damage and finally the physical repair of the lesion. Most of the subtle changes to DNA, such as oxidative lesions, alkylation products and SSBs, are repaired through a series of mechanisms that is termed base excision repair (BER). In BER, damaged bases are first removed from the double helix, and the 'injured' section of the DNA backbone is then excised and replaced with newly synthesized DNA[6]. Key to this process are the enzymes poly(ADP-ribose) polymerase 1 and 2 (PARP1 and PARP2), which act as sensors and signal transducers for lesions, such as SSBs. In addition to BER, the pool of deoxynucleotides (deoxyadenosine triphosphate (dATP), deoxythymidine triphosphate (dTTP), deoxyguanosine triphosphate (dGTP) and deoxycytidine triphosphate (dCTP)) that provide the building blocks of DNA can be chemically modified before they are incorporated into the double helix. The nucleotide pool is, therefore, continually 'sanitized' by enzymes such as nudix-type motif 5 (NUDT5)[7].

Whereas small base adducts are repaired by BER, some of the bulkier single-strand lesions that distort the DNA helical structure, such as those caused by ultraviolet light, are processed by nucleotide excision repair (NER)[8]. NER is often subclassified into transcription-coupled NER, which occurs where the lesion blocks, and is detected

[1]The Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK. *These authors contributed equally to this work.

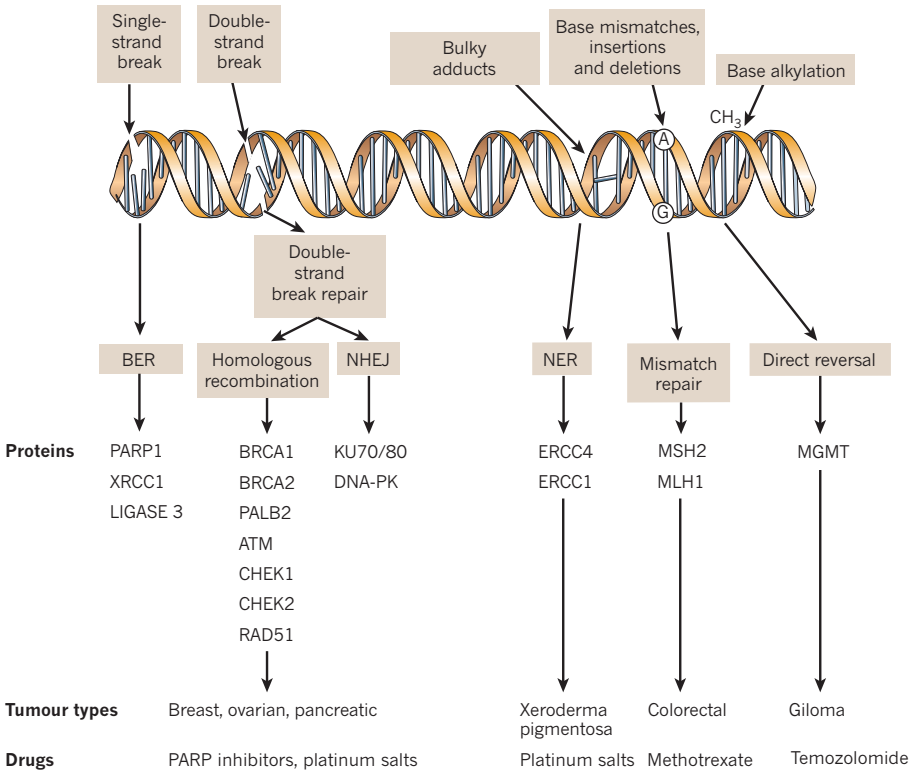


Figure 1 | A panoply of DNA repair mechanisms maintains genomic stability. DNA is continually exposed to a series of insults that cause a range of lesions, from single-strand breaks (SSBs) to base alkylation events. The choice of repair mechanism is largely defined by the type of lesion, but factors such as the stage of the cell cycle also have a role. Key proteins involved in each DDR mechanism, the tumour types usually characterized by DDR defects and the drugs that target these defects are shown. BER, base excision repair; NER, nucleotide excision repair; NHEJ, non-homologous end-joining. Figure modified, with permission, from ref. 72.

by elongating RNA polymerase; and global-genome NER, in which the lesion is detected not as part of a blocked transcription process but because it disrupts base pairing and distorts the DNA helix. Although these processes detect lesions using different mechanisms, they repair them in a similar way: DNA surrounding the lesion is excised and then replaced using the normal DNA replication machinery. Excision repair cross-complementing protein 1 (ERCC1) is key to this excision step.

The major mechanisms that cope with DSBs are homologous recombination[9] and non-homologous end joining (NHEJ)[10]. Homologous recombination acts mainly in the S and G2 phases of the cell cycle and is a conservative process in that it tends to restore the original DNA sequence to the site of damage. Part of the DNA sequence around the DSB is removed (known as resection) and the DNA sequence on a homologous sister chromatid is used as a template for the synthesis of new DNA at the DSB site. Crucial proteins involved in mediating homologous recombination include those encoded by the *BRCA1*, *BRCA2*, *RAD51* and *PALB2* genes. In contrast to homologous recombination, NHEJ occurs throughout the cell cycle. Rather than using a homologous DNA sequence to guide DNA repair, NHEJ mediates repair by directly ligating the ends of a DSB together. Sometimes this process can cause the deletion or mutation of DNA sequences at or around the DSB site. Therefore, compared with homologous recombination, NHEJ, although mechanistically simpler, can often be mutagenic.

Mismatch repair[11] is crucial to the DDR. It deals primarily with dNTP misincorporation and formation of 'insertion and deletion' loops that form during DNA replication. These errors cause base 'mismatches' in the DNA sequence (that is, non-Watson-Crick base pairing) that distort the helical structure of DNA and so are recognized as DNA lesions. The recognition of this distortion triggers a procession of events resulting in the excision of newly synthesized DNA encompassing the mismatch site and the resynthesis of DNA in its place. Key to the process of mismatch repair are proteins encoded by the mutS and mutL homologue genes, such as *MSH2* and *MLH1*.

Finally, translesion synthesis and template switching allow DNA to continue to replicate in the presence of DNA lesions that would otherwise halt the process. Translesion synthesis and template switching are therefore usually considered to be part of the DDR. In translesion synthesis, relatively high-fidelity DNA replication polymerases are transiently replaced with low-fidelity 'translesion' polymerases that are able to synthesize DNA using a template strand encompassing a DNA lesion. Once the replication fork passes the site of the lesion, the low-fidelity DNA polymerases are normally replaced with the usual high-fidelity enzyme, which allows DNA synthesis to continue as normal. In template switching, the DNA lesion is bypassed at the replication fork by simply leaving a gap in DNA synthesis opposite the lesion. After the lesion has passed the replication fork, the single-strand gap is repaired using template DNA on a sister chromatid, similar to the process used during homologous recombination.

Although sometimes considered distinct from the DDR, the mechanisms that control the integrity of telomeric DNA at the end of each human chromosome also act as a barrier against genomic instability and mutation[12]. Rather than being an exposed DNA double-helix structure at the end of the chromosome, telomeric DNA comprises a series of guanine-rich, repetitive DNA sequences. These enable the telomeric DNA to be bound in a loop-like structure with a series of proteins (telomere repeat binding factor 1 and 2 (TERF1 and TERF2), protection of telomeres 1 (POT1), TERF1-interacting nuclear factor 2 (TINF2)) that form a shelterin complex. This 'capping' structure prevents the otherwise exposed ends of different chromosomes from becoming fused together (a process known as end–end fusion) by the DDR. In most somatic cells, the length of telomeric DNA is reduced at each cell cycle, a process termed telomere attrition. Eventually, telomeres reach a crucial length that precludes the formation of an effective shelterin complex. In normal cells, this failure in telomere capping induces a p53-mediated response that results in cellular senescence, a mechanism that ultimately prevents unlimited cell proliferation. When the p53 response is abrogated, end–end fusions occur, which leads to the formation of chromosomes with two centromeres. At mitosis, each centromere in a dicentric chromosome attaches to an opposite spindle pole. The physical stress of opposing forces during chromosome segregation shears dicentric chromosomes, resulting in broken chromosome ends. This process is an ideal substrate for the chromosome translocation, focal DNA amplification and deletion events that potentially drive tumorigenesis. Sheared chromosomes could also be the source of

the mutagenic 'big bang' or, chromothripsis[13], that has been posited to occur in some tumours. Whether or not this is the case, compared with healthy cells, many pre-invasive and invasive cancers have significantly shorter telomeres, a situation in which end–end fusions and dicentric chromosome shearing could easily occur[12].

The core DDR machinery does not work alone, but is coordinated with a set of complementary mechanisms that are also crucial to maintaining the integrity of the genome. For example, chromatin-remodelling proteins allow the DNA repair apparatus to gain access to the damaged DNA[14]. Similarly, processes such as homologous recombination are dependent on the close proximity of a sister chromatid, and probably work in tandem with the chromosome cohesion machinery to ensure this is the case. Crucially, DDR core components interact with the cell-cycle-checkpoint and chromosome-segregation machinery. These interactions allow DNA repair to occur before mitosis takes place and ensure that the correct complement of genetic material is passed on to daughter cells[15].

Cancer

One of the most pervasive characteristics of human tumours is genomic instability[16]. Although the specific DDR defects are not known in most cancers, there are several examples in which there is an incontrovertible link between a particular DDR dysfunction and the neoplastic phenotype. For example, 15% of sporadic colorectal tumours show an abnormal shortening or lengthening of dinucleotide repeat sequences. This DNA mutation pattern, known as microsatellite instability, is probably caused by an inability to repair DNA replication errors when mismatch repair is defective. Microsatellite instability is observed not only in sporadic colorectal tumours but also in a familial form of the disease known as hereditary non-polyposis colorectal cancer (HNPCC). HNPCC is associated with loss-of-function mutations in mismatch repair genes, such as *MSH2* and *MLH1* (ref. 11).

Similarly, a subset of both sporadic and familial cancers seems to be associated with defects in homologous recombination. Integrated (genomic, transcriptomic, methylomic) analysis of just under 500 samples of high-grade serous ovarian adenocarcinoma has estimated that homologous recombination is probably defective in approximately half. These homologous recombination defects are largely driven by DNA mutations or epigenetic silencing of genes such as *BRCA1* and *BRCA2* (ref. 17). Familial forms of breast, ovarian and pancreatic cancer are also associated with loss-of-function mutations in homologous recombination modifying genes, such as *BRCA1*, *BRCA2*, *PALB2*, *ATM*, *RAD51C* and *RAD51D.* A number of rare, syndromic conditions also exist in which the underlying genetic determinant is a loss-of-function mutation in a gene that modulates the response to DSBs[9]. These conditions include ataxia telangiectasia (ATM), Bloom's syndrome, Fanconi anaemia, Rothmund–Thomson syndrome, Werner syndrome and Nijmegen breakage syndrome, all of which have variable clinical phenotypes but are similar in that they cause extreme radiosensitivity, a characteristic of the inability to process DSBs effectively.

Evidence is also building that the DDR is not only invoked but also dysfunctional at an early stage in the development of neoplasia. Markers of DSBs, such as nuclear γH2AX foci (a histone phosphorylation event that occurs on chromatin surrounding a DSB), are markedly elevated in some precancerous lesions[18,19]. One hypothesis[19] proposes that the original cause of these effects is oncogene activation. The activation of oncogenes such as *MYC* and *RAS* stimulates the firing of multiple replication forks as part of a proliferative program. These forks rapidly stall, collapse and form DSBs because they exhaust the available dNTP pool or because multiple forks collide on the same chromosome. Regardless of the mechanism, stalled and collapsed forks normally invoke the DDR and cell-cycle checkpoints that enable DNA lesions to be repaired before mitosis takes place. For precancerous lesions to progress to mature tumours, it is thought that critical DSB signal transduction and cell-cycle checkpoint proteins, such as ATM, ataxia telangiectasia and Rad3 related (ATR), and the master 'gatekeeper' protein p53 become inactivated. With these DDR components rendered dysfunctional, collapsed forks are not effectively repaired, and cells proceed through the cell cycle with DNA lesions intact, increasing the chance of mutagenesis[18,19].

Selection of cancer treatment

Historically, DDR defects have already been exploited in the form of many commonly used chemotherapy and radiotherapy regimens. One example is the use of platinum salts (carboplatin or cisplatin) frequently given in combination with the taxane paclitaxel in patients with advanced ovarian cancer. Platinum salts cause DNA inter- and intrastrand crosslinks — lesions that are recognized by the DDR and repaired by a combination of NER and homologous recombination. These agents may be effective in patients with ovarian cancer because these tumours commonly harbour defects in homologous recombination[20], especially high-grade serous ovarian cancers. Estimates vary, but somewhere between 37% and 55% of high-grade serous ovarian cancers have germ line or somatic mutations in *BRCA1* or *BRCA2* and some have epigenetic inactivation of *BRCA1* (ref. 21). Cisplatin also improves survival in patients with ERCC1-negative non-small-cell lung cancers, but not in those with ERCC1-positive tumours[22]. The NER pathway, and in particular the ERCC1–ERCC4 complex, is essential for the removal of platinum adducts, and low levels of ERCC1 are associated with an accumulation of adducts and an increased apoptotic response to cisplatin. Finally, where alkylating agents work, they probably also exploit a DDR defect. Normally, O6-methyltransferase (MGMT) directly reverses DNA alkylation events caused by drugs such as temozolomide when they occur on the O6 position of guanine residues. However, in some tumours, the promoter of the *MGMT* gene is methylated, causing a reduction in MGMT expression. This defect causes O^6 alkylation events formed by temozolomide to persist so that they eventually pair incorrectly with thymidine residues during replication. The resultant DNA mismatch is recognized by the mismatch repair apparatus, which ultimately stimulates cell death. *MGMT* methylation has also been implicated in the response of tumours (such as gliomas) to temozolomide, and is now being used as a biomarker to direct the use of therapy[23] (Table 1).

An alternative approach to chemotherapy is the design of drugs that target specific DDR components. Topoisomerase inhibitors, such as irinotecan (a topoisomerase I inhibitor) and etoposide (a topoisomerase II inhibitor), could be regarded as the first generation of DDR 'targeted' agents[24]. Topoisomerases catalyse the breaking and rejoining of the phosphodiester backbone as part of the process that unwinds the torsional structure of DNA before processes such as transcription and DNA replication; agents that inhibit this function, such as irinotecan and etoposide, leave DNA breaks across the genome and have been used effectively in some cancers (Table 1). The development of the next generation of DDR inhibitors, particularly PARP inhibitors, has given considerable impetus to this area.

PARP inhibitors as targeted therapy

PARP1 and PARP2 are members of the PARP protein superfamily. These enzymes catalyse the polymerization of ADP-ribose moieties onto target proteins (PARsylation) using NAD^+ as a substrate, releasing nicotinamide in the process. This modification often modulates the conformation, stability or activity of the target protein[25]. The best understood role of PARP1 is in SSB repair ((SSBR) a form of BER). PARP1 initiates this process by detecting and binding SSBs through a zinc finger in the PARP protein. Catalytic activity of PARP1 results in the PARsylation of PARP1 itself and the PARsylation of a series of additional proteins, such as XRCC1 and the histones H1 and H2B; when PARP activity is inhibited, SSB repair is compromised[25]. The development of PARP inhibitors began with the observation that nicotinamide, a product of PARP catalytic activity, is itself a weak PARP inhibitor, as are nicotinamide analogues, such as 3-aminobenzamide. The screening of chemical libraries and the subsequent chemical refinement

Table 1 | Examples of DDR inhibitors in clinical use or in development

Class	Name	Stage of development	Cancer type
Topoisomerase I inhibitor	Irinotecan	Licensed for use	Mainly colorectal cancers
Topoisomerase II inhibitor	Etoposide phosphate	Licensed for use	For example, testicular cancer, Ewing's sarcoma, lung cancer, lymphoma, lymphocytic leukaemia and glioblastoma multiforme
DNA protein kinase inhibitor	CC-115	Phase I study under way	Advanced solid tumours, non-Hodgkin's lymphoma and multiple myeloma
PARP inhibitor	Olaparib	2 phase II studies completed, 9 additional phase II studies under way	Ovarian, breast, gastric, colorectal and pancreatic tumours and a range of other solid tumours
	Veliparib	16 phase II studies under way	Ovarian, breast, gastric, colorectal and pancreatic tumours and a range of other solid tumours
	Rucaparib	2 phase II studies recruiting	Breast and other solid tumours
	MK4827	Phase I study completed, phase Ib ongoing	Glioblastoma multiforme, melanoma and other solid tumours and leukaemias
CHEK1 inhibitor	AZD7762	Phase I study completed	Range of solid tumours
MGMT inhibitor	Lomeguatrib	Phase I study under way	Colorectal cancer and other solid tumours
Telomerase inhibitor	Imetelstat	4 phase II studies under way	Solid tumours and leukaemias

of compounds have led to second generation PARP inhibitors (for example, the benzimidazole-4-carboxamide NU1025), and to a third generation of clinically usable inhibitors, such as the tricyclic lactam indole, AG014699, and phthalazinones, such as AZD2281 (olaparib)[26].

For cancer therapy, PARP inhibitors are potent chemosensitizers. For example, 3-aminobenzamide enhances the cytotoxic effect of the DNA-methylating agent dimethyl sulphate, and several preclinical studies have shown that clinical PARP inhibitors increase the effects of another DNA-methylating agent, temozolomide[25,26]. In 2005, potent PARP inhibitors were shown to selectively inhibit the growth of cells with defects in either the *BRCA1* or *BRCA2* genes, suggesting a new use for these agents[27,28]. PARP inhibition probably works by allowing the persistence of spontaneously occurring SSBs, or by inhibiting PARP release from a DNA lesion. Whichever is the case, both of these DNA lesion types could conceivably stall and collapse replication forks, potentially creating lethal DSBs[29]. In normal cells, the effects of PARP inhibition are buffered by homologous recombination, which repairs the resultant DSB. However, effective homologous recombination is reliant on functioning *BRCA1* and *BRCA2*, so when these genes are defective — as they are in tumours of germline *BRCA*-mutant carriers — DSBs are left unrepaired, and potent PARP inhibitors can cause cell death. In some *in vitro* models, cells with *BRCA* mutations are more than 1,000 times more sensitive to PARP inhibitors than cells with wild-type alleles[27]. This genotype-specific selectivity provides a strong argument for these agents to be tested clinically as single agents, rather than in combination with chemotherapies. Moreover, the sensitivity to PARP inhibitors seems to be defined more by the *BRCA* genotype of a cancer cell than by its tissue of origin. Breast, ovarian and prostate cancers with *BRCA* mutations all seem to be profoundly sensitive to these drugs. The demonstration that *BRCA*-mutant cells are sensitive to PARP inhibitors has now become the emblematic 'poster child' for exploiting synthetic lethality in cancer therapy. Synthetic lethality, a genetic concept developed by Dobzhansky[30], Lucchesi[31] and others and championed as a therapeutic approach for cancer by Hartwell and colleagues[32] and Kaelin[33], is the process in which different single gene defects are compatible with cell viability, but the synthesis or combination of these gene defects results in cell death. In this case, *BRCA* dysfunction is synthetically lethal when PARP is inhibited. Furthermore, the demonstration that one DDR pathway (homologous recombination) could be targeted by inhibition of another (PARP-mediated BER and SSB repair) suggests that DDR pathways functionally buffer one another. This hypothesis has long been established in model organisms (such as yeast) in which synthetic lethalities can be easily identified using genetic screens. These studies and the BRCA/PARP paradigm in human tumour cells has inspired the search for other DDR synthetic lethalities that could be exploited in the treatment of cancer, including other synthetic lethal approaches for *BRCA*-mutant tumour cells[34], as well as for tumour cells with other gene defects. For example, the mismatch repair gene defects found in colorectal cancer are synthetically lethal with inhibition of genes controlling the response to oxidative DNA damage, such as *POLB*[35], *POLG*[35] and *PINK1* (ref. 36). The first therapeutic test of the hypothesis that mismatch-repair defective cancers can be treated with a synthetic lethality approach[37] is now in a proof-of-concept clinical trial (http://www.clinicaltrials.gov, NCT00952016).

The preclinical work described above has paved the way for subsequent clinical studies of PARP inhibitors (Table 1). Phase I studies have established the safety of olaparib, a potent PARP inhibitor, as a single agent and shown that significant and durable antitumour responses can be established in patients with *BRCA*-mutant breast, ovarian or prostate tumours[38]. Furthermore, olaparib does not seem to cause many of the side effects associated with standard chemotherapies. In subsequent phase II clinical studies, 40% of patients with breast or ovarian cancer with germline BRCA mutations had a favourable response to the drug. This is a particularly high response given that the patients in these trials had been heavily pretreated and had become resistant to a range of chemotherapies. Such work has led to a series of additional PARP inhibitors being tested in clinical trials[39].

Recently, attention has focused on clinical trials assessing iniparib. Unlike drugs such as olaparib and AG014699, iniparib was originally proposed to be a non-competitive inhibitor of PARP1 that disrupts the interaction between PARP1 and DNA[40]. In a phase II trial, iniparib elicited a significant response in patients with triple negative (oestrogen-receptor, progesterone-receptor and *HER2*-amplification negative) breast cancer when used in combination with gemcitabine and carboplatin. However, a subsequent phase III trial failed to deliver the predefined improvements in progression-free or overall survival. The reasons for this failure are unclear, although differences in the patient populations and study design may have contributed. However, the clinical activity of iniparib may not be related to PARP inhibition[41]; this drug has very low PARP inhibition *in vitro,* and its mechanism of action *in vivo* is unclear[40,41]. We therefore believe that there is little evidence to regard the relative failure of iniparib in phase III trials as a PARP-inhibitor class effect. In fact, the clinical observations with potent catalytic-site PARP inhibitors have been promising when used in appropriate populations[39].

Alongside the clinical trials of PARP inhibitors, preclinical research has extended our understanding of how these agents work and how they might be used most effectively in the clinic. Since the initial reports of

BRCA/PARP synthetic lethality, several other determinants of PARP-inhibitor sensitivity have been identified. As expected, most of these determinants cause PARP-inhibitor sensitivity by modulating homologous recombination (for example, RAD51, ATRX, SHFM1, RPA1, NBN, ATR, ATM, CHEK1, CHEK2, miR-182, CDK1, SWI5–SFR1, USP1/UAF1 and several Fanconi anaemia proteins[42–47]). Of particular interest is the observation that cyclin dependent kinase 1 (CDK1) inhibition suppresses homologous recombination and may therefore elicit PARP-inhibitor sensitivity both *in vitro* and *in vivo*[46]. More extensive studies are now required to identify the patient subgroup, and specifically the genetic context, in which this combination could be used. In addition, mutations in the tumour-suppressor gene *PTEN*[48] and the *TMPRSS2–ERG* fusion[49] can sensitize cells to clinical PARP inhibitors; both of these changes are relatively common in human tumours so could be used as biomarkers to direct the use of these agents. In the case of the PTEN/PARP synthetic lethality, one model suggests that PTEN deficiency causes a downregulation in RAD51 levels, therefore disrupting homologous recombination[49]. Finally, several PARP-inhibitor-resistance-causing mechanisms have been identified, particularly in *BRCA1*- or *BRCA2*-mutant tumour models. In two studies, PARP-inhibitor resistance was caused by secondary mutations in the *BRCA1* or *BRCA2* genes themselves[50,51]. These mutations restored the open reading frame of the *BRCA1* or *BRCA2* gene, enabled the generation of functional proteins and restored the ability to repair DNA damage caused by PARP inhibitors. Importantly, these secondary mutations were observed not only in *in vitro* models but also in ovarian tumours from patients who had become resistant to platinum salts such as carboplatin[50–53]. These drugs target cells with deficient homologous recombination, albeit with less efficiency than PARP inhibitors. Subsequent studies have suggested that these reversion events are relatively common in *BRCA*-mutant cancers that do not respond to platinum salts. This is especially true in ovarian cancer, in which more than 42% of the carboplatin-resistant tumours tested had secondary mutations that restored the *BRCA1* or *BRCA2* open reading frames[54]. Taken together, these observations could help to explain why some platinum-refractory *BRCA*-mutant patients in PARP-inhibitor clinical trials failed to respond and strengthen the argument for the inclusion of biomarkers that monitor *BRCA1* and *BRCA2* sequence and homologous recombination function in future trials. Furthermore, the restoration of homologous recombination and drug resistance by secondary mutations provides strong confirmation that, at least in this setting, PARP inhibitors and platinum salts require defective homologous recombination for antitumour activity. Other identified mechanisms of PARP-inhibitor response include the upregulation of genes that encode P-glycoprotein efflux pumps[55] and the loss of TP53BP1 (a DDR protein) in *BRCA1*-deficient cells, suggesting that these mechanisms should also be assessed in clinical biopsies[56,57].

In addition to dissecting the mechanisms behind PARP-inhibitor sensitivity and resistance, there has been considerable debate regarding the potential of PARP inhibitors in non-familial cancers. Germline mutations in *BRCA1* or *BRCA2* are common in hereditary breast cancer, but relatively rare in sporadic breast cancer. However, many sporadic tumours have molecular and histopathological features in common with BRCA-deficient tumours, a concept originally termed BRCAness[20]. For example, BRCAness may be caused by methylation of the *BRCA1* gene promoter (which causes epigenetic silencing of BRCA1 expression) or even somatic mutation of genes that modulate homologous recombination. Furthermore, traditional histopathological methods and gene-expression profiling approaches have suggested some level of phenotypic or molecular overlap between triple-negative breast cancers, basal-like breast cancers and *BRCA1*-mutant familial breast cancers[20]. Although the overlap is not absolute, it has led to the hypothesis that there may be a subset of sporadic breast cancers that exhibit features of BRCAness and may respond favourably to PARP inhibitors[20]. This hypothesis has yet to be fully tested experimentally, in part because existing tumour-cell models of triple-negative breast cancer do not encompass the molecular and functional heterogeneity of triple-negative breast cancer. In the clinical setting, patients with triple-negative breast cancer do not show a frequent response to either olaparib or veliparib[39], although the number of patients tested is still small. Needless to say, larger studies are needed, as are clinical trials that lead to a full drug licence for PARP inhibitors in *BRCA*-mutation-related breast cancer.

Future issues

More research is needed into how DDR activity affects the clinical efficacy of DNA-damaging chemotherapies, and here we highlight some areas that need to be addressed in the future.

Pharmacological targeting

Although the DDR has been the subject of intense study, the biochemical and genetic interactions between the various DNA repair pathways is not completely understood. This understanding would be facilitated by the availability of potent and specific DDR-pathway small-molecule inhibitors. Past efforts have focused on the protein kinases that transduce DDR signals (such as ATM, ATR and DNA-PK) and on the PARP superfamily members that detect damage and act as signal transduction mediators. For example, synthetic lethalities with ATM and DNA-PK suggest that ATM inhibitors could increase the potential of DNA-damaging chemotherapy in *TP53*-mutant tumours, and DNA-PK inhibitors could be used in combination with chemotherapy in *ATM*-mutant tumours[58]. A number of other enzyme classes have also been identified as having a role in the DDR, including enzymes that mediate SUMOylation, histone modification, acetylation and ubiquitylation[59]. The availability of pharmacological inhibitors of these proteins would facilitate the search for additional synthetic lethal interactions. An example is the identification of small-molecule inhibitors (such as MLN4924) that modulate neddylation, a post-translational modification process analogous to ubiquitylation[60]. Inhibition of neddylation, which is induced by MLN4924, causes stabilization of the DNA-replication licensing factor CDT1, DNA synthesis in the absence of cell-cycle progression and activation of the DDR[60]. These agents could potentially target tumour cells that have existing DNA repair defects or are under replicative stress.

The clinical use of DDR-pathway drugs remains a challenge. Addition of such agents to standard chemotherapy regimens will probably lead to greater side effects, and therefore more work in this area is needed on the appropriate dosing and scheduling of inhibitors. These studies should also address the possibility that targeting the DDR could also increase the rate of second cancers or leukaemias, something that has already been observed for some chemotherapies. In principle, these iatrogenic effects might be avoided if DDR inhibitors could be administered for a relatively short duration, a hypothesis that should be tested clinically as well as in preclinical models. Finally, combining DDR inhibitors with other targeted agents, such as signal transduction inhibitors or targeted radiation, has been underexplored and could form the basis of effective treatment regimens.

Cancer stem cells and the tumour micro-environment

Many human cancers contain cells that share some of the self-renewal and differentiation properties of stem cells. These cancer stem cells are thought to act as tumour-initiating cells that, after treatment, may potentially repopulate the therapy-resistant tumour. With this role in mind, understanding the response to DNA-damaging therapy in cancer stem cells may prove essential in understanding whether a treatment will ultimately work. Evidence is accumulating that the DDR has distinct properties in cancer stem cells. These may include controversial and unproven concepts such as the 'immortal strand' model in which parental, but not newly synthesized DNA, is retained in stem cells after mitosis; retaining only the parental strand of DNA is thought to limit the impact of DNA replication errors to the stem cell population[61]. Whether this is the case or not, a number of studies have suggested that resistance to therapy is mediated by altered DDR activity in cancer stem cells and tumour-initiating cells. For example, a comparison of

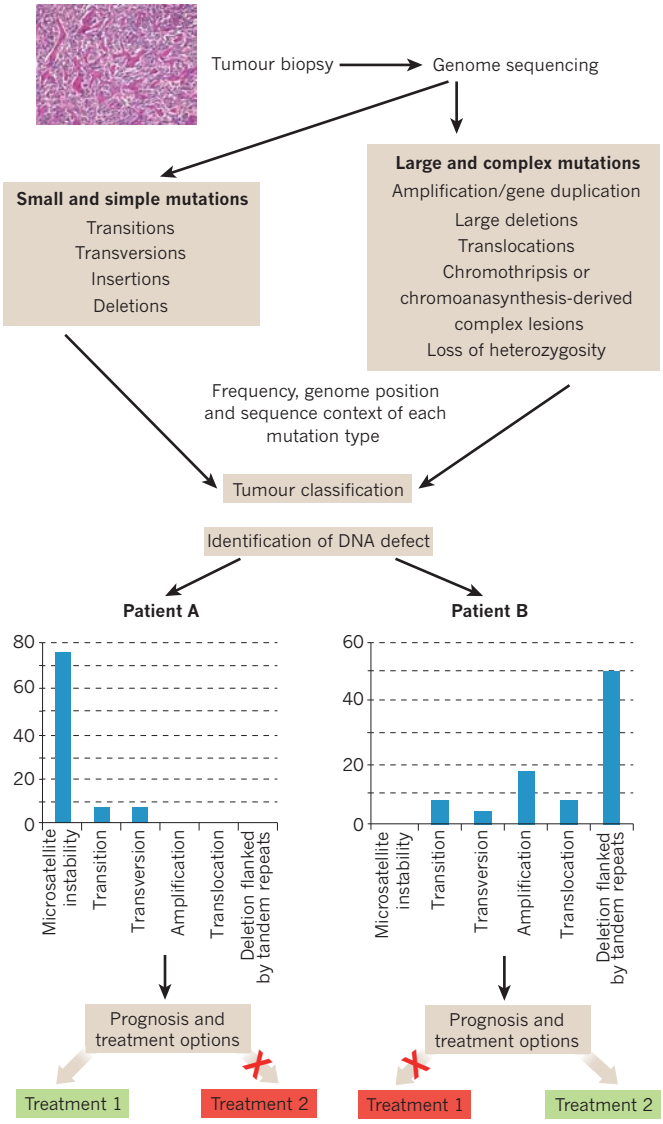


Figure 2 | Genomic scars in cancer. Modern sequencing approaches mean that the nucleotide sequence and mutational spectrum of tumour biopsies can be rapidly determined. This spectrum, or sequence scar, could be used to classify each tumour. Classifiers could take into account the frequency and location of mutations, as well as their predominant sequence context (for example, insertions and deletions that are flanked by small regions of microhomology[13,73]). Such classifiers could then be used to estimate the probable outcome for the patient and to select the most appropriate therapy. The choice of therapy is defined by the balance of sequence scars found in tumour genome biopsies.

breast tumour biopsies before and after chemotherapy treatment has shown an enrichment of cancer stem cells, which suggests an intrinsic resistance of these cells. In mouse models, cancer stem cells deficient in p53 showed accelerated DNA repair activity[62], as well as high levels of AKT and WNT signalling, both of which promoted cancer-stem-cell survival after exposure to ionizing radiation[63]. In glioblastoma multiforme, cancer stem cells have been characterized on the basis of expression of prominin (CD133; now known as PROM1)[64]. $CD133^+$ cells are enriched after irradiation of human glioblastoma multiforme xenografts, suggesting that cancer stem cells are necessary for tumour regrowth after radiotherapy. Cancer stem cells showed more robust activation of the DDR and higher intrinsic DNA repair activity. Blocking the kinases CHEK1 and CHEK2 led to enhanced killing of cancer stem cells by ionizing radiation[65], suggesting that a pharmacological approach may overcome this intrinsic resistance. Much more detailed and systematic knowledge of the consequences of DNA damage and the DDR in stem cells is required to exploit this important area.

If we are to optimize therapies that exploit the DDR, then the impact of the tumour microenvironment and the non-tumour cells that sit within a neoplastic lesion must be considered. For example, non-tumour stromal cells form at least half of the cellular complement of pancreatic ductal adenocarcinomas. This stromal compartment contains pancreatic stellate cells, which, in *in vitro* systems at least, cause tumour cells to become resistant to ionizing radiation; this radioprotective effect could explain the relative radioresistance of pancreatic ductal adenocarcinomas. Already, there is a preliminary understanding of how stromal and tumour cells interact and communicate to protect against ionizing radiation[66], but the understanding of how the microenvironment modulates the DDR is still limited. Given the complexity of studying molecular mechanisms *in vivo* and of effectively modelling cell–cell interactions *in vitro*, this is not surprising. These technical problems are not insurmountable, but they do raise an important issue: historically, much of our knowledge of the DDR has originated from studies of cells *in vitro*. Understanding the DDR *in vivo* is critical if cancer therapies are to be modelled correctly.

The genomic scar

Historically, cancer taxonomy has mainly been based on the anatomical site of a tumour and its clinical and histopathological features. Cancers are now increasingly classified according to their underlying pattern of pathogenic mutations, a process which is being facilitated by exome or whole-genome sequencing of each cancer[67]. Treatment decisions can then be made on the basis of pharmacologically targetable drivers of the disease. In addition to identifying the driver genes affected, examining the collateral genomic damage caused by DNA repair deficiency may help in the classification of cancer and could provide biomarkers to inform therapy. In essence, the mutational spectrum of a tumour, revealed by its DNA sequence, is a sequence 'scar' that reflects both the mutagens that a tumour has been exposed to and the repair processes that have operated to mitigate their impact[67]. Already, there are some precedents to suggest that this approach could have merit. For example, homologous-recombination deficient tumours have a characteristic mutational spectrum defined by the use of error-prone DSB repair mechanisms[68,69]. Therefore, the wider tumour genome sequence or sequence scar in a tumour with defective homologous recombination might be used to decide which patients should be treated with PARP inhibitors. Microsatellite instability is another example. In many cases, microsatellite instability in a tumour reveals both the mutagenic nature of DNA replication and the failure to reverse these sequence errors when mismatch repair is defective[11]. Given the impact that the DDR can have on both disease development and response to treatment, it seems reasonable to consider categorizing tumours according to their DNA repair defects and to use this information to personalize treatment (Fig. 2). As a result of the speed at which in-depth, whole-genome tumour sequences are now being generated, we will soon have a first draft of the mutational landscape of all human cancers, which will allow these possibilities to be addressed.

To match these molecular signatures of DNA mutation and repair, we require functional biomarkers that identify DDR defects and aid in the selection of therapy. Again, there are preliminary signs that this could be achievable: a mechanistic determinant of PARP-inhibitor sensitivity, RAD51 foci deficiency, can be measured by immunohistochemical staining in tumour biopsies[70], and the ability to detect other DNA repair complexes by immunohistochemistry is improving[71]. In principle, functional biomarkers should be identifiable for most DDR pathways even though significant technical and practical issues need to be resolved before such tests can be used in the clinic. Because the DDR pathway is fully engaged only after DNA is damaged, standard tumour biopsies or specimens are not particularly useful for measurement of DDR-pathway activity. A functional test of DDR-pathway deficiency might therefore need to include some

form of genotoxic stress, followed by a measurement of the molecular response. For example, to test the homologous recombination defect in tumour cells by measuring the RAD51 response, cells must first be exposed to an agent that causes DNA damage. These agents could include ionizing radiation, chemotherapy or the PARP inhibitor itself. Clinically, this could mean that patients need to be treated with a DNA-damaging agent before biopsy for the purposes of biomarker assessment or alternatively tumour biopsies would have to be exposed to DNA-damaging agents *ex vivo*. Both scenarios have practical or logistical issues. For instance, tumour cells may not respond in the same way *ex vivo* as they do *in vivo*, and pretreating a patient to assess tumour response might not always be feasible from a clinical perspective. Nevertheless, molecular and functional assays could mean that a tumour could be defined as, for example, 'deficient for homologous recombination and potentially sensitive to PARP inhibitor' before treatment rather than simply being classified as a 'non-familial, triple-negative breast cancer'. The development of functional biomarkers of other DDR pathways would considerably advance this area.

The genomic instability of tumour cells is one of their most pervasive characteristics and presents a therapeutic opportunity. Already, DNA-damaging chemotherapies have provided the core of cancer treatment for the past half-century and, if used appropriately, targeted therapies that exploit the DDR could deliver better therapeutic responses in the future. Nevertheless, there are still significant unexplored territories for the field; an understanding of the DDR *in vivo,* the definition of clinical biomarkers that allow the selection of appropriate therapy and an understanding of how the DDR interacts with the other homeostatic systems in the cell are relative unknowns. We believe that the crucial factors that have previously allowed this field to progress — the comprehensive study of the basic biology of the DDR combined with a desire to apply this information in the clinic — will drive significant advances in the future. ■

1. Boveri, T. *Zur Frage der Entstehung Maligner Tumoren* (Gustav Fischer, 1914).
2. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell Sci.* **121,** 1–84 (2008). **A translation of Boveri's original work that proposed the hypothesis that a form of genomic dysregulation characterizes tumour cells.**
3. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **40,** 179–204 (2010).
4. Hoeijmakers, J. H. DNA damage, aging, and cancer. *N. Engl. J. Med.* **361,** 1475–1485 (2009).
5. Phillips, D. H., Hewer, A., Martin, C. N., Garner, R. C. & King, M. M. Correlation of DNA adduct levels in human lung with cigarette smoking. *Nature* **336,** 790–792 (1988).
6. David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447,** 941–950 (2007).
7. Nakabeppu, Y. *et al.* Mutagenesis and carcinogenesis caused by the oxidation of nucleic acids. *Biol. Chem.* **387,** 373–379 (2006).
8. Cleaver, J. E., Lam, E. T. & Revet, I. Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. *Nature Rev. Genet.* **10,** 756–768 (2009).
9. Moynahan, M. E. & Jasin, M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature Rev. Mol. Cell Biol.* **11,** 196–207 (2010).
10. Lieber, M. R. NHEJ and its backup pathways in chromosomal translocations. *Nature Struct. Mol. Biol.* **17,** 393–395 (2010).
11. Jiricny, J. The multifaceted mismatch-repair system. *Nature Rev. Mol. Cell Biol* **7,** 335–346 (2006).
12. Artandi, S. E. & DePinho, R. A. Telomeres and telomerase in cancer. *Carcinogenesis* **31,** 9–18 (2010).
13. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
14. Bell, O., Tiwari, V. K., Thoma, N. H. & Schubeler, D. Determinants and dynamics of genome accessibility. *Nature Rev. Genet.* **12,** 554–564 (2011).
15. Warmerdam, D. O. & Kanaar, R. Dealing with DNA damage: relationships between checkpoint and repair pathways. *Mutat. Res.* **704,** 2–11 (2010).
16. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
17. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
18. Bartkova, J. *et al.* Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444,** 633–637 (2006).
19. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319,** 1352–1355 (2008). **References 18 and 19 describes how oncogenes are able to induce the DDR early in oncogenesis.**
20. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. *Nature Rev. Cancer* **4,** 814–819 (2004).
21. Banerjee, S., Kaye, S. B. & Ashworth, A. Making the best of PARP inhibitors inovarian cancer. *Nature Rev. Clin. Oncol.* **7,** 508–519 (2010).
22. Friboulet, L. *et al.* Molecular characteristics of ERCC1-negative versus ERCC1-positive tumors in resected NSCLC. *Clin. Cancer Res.* **17,** 5562–5572 (2011).
23. Weller, M. *et al. MGMT* promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Rev. Neurol.* **6,** 39–51 (2010).
24. Pommier, Y., Leo, E., Zhang, H. & Marchand, C. DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem. Biol.* **17,** 421–433 (2010).
25. Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H. & Poirier, G. G. PARP inhibition: PARP1 and beyond. *Nature Rev. Cancer* **10,** 293–301 (2010).
26. Zaremba, T. & Curtin, N. J. PARP inhibitor development for systemic cancer targeting. *Anticancer Agents Med. Chem.* **7,** 515–523 (2007).
27. Farmer, H. *et al.* Targeting the DNA repair defect in *BRCA* mutant cells as a therapeutic strategy. *Nature* **434,** 917–921 (2005).
28. Bryant, H. E. *et al.* Specific killing of *BRCA2*-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434,** 913–917 (2005). **References 27 and 28 describe the synthetic lethal interaction between *BRCA* mutations and PARP inhibition.**
29. Lord, C. J. & Ashworth, A. Targeted therapy for cancer using PARP inhibitors. *Curr. Opin. Pharmacol.* **8,** 363–369 (2008).
30. Dobzhansky, T. Genetics of natural populations. Xiii. Recombination and variability in populations of *drosophila pseudoobscura*. *Genetics* **31,** 269–290 (1946).
31. Lucchesi, J. C. Synthetic lethality and semi-lethality among functionally related mutants of *Drosophila melanogaster*. *Genetics* **59,** 37–44 (1968).
32. Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278,** 1064–1068 (1997).
33. Kaelin, W. G. Jr. Synthetic lethality: a framework for the development of wiser cancer therapeutics. *Genome Med.* **1,** 99 (2009).
34. Issaeva, N. *et al.* 6-Thioguanine selectively kills BRCA2-defective tumors and overcomes PARP inhibitor resistance. *Cancer Res.* **70,** 6268–6276 (2010).
35. Martin, S. A. *et al.* DNA polymerases as potential therapeutic targets for cancers deficient in the DNA mismatch repair proteins MSH2 or MLH1. *Cancer Cell* **17,** 235–248 (2010).
36. Martin, S. A., Hewish, M., Sims, D., Lord, C. J. & Ashworth, A. Parallel high-throughput RNA interference screens identify PINK1 as a potential therapeutic target for the treatment of DNA mismatch repair-deficient cancers. *Cancer Res.* **71,** 1836–1848 (2011).
37. Martin, S. A. *et al.* Methotrexate induces oxidative DNA damage and is selectively lethal to tumour cells with defects in the DNA mismatch repair gene *MSH2*. *EMBO Mol. Med.* **1,** 323–337 (2009).
38. Fong, P. C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from *BRCA* mutation carriers. *N. Engl. J. Med.* **361,** 123–134 (2009). **This article describes the results of the first clinical trial to demonstrate the clinical potential of a synthetic lethal approach for therapy.**
39. Balmaña, J., Domchek, S. M., Tutt, A. & Garber, J. E. Stumbling blocks on the path to personalized medicine in breast cancer: the case of PARP inhibitors for BRCA1/2-associated cancers. *Cancer Discov.* **1,** 29–34 (2011).
40. Mendeleyev, J., Kirsten, E., Hakam, A., Buki, K. G. & Kun, E. Potential chemotherapeutic activity of 4-iodo-3-nitrobenzamide. Metabolic reduction to the 3-nitroso derivative and induction of cell death in tumor cells in culture. *Biochem. Pharmacol.* **50,** 705–714 (1995).
41. Liu, X. et al. Iniparib non-selectively modifies cysteine-containing proteins in tumor cells and is not a bona fide PARP inhibitor. *Clin. Cancer Res.* advance ahead of print http://dx.doi.org/10.1158/1078-0432.CCR-11-1973 (29 November 2011).
42. McCabe, N. *et al.* Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Res.* **66,** 8109–8115 (2006).
43. Peasland, A. *et al.* Identification and evaluation of a potent novel ATR inhibitor, NU6027, in breast and ovarian cancer cell lines. *Br. J. Cancer* **105,** 372–381 (2011).
44. Murai, J. *et al.* The USP1/UAF1 complex promotes double-strand break repair through homologous recombination. *Mol. Cell. Biol.* **31,** 2462–2469 (2011).
45. Moskwa, P. *et al.* miR-182-mediated downregulation of *BRCA1* impacts DNA repair and sensitivity to PARP inhibitors. *Mol. Cell* **41,** 210–220 (2011).
46. Johnson, N. *et al.* Compromised CDK1 activity sensitizes BRCA-proficient cancers to PARP inhibition. *Nature Med.* **17,** 875–882 (2011).
47. Akamatsu, Y. & Jasin, M. Role for the mammalian Swi5–Sfr1 complex in DNA strand break repair through homologous recombination. *PLoS Genet.* **6,** e1001160 (2010).
48. Mendes-Pereira, A. M. *et al.* Synthetic lethal targeting of *PTEN* mutant cells with PARP inhibitors. *EMBO Mol. Med.* **1,** 315–322 (2009).
49. Brenner, J. C. *et al.* Mechanistic rationale for inhibition of poly(ADP-ribose) polymerase in *ETS* gene fusion-positive prostate cancer. *Cancer Cell* **19,** 664–678 (2011).

An important paper suggesting that *ETS* gene fusions may confer PARP-inhibitor sensitivity in prostate cancer.
50. Edwards, S. L. *et al.* Resistance to therapy caused by intragenic deletion in BRCA2. *Nature* **451,** 1111–1115 (2008).
51. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in *BRCA2*-mutated cancers. *Nature* **451,** 1116–1120 (2008).
References 50 and 51 are the first demonstrations of synthetic lethal resistance to PARP inhibitors and platinum drugs in *BRCA*-mutant cells.
52. Sakai, W. *et al.* Functional restoration of BRCA2 protein by secondary *BRCA2* mutations in *BRCA2*-mutated ovarian carcinoma. *Cancer Res.* **69,** 6381–6386 (2009).
53. Swisher, E. M. *et al.* Secondary *BRCA1* mutations in *BRCA1*-mutated ovarian carcinomas with platinum resistance. *Cancer Res.* **68,** 2581–2586 (2008).
54. Norquist, B. *et al.* Secondary somatic mutations restoring *BRCA1/2* predict chemotherapy resistance in hereditary ovarian carcinomas. *J. Clin. Oncol.* **29,** 3008–3015 (2011).
55. Rottenberg, S. *et al.* High sensitivity of BRCA1-deficient mammary tumors to the PARP inhibitor AZD2281 alone and in combination with platinum drugs. *Proc. Natl Acad. Sci. USA* **105,** 17079–17084 (2008).
56. Bouwman, P. *et al.* 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and *BRCA*-mutated breast cancers. *Nature Struct. Mol. Biol.* **17,** 688–695 (2010).
57. Bunting, S. F. *et al.* 53BP1 inhibits homologous recombination in *Brca1*-deficient cells by blocking resection of DNA breaks. *Cell* **141,** 243–254 (2010).
58. Jiang, H. *et al.* The combined status of ATM and *p53* link tumor development with therapeutic response. *Genes Dev.* **23,** 1895–1909 (2009).
59. Galanty, Y. *et al.* Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. *Nature* **462,** 935–939 (2009).
60. Soucy, T. A. *et al.* An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature* **458,** 732–736 (2009).
This paper provides proof of the concept that targeting additional elements of the DDR and replication machinery, such as neddylation complexes, could provide an alternative approach to cancer therapy.
61. Lansdorp, P. M. Immortal strands? Give me a break. *Cell* **129,** 1244–1247 (2007).
62. Diehn, M. *et al.* Association of reactive oxygen species levels and radioresistance in cancer stem cells. *Nature* **458,** 780–783 (2009).
63. Zhang, M., Atkinson, R. L. & Rosen, J. M. Selective targeting of radiation-resistant tumor-initiating cells. *Proc. Natl Acad. Sci. USA* **107,** 3522–3527 (2010).
64. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432,** 396–401 (2004).
65. Bao, S. *et al.* Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* **444,** 756–760 (2006).
66. Mantoni, T. S., Lunardi, S., Al-Assar, O., Masamune, A. & Brunner, T. B. Pancreatic stellate cells radioprotect pancreatic cancer cells through beta 1-integrin signaling. *Cancer Res.* **71,** 3453–3458 (2011).
67. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).
68. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009).
69. Barber, L. J. *et al.* Comprehensive genomic analysis of a *BRCA2* deficient human pancreatic cancer. *PLoS ONE* **6,** e21639 (2011).
70. Graeser, M. *et al.* A marker of homologous recombination predicts pathologic complete response to neoadjuvant chemotherapy in primary breast cancer. *Clin. Cancer Res.* **16,** 6159–6168 (2010).
71. Polo, S. E. & Jackson, S. P. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev.* **25,** 409–433 (2011).
72. Hoeijmakers, J. H. Genome maintenance mechanisms for preventing cancer. *Nature* **411,** 366–374 (2001).
73. Liu, P. *et al.* Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146,** 889–903 (2011).

Acknowledgements We thank Cancer Research UK, The Wellcome Trust, Breakthrough Breast Cancer, AACR, The Komen Foundation, The Breast Cancer Research Foundation, the European Union and The Breast Cancer Campaign for funding the work in our laboratory. We also acknowledge NHS funding to the Royal Marsden Hospital NIHR Biomedical Research Centre.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to C.J.L. (chris.lord@icr.ac.uk) or A.A. (alan.ashworth@icr.ac.uk)

REVIEW

doi:10.1038/nature10761

The promise of induced pluripotent stem cells in research and therapy

Daisy A. Robinton[1–5] & George Q. Daley[1–5]

The field of stem-cell biology has been catapulted forward by the startling development of reprogramming technology. The ability to restore pluripotency to somatic cells through the ectopic co-expression of reprogramming factors has created powerful new opportunities for modelling human diseases and offers hope for personalized regenerative cell therapies. While the field is racing ahead, some researchers are pausing to evaluate whether induced pluripotent stem cells are indeed the true equivalents of embryonic stem cells and whether subtle differences between these types of cell might affect their research applications and therapeutic potential.

After a decade of constraints, pluripotent stem-cell biology is now a flourishing research area, following the achievement of a long-standing ambition — the successful derivation of pluripotent stem cells from a patient's cells. In a momentous contribution, in 2006 Takahashi and Yamanaka illustrated how cell fates can be altered by the ectopic co-expression of transcription factors[1]. The manipulation of cell fates through reprogramming has altered fundamental ideas about the stability of cellular identity, stimulating major new directions in research into human disease modelling, tissue differentiation *in vitro* and cellular transdifferentiation. Despite heady progress, a major question remains: are the new induced pluripotent stem (iPS) cells equivalent to the classic embryonic stem (ES) cells and thus a suitable alternative for research and therapy? Whereas the initial wave of papers argued convincingly that the two cell types were functionally equivalent, a more refined analysis of how iPS cells behave *in vitro*, coupled with genome-wide genetic and epigenetic analysis, has revealed numerous subtle but substantial molecular differences, probably owing to technical limitations inherent in reprogramming. In this Review, we describe the derivation of iPS cells, outline the functional assessments of pluripotency, and then recount how global assessments of gene expression, gene copy number variation, DNA methylation and chromatin modification provide a more nuanced comparison of iPS cells and ES cells. We detail how these features influence the utility of each of these cell types for disease modelling and therapeutics, and offer predictions for the evolution of the art of reprogramming somatic cells.

Pluripotent stem cells

The years since Takahashi and Yamanaka's breakthrough have seen a flood of papers touting advances in reprogramming technology, including alternative methods for reprogramming and the successful derivation of iPS cells from various cell types. Although the field has advanced at a breathtaking pace, investigators have recently taken a step back to more critically evaluate iPS cells relative to ES cells and have endeavoured to fully understand how these cell populations differ from one another in the hope of closing the gap between the two populations. Taking clues from the data, it seems that researchers should attempt to define each cell type more accurately and to understand its inherent properties rather than ask whether these two classes of pluripotent cell are identical. Although ES cells and iPS cells are arguably equivalent in all their functions, these cells are bound to harbour subtle differences and to have distinct but complementary roles in research because of their distinct origins and modes of derivation. To appreciate the differences between ES cells and iPS cells, we must first define what it means to be pluripotent.

The term pluripotency has been assigned to a variety of cell types with a wide range of functional capacities. In its loosest sense, pluripotent describes a cell that can generate cell types from each of the three embryonic germ layers: the endoderm, mesoderm and ectoderm. At the strict end of the range of definitions, however, pluripotent describes a cell that can give rise to an entire organism, generating every cell type within that organism. The property of cell pluripotency was first exposed by Driesch in 1891, when he separated the two cells of an early sea urchin blastocyst and observed the development of two complete sea urchins[2]. Many decades later, studies of embryo aggregation and blastocyst chimaerism by Mintz and colleagues[3], Gardner[4] and Brinster[5], in the 1960s and 1970s, solidified the idea that the cells of the inner cell mass of the mouse blastocyst were pluripotent, and the isolation of mouse teratocarcinoma stem cells and native ES cells by Evans and Kaufman[6] and Martin[7], in 1981, ushered in the era of culturing pluripotent stem cells in a dish. The first successful isolation of human ES cells, by Thomson and colleagues in 1998, brought forth a surge of excitement in the scientific community and beyond[8]. The potential to understand early human development, tissue formation and differentiation *in vitro* through studying ES cells seemed to offer limitless possibilities. The opportunity to model diseases, discover disease mechanisms and, ultimately, use cell therapy for previously untreatable conditions was particularly alluring.

The derivation of ES cells from the human embryo, however, sparked controversy in the United States and led to a presidential executive order that restricted government funding[9]. The limited numbers of stem cell lines that were approved for research lacked the diversity necessary to address some of the most compelling questions, particularly those related to modelling and treating disease[10]. Most ES cells represented generic cells isolated from presumably normal embryos — except for those from embryos that had been tested by pre-implantation diagnostics and found to carry genetic

[1]Stem Cell Transplantation Program, Division of Pediatric Hematology/Oncology, Manton Center for Orphan Disease Research, Howard Hughes Medical Institute, Children's Hospital Boston and Dana Farber Cancer Institute, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. [2]Division of Hematology, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Broad Institute, Cambridge, Massachusetts 02142, USA. [5]Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA.

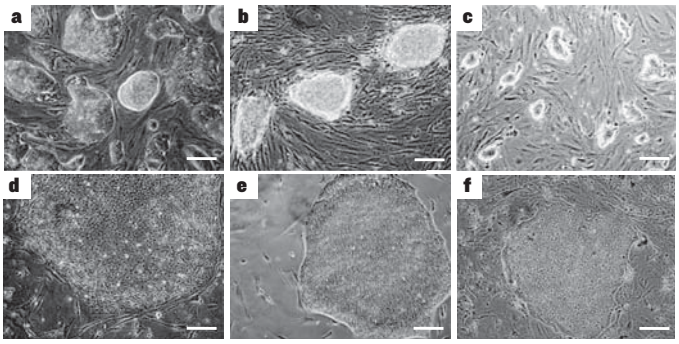


Figure 1 | Morphology of pluripotent stem cell types. Mouse ES (**a**) and iPS (**b**) cells form dome-shaped, refractile colonies. These colonies are in contrast to the flat morphology of mouse epiblast-derived stem cells (**f**), which resemble human ES (**d**) and iPS (**e**) cells. Human iPS cells induced into a naive pluripotent state by treatment with chemical inhibitors[97–100] (**c**) parallel the morphology of mouse ES and iPS cells. Scale bars, 50 μm.

diseases. The generic lines were not matched to a particular patient, so products derived from them for transplantation purposes would face rejection by the transplant recipient's immune system or necessitate that the recipient receive lifelong therapy with toxic immunosuppressive medication. To compound these limitations, when human ES cells are cultivated on mouse feeder cells, the human cells can incorporate mouse components that render the ES cells subject to immune rejection.

To realize the full potential of ES cells, researchers foresaw that customized, personalized pluripotent stem cells specific to each patient would be generated by using somatic-cell nuclear transfer (SCNT) — the procedure that had been used successfully to clone Dolly the sheep from adult mammary cells. Nuclear-transfer-generated ES (ntES) cell lines would capture a patient's complete genome in a cell that could be induced to become any tissue, thus allowing differentiation into disease-relevant cells for analysis or cell-replacement therapy. Despite successful proof of principle in mouse studies[11], and the clear distinctions between generating ntES cells for medical research and creating cloned blastocysts for reproduction, the ethical controversy driven by widespread opposition to human cloning has severely curtailed research into human SCNT. Only this year, when investigators gained access to a large number of human oocytes, was the derivation of pluripotent stem cells through human SCNT accomplished[12]. However, the investigators in this study needed to leave the oocyte nucleus intact to derive pluripotent stem cells, so the resultant cells were triploid, thus affording research applications for these cells but limiting their suitability for therapeutic use[12].

Despite the many hindrances to the study and derivation of human ES and ntES cells over the past decade, great strides were being made in understanding the pathways that regulate the maintenance and pluripotency of ES cells. This progress was not lost on those seeking an alternative source of personalized patient-specific stem cells, and in 2006 Takahashi and Yamanaka announced the successful derivation of iPS cells from adult mouse fibroblasts through the ectopic co-expression of only four genes[1]. In an elegant screen of 24 gene candidates selected for their links to ES-cell pluripotency, these researchers found four factors that were sufficient to reprogram adult fibroblasts into iPS cells: OCT4 (also known as POU5F1), SOX2, Krüppel-like factor 4 (KLF4) and c-MYC. This historic contribution inspired an astonishing flurry of follow-up studies, with successful reprogramming quickly translated to human fibroblasts[13–15] and then to a wide variety of other cell types, including pancreatic β cells[16], neural stem cells[17,18], mature B cells[19], stomach and liver cells[20], melanocytes[21], adipose stem cells[22] and keratinocytes[23], demonstrating the seemingly universal capacity to alter cellular identity.

Mouse and human iPS cells differ in appearance. Mouse iPS-cell colonies appear more dome-like and refractile than human iPS-cell colonies. Human iPS-cell colonies are flatter than those of mice and are akin to a distinct type of pluripotent stem cell that is derived from the epiblast of the early mouse embryo[24], a feature that indicates that mouse and human iPS cells, like mouse and human ES cells, probably reflect distinct developmental states (Fig. 1). The pluripotent state of mouse stem cells is called a 'naive' state because it closely resembles the most primitive state, or ground state, of the mouse inner cell mass; this is different from the more 'primed' state of human stem cells, which proliferate in response to different cytokines, reflecting the distinct developmental states of these populations[25]. Regardless of the method of derivation, iPS cells maintain the key features of ES cells, including the ability to propagate in culture indefinitely and the capacity to generate cells from each of the three embryonic germ layers (see ref. 26 for a review). Such broad similarities are not proof that iPS cells are molecularly or functionally equivalent to ES cells. Yamanaka's intention was to derive an alternative source of pluripotent stem cells with the same range of functions as ES cells but offering even greater potential for clinical use. To determine the degree of success garnered by reprogramming, we must explore the set of assays that were developed to assess the key characteristic of ES cells: pluripotency.

Assessment of pluripotency

In the past few years, consistent standards for the identification and evaluation of iPS cells and for the assessment of their functional equivalence to ES cells have become widely accepted[27]. A variety of reprogramming methods have been developed to derive iPS cells, and each has advantages and disadvantages (Table 1). Assessing reprogramming begins with identifying compact colonies that have distinct borders and well-defined edges, and are comprised of cells with a large nucleus, large nucleoli and scant cytoplasm. A wide range of colony morphologies result from reprogramming, and although many colonies appear morphologically similar to ES cells, only a subset of these have comparable molecular and functional features. To accurately distinguish reprogrammed, bona fide iPS cells from those that are only partially reprogrammed, investigators look for a series of molecular hallmarks.

Markers of pluripotency

Fully reprogrammed cells express a network of pluripotency genes, including *OCT4*, *SOX2* and *NANOG*, in levels comparable to ES cells, and they reactivate telomerase gene expression, downregulate *THY1* and upregulate *SSEA1* (ref. 28). Positive staining for alkaline phosphatase activity has been widely used as a marker of pluripotency; however, recently published data have shown this to be insufficient as a test for true iPS cells, because intermediately reprogrammed cells also stain positively[29]. The same report shows that iPS cells that are generated by virus-mediated reprogramming silence proviral genes when the endogenous pluripotency genes are activated, and that this event is paired with the expression of the embryonic antigens SSEA3, TRA-1-60, TRA-1-81, DNA methyltransferase 3β (DNMT3β) and REX1 (ref. 29). Genome-wide epigenetic reprogramming is crucial for deriving fully reprogrammed cells, and the degree of success is measured, in part, by evaluating the methylation status at the promoters of the genes responsible for maintaining pluripotency, as well as at the genes important for driving differentiation[30]. A crucial event during epigenetic reprogramming is the reactivation of the silent X chromosome, which occurs late in reprogramming and represents a hallmark of ground-state pluripotency[28,30,31]. If iPS cells acquire all of these molecular features, they are expected to behave like ES cells and to demonstrate reprogramming-factor independence, which is marked by silencing of the proviral transgenes. Variations in epigenetic reprogramming, the extent of methylation, the persistence of expression of integrated proviruses and other known and unknown factors can alter the differentiation potential of iPS cells. Because of the potential for heterogeneity, it is essential to know as much as possible about the nature of a cell line before labelling it pluripotent.

Table 1 | Methods for reprogramming somatic cells to iPS cells

Vector type		Cell types	Factors*	Efficiency (%)	Advantages	Disadvantages
Integrating	Retroviral[1,14,82,83]	Fibroblasts, neural stem cells, stomach cells, liver cells, keratinocytes, amniotic cells, blood cells and adipose cells	OSKM, OSK, OSK + VPA, or OS + VPA	~0.001–1	Reasonably efficient	Genomic integration, incomplete proviral silencing and slow kinetics
	Lentiviral[15,16,84,85]	Fibroblasts and keratinocytes	OSKM or *miR302/367* cluster + VPA	~0.1–1.1	Reasonably efficient and transduces dividing and non-dividing cells	Genomic integration and incomplete proviral silencing
	Inducible lentiviral[23,28]	Fibroblasts, β cells, keratinocytes, blood cells and melanocytes	OSKM or OSKMN	~0.1–2	Reasonably efficient and allows controlled expression of factors	Genomic integration and requirement for transactivator expression
Excisable	Transposon[86]	Fibroblasts	OSKM	~0.1	Reasonably efficient and no genomic integration	Labour-intensive screening of excised lines
	loxP-flanked lentiviral[87]	Fibroblasts	OSK	~0.1–1	Reasonably efficient and no genomic integration	Labour-intensive screening of excised lines, and *loxP* sites retained in the genome
Non-integrating	Adenoviral[88,89]	Fibroblasts and liver cells	OSKM	~0.001	No genomic integration	Low efficiency
	Plasmid[90,91]	Fibroblasts	OSNL	~0.001	Only occasional genomic integration	Low efficiency and occasional vector genomic integration
DNA free	Sendai virus[92]	Fibroblasts	OSKM	~1	No genomic integration	Sequence-sensitive RNA replicase, and difficulty in purging cells of replicating virus
	Protein[93,94]	Fibroblasts	OS	~0.001	No genomic integration, direct delivery of transcription factors and no DNA-related complications	Low efficiency, short half-life, and requirement for large quantities of pure proteins and multiple applications of protein
	Modified mRNA[95]	Fibroblasts	OSKM or OSKML + VPA	~1–4.4	No genomic integration, bypasses innate antiviral response, faster reprogramming kinetics, controllable and high efficiency	Requirement for multiple rounds of transfection
	MicroRNA[96]	Adipose stromal cells and dermal fibroblasts	miR-200c, miR-302s or miR-369s	~0.1	Efficient, faster reprogramming kinetics than commonly used lentiviral or retroviral vectors, no exogenous transcription factors and no risk of integration	Lower efficiency than other commonly used methods

*OSKM and similar factor names represent combinations of reprogramming factors: K, KLF4; L, LIN28; M, c-MYC; N, NANOG; O, OCT4; S, SOX2; and VPA, valproic acid.

Functional assays of pluripotency

When iPS cell lines are isolated and documented to carry the molecular features of fully reprogrammed cells, they are typically also assessed in functional assays. Characterization of the functional abilities of iPS cells begins with *in vitro* differentiation. The cells can be differentiated as embryoid bodies — compact balls of loosely organized tissues that resemble the gastrulating embryo — or through two-dimensional directed differentiation in a culture dish. Such cultures can then be assessed for markers of each of the three germ layers. Analysis of the pluripotency of mouse cells typically entails the development of a chimaera, which evaluates the potential of iPS cells to contribute to the normal development of adult tissues after injection into the blastocyst. Whether germline transmission occurs after blastocyst chimaerism is measured by the ability of chimaeras to produce all-iPS-cell mice in their offspring. These offspring have the genomic integrity of the injected iPS cell line, as well as the ability to form functional germ cells. The highest stringency test for mouse iPS cells — tetraploid complementation — entails the injection of iPS cells into tetraploid blastocysts to measure the ability of the iPS cells to direct the normal development of an entire organism. This test has been accomplished for only a limited subset of iPS cells[32–34], although with an efficiency that parallels tetraploid complementation carried out with ES or ntES cells[33,35,36].

The current functional gold standard for human iPS cells involves the evaluation of teratoma formation. In this assay, the *in vivo* differentiation potential of human iPS cells is measured after their injection subcutaneously or intramuscularly into immunodeficient mice[37,38]. If the cells are truly pluripotent, they will form well-differentiated tumours comprised of elements from each of the three germ layers. This assay provides information about the spontaneous differentiation potential of the injected iPS cells. Although it is the most stringent assay available for human iPS cells, it is not powerful enough to assess whether iPS cells can produce all the cell types of the human body, and it cannot assess the contribution of iPS cells to the germ line. The caveat to all these functional assays lies in the fact that the standards for iPS cells are still hotly debated, especially when anticipating the use of iPS cells for therapy[39]. Adopting a consistent set of standards that can be applied uniformly worldwide is essential as stem-cell research and applications move forward.

Functional differences between iPS cells and ES cells

Despite the multitude of assays used to evaluate pluripotency, and although many parallels have been found between iPS cells and ES cells, there is a wide range of evidence showing that there are subtle yet substantial differences between these cell types. Disparities were first observed in the differentiation abilities of iPS cells and ES cells in both teratoma-forming and *in vitro* differentiation assays. Some mouse iPS cells showed lower efficiencies of teratoma formation than mouse ES cells, whereas some human iPS cells showed less propensity

to differentiate along haematopoietic, neuroepithelial and neuronal lineages than human ES cells[40–42]. Some researchers interpreted these findings to mean that iPS cells have an intrinsically lower differentiation capacity than ES cells[41], whereas other research groups have offered different explanations, including that the cell of origin might have a specific effect on the differentiation capacities of the derived iPS cells.

The results from cell-of-origin studies indicate that the parental cell can influence the differentiation capacity of the resultant iPS cells. In one study, mouse bone-marrow-derived and B-cell-derived iPS cells showed more efficient differentiation along haematopoietic lineages than did fibroblast-derived iPS cells or neural-progenitor-derived iPS cell lines. Interestingly, treatment of the neural-progenitor-derived iPS cells with trichostatin A, a potent histone-deacetylase inhibitor, plus 5-azacytidine, a methylation-resistant cytosine analogue, increased the blood-forming capacity of these cells, suggesting that their limitations were due to epigenetic modifications. Whereas the bone-marrow-derived and neural-progenitor-derived iPS cells contributed well to all tissues in the chimaera assay, including to the germ line, the fibroblast-derived iPS cells contributed only poorly[43]. This study laid some of the early groundwork for later lines of investigation that probed the molecular differences between iPS cells and ES cells, and provided an explanation for the functional differences between these cells.

One investigation found that some iPS cells derived from human retinal-pigment epithelial cells show an increased propensity to differentiate back into this cell type than do ES cells or iPS cells derived from other tissues[44]. More recently, Bar-Nur and colleagues showed that iPS cells generated from human pancreatic islet β cells retain open chromatin at the loci of key β-cell genes and that this correlates with a greater capacity to differentiate into insulin-producing cells both *in vivo* and *in vitro* than that of ES cells or isogenic non-β-cell-derived iPS cells[45]. These functional differences extend beyond differentiation and potency to disease modelling. For example, fragile X syndrome is caused by aberrant silencing of the *FMR1* gene during human development; iPS cells that were reprogrammed from adult skin fibroblasts from an individual with fragile X syndrome failed to reactivate the *FMR1* gene, whereas ES cells derived from embryos with this syndrome, as diagnosed by pre-implantation testing, expressed *FMR1* (ref. 46). Consequently, the potential for epigenetic memory in the fragile-X-syndrome-derived iPS cells, and substantial differences between fragile-X-syndrome-derived iPS and ES cells, must be considered when studying this condition and potentially many other conditions. To determine whether the pluripotent cells being used are appropriate to address a particular question or to use in a given application, it is crucial to compare not only the *in vivo* and *in vitro* differentiation potentials but also the genetic and epigenetic disparities that underpin these functional differences.

iPS cells versus ES cells

Refined analyses, described in this section, have addressed whether iPS cells are suitable alternatives to classic ES cells for use in research and therapy.

Genetics and epigenetics

Global gene-expression analysis and bisulphite genomic sequencing accompanying early derivations of iPS cells provided the initial evidence for differences between iPS cells and ES cells at the epigenetic level[14]. Further exploration of these differences led to the identification of only a few, seemingly consistent, differences in global gene expression that were more pronounced in earlier passages of iPS cells[47]. Many of the differentially expressed genes were imprinted in ES cells[48].

Looking beyond the expression patterns to the DNA sequence itself has revealed genetic variation between iPS cells and ES cells. A recent publication suggested that chromosomal aberrations are a common feature of stem-cell populations that are propagated *in vitro*, with each type of stem cell — whether ES cell, iPS cell or multipotent stem cell — being prone to distinct abnormalities[49]. Both human iPS and ES cells showed a tendency for gains at chromosomes 12 and 17. Whereas iPS cells had additional gains at chromosomes 1 and 9, ES cells had additional gains at chromosomes 3 and 20. Other work identified an accumulation of point mutations in reprogrammed cells, particularly occurring in oncogenic pathways[45], whereas another study noted an increase in copy number variants (CNVs) in early-passage human iPS cells relative to intermediate-passage iPS cells or ES cells[50]. The number and size of these CNVs were negatively correlated with the passage number in iPS cells, suggesting that a selective disadvantage is conferred by these aberrations. A comparison of iPS cells and their parental cell of origin showed that the majority of CNVs were created *de novo* in fragile regions of the genome[50]. A comprehensive study by Laurent and colleagues found a higher frequency of subchromosomal CNVs in pluripotent cell samples than in non-pluripotent cell samples[51]. This work uncovered variation between genomic regions enriched for CNVs in human ES cells and iPS cells. A small subset of samples of ES cells harboured a large number of duplications, whereas several iPS-cell samples contained moderate numbers of deletions. Reprogramming was associated with deletions in tumour-suppressor genes, whereas extended time in culture led to duplications of oncogenes in human iPS cells.

The finding that human iPS cells derived from a variety of tissues have residual, persistent donor-cell-specific gene-expression patterns sparks the question of whether the current measure of a fully reprogrammed cell is sufficient[52] or whether iPS cells retain some type of 'somatic memory' from their past identity. To understand this observation better, a more detailed analysis at the epigenetic level is required.

Reprogramming cells to a pluripotent state entails global epigenetic remodelling and introduces epigenetic changes, some of which are necessary for reprogramming to occur and others of which are inadvertently introduced during the process. A failure to demethylate pluripotency genes is associated with partial reprogramming in iPS cells. Whole-genome profiling of the DNA methylomes of multiple human iPS and ES cell lines, as well as somatic and progenitor cell lines, from different laboratories using different reprogramming techniques and with a variety of cells derived from distinct germ layers has shown that although overall iPS-cell DNA methylomes closely resemble human ES-cell DNA methylomes, iPS cells have significant variability in their somatic memory, as well as aberrant iPS-cell-specific differential methylation. Some studies have suggested that this occurs in a passage-dependent manner, but others have shown that differentially methylated regions (DMRs) in iPS cells are transmitted to differentiated progeny at a high frequency and cannot be erased through passaging[29,53–55]. Overall, there are remarkable global similarities between the DNA methylomes of generic iPS and ES cells; however, a core set of DMRs that seems to represent hot spots of failed epigenetic reprogramming has been identified[55]. These DMRs are enriched for genes that are important for developmental processes[51,55]. The high incidence of unique DMRs in iPS cells compared with progenitor somatic cells or ES cells suggests that these patterns are stochastic and arise during reprogramming. In the most exhaustive comparison so far, Kim and colleagues reported that more DMRs were present in mouse iPS cells than in ntES or embryo-derived ES cells[43]. However, these DMRs did not pertain to specific loci and thus do not represent consistent differences between iPS cells and ES cells. This lack of consistency suggests that aberrant DMRs in mouse iPS cells reflect the technical limitations inherent in reprogramming, rather than indicating loci that can reliably distinguish ES cells from iPS cells[43].

In addition, the residual iPS-cell-specific methylation in many iPS-cell isolates links these cells to their tissue of origin and, ultimately, affects their differentiation propensity[43,55]. Residual signatures can be distinct enough to enable the myeloid and lymphoid origins of blood-derived iPS cells to be discerned[43]. In iPS cells derived from non-haematopoietic cells, such as fibroblasts and neural progenitors, there can be residual repressive methylation at loci that are required for haematopoietic fates, reducing the blood-forming potential *in vitro*[43]. Exogenous supplementation of neural-progenitor-derived iPS cells with the cytokine WNT3A can increase the blood-forming potential of these cells, supporting the idea that incomplete reprogramming owing to epigenetic marks can be

overcome by manipulating the culture conditions. Treatment of cultures with demethylating agents or knockdown of *DNMT1* expression has been shown to convert intermediately reprogrammed cells into fully pluripotent cells, further supporting this idea[48]. When iPS cells are forced to differentiate along a particular lineage, they become more amenable to generating cells of that lineage after another round of reprogramming[43]. This finding shows that the differentiation propensity and DNA methylation profile can be reset, and it suggests that the 'epigenetic memory' of the donor cell can be exploited, especially in cases in which directed differentiation is particularly challenging[43].

Another important feature of epigenetic reprogramming is the reactivation of the inactivated X chromosome. During normal development in eutherian mammals (those with a placenta), one X chromosome is randomly inactivated in each cell in females. Whether this epigenetic silencing event is reset in iPS cells remains an area of controversy, in part because of the poor fidelity of X-inactivation markers in pluripotent cells[56]. Some studies have shown that the majority of female human iPS clones retain an inactivated X chromosome (which is transcriptionally silent)[57], whereas others have indicated that some human iPS clones lose immunostaining for trimethylated H3K27 on the X chromosome (a marker of epigenetic silencing), indicating X reactivation[58]. In addition, some of the earliest studies of iPS cells showed X reactivation in reprogrammed female mouse fibroblasts[31,58]. However, recently published data support the finding that X reactivation does not occur in human iPS cells and, interestingly, reprogramming was found to favour expression of a particular X chromosome when induced from a mixed X-inactivated population of fibroblasts[59]. Finally, epigenetic reprogramming sometimes fails to properly restore bivalent domains, which mark developmental loci with active and repressive histone modifications[60].

Although many of the studies cited here have generated data suggesting that there are epigenetic differences between iPS cells and ES cells, there are several limitations on extending these data to all iPS cells in a more general (and more useful) sense. The published comparisons were often made using iPS cells derived from a multitude of laboratories by a variety of methodologies, and reanalysis of the gene-expression microarray data using an unsupervised clustering algorithm shows a strong correlation between transcriptional signatures and specific laboratories for both iPS cells and ES cells. This finding indicates that specific culture protocols and laboratory environments can affect the transcriptional profile of iPS and ES cells. Therefore, the data produced in a particular laboratory might be specific to the cells derived there[61].

In addition, most iPS colonies are clones derived from a single reprogrammed cell, whereas ES cells used for analysis are typically non-clonal. The subcloning of ES cells has revealed genetic and epigenetic anomalies that would probably have otherwise gone undetected in the heterogeneous ES-cell population[62]. With regard to somatic memory, there is poor overlap between the gene sets that have been reported to be characteristic of a particular cell type of origin, suggesting that the retention of somatic memory is stochastic and is a reflection of the technical failure of reprogramming to fully erase the somatic epigenome. To exacerbate the issue, the iPS cells used for comparison often have different genetic backgrounds and have frequently been derived from fibroblasts that were already heterogeneous in their make-up, affecting both the gene-expression patterns and the functionality of the iPS cells.

Throughout the literature, many publications lack correlation between the gene-expression patterns and the epigenetic patterns observed. An additional consideration is the presence of different viral insertions in individual iPS cell lines, which can also affect the functionality of the derived cells[1]. Evidence to support this idea is provided by the reduced number of differences observed among iPS cells and between iPS and ES cells when transgenes are removed[63]. Many of the aforementioned studies have focused on differences in either transcriptional profiles or changes in epigenetic marks; however, the most recent studies have evaluated iPS cells and ES cells from both of these angles in parallel, together with their *in vitro* differentiation potential, generating the most comprehensive and compelling data that have been published so far.

Holistic analysis

Stadtfeld and colleagues explored the epigenetic and functional discrepancies between iPS cells and ES cells using a new reprogramming strategy that allowed direct comparison of genetically matched cells derived from the same source[32]. These authors derived iPS and ES cells from mice carrying an integrated doxycycline-inducible reprogramming cassette in every cell, a strategy that sidesteps the confounding effects of variable genetic backgrounds and viral integration that have been observed in other studies. The overall messenger RNA and microRNA expression patterns of iPS cells and ES cells were indistinguishable except for the aberrant silencing of a few transcripts localized to the imprinted *Dlk1–Dio3* gene cluster on chromosome 12qF1, a region that is important for development. A failure to reactivate this locus meant that iPS cells contributed poorly to chimaeras and were unable to generate all-iPS-cell mice. By contrast, iPS cells with normal *Dlk1–Dio3* expression contributed to high-grade chimaeras and supported the development of viable all-iPS-cell mice. The treatment of iPS cells that failed to reactivate *Dlk1–Dio3* with a histone-deacetylase inhibitor rescued the ability of these clones to support the development of all-iPS-cell mice by relieving this region of aberrant hypermethylation. However, recent data from iPS and ES cells derived from a mouse strain carrying a distinct reprogramming cassette suggest that different levels of expression of reprogramming factors, rather than aberrant silencing of *Dlk1–Dio3*, account for the different behaviour of the cell lines[64]. The disparate results from these studies highlight that iPS cells can behave differently based on subtle variations in the expression of only a few loci.

To systematically compare human iPS cells derived from different somatic cell types and ES cells, Ohi and colleagues compared ES cells with iPS cells reprogrammed from somatic cells representative of the three embryonic germ layers[65]. Transcriptional and epigenetic profiling of these cells showed transcriptional differences, owing, in part, to incomplete promoter methylation, which enabled iPS cells to be discerned on the basis of their cell of origin. The differential methylation between iPS cells and ES cells did not correlate with varying levels in DNA methyltransferases; however, the authors found a nonrandom pattern of incompletely silenced genes in genetic regions that are isolated from other genes that undergo silencing during reprogramming. This finding could be explained by inefficient or delayed recruitment of the silencing machinery and DNA methyltransferases to particular somatic genes because of the isolated location of these genes[65].

In another comprehensive study, Polo and colleagues evaluated the effect of cellular origin on the gene-expression pattern, epigenetic properties and functional abilities of genetically matched mouse iPS cells[66]. Using the same 'secondary' reprogramming strategy used by Wernig and colleagues[53], whereby reprogramming is assessed using tissues from a mouse generated from iPS cells carrying integrated, doxycycline-inducible reprogramming factors, Polo and colleagues generated iPS cells from tail-tip fibroblasts, splenic B cells, bone-marrow-derived granulocytes and skeletal muscle precursors, and showed that each iPS cell line retained a transcriptional memory of its cell of origin. This memory was evident in that markers for each respective cell of origin remained actively expressed, and was supported by the finding that iPS cell lines derived from the same cell of origin clustered together on the basis of global transcriptional data. A similar correlation was found on evaluation of methylation patterns, which showed subtle but substantial differences, reflecting the consequences of different histone marks. The effects of somatic memory extended beyond the genetic and epigenetic levels to functional significance, affecting the autonomous differentiation potential of the different iPS cell lines after embryoid-body formation. A clear bias that reflected the cells of origin was observed in the iPS cell lines. Notably, the transcriptional, epigenetic and functional effects evaluated in early-passage iPS cell lines became

less significant with continued passaging. This finding indicated that complete reprogramming is a gradual process that extends beyond the time frame necessary to observe the activation of endogenous pluripotency genes, transgene-free growth and differentiation into cell types from each of the three germ layers.

Exploring equivalence

Having considered data from a multitude of published studies, generated by the painstaking efforts of many research groups, we return to our earlier question: are iPS cells equivalent to ES cells? The answer is not straightforward. Rather, there is an emerging consensus that iPS cells and ES cells are neither identical nor distinct populations. Instead, they are overlapping, with greater variability inherent within each population than between the populations. The heterogeneity and behaviour of each class of cells is more complex than has previously been thought. The two pluripotent stem-cell types are, in theory, functionally equivalent; however, in practice, they harbour genetic and epigenetic differences that reflect their different histories. It remains to be seen whether there are any consistent molecular distinctions between iPS cells and ES cells.

It is also important to consider that, in contrast to long-standing belief, ES cells themselves have considerable epigenetic heterogeneity and have differing propensities for differentiation — much like those found in iPS cells[67,68]. These observations, paired with those discussed in this Review, are a call for researchers to take a step back from the direct comparison of iPS cells and ES cells, and they highlight the need to redefine what it means to belong to either of these cell classes. Some researchers have already taken heed of this message and generated a bioinformatics assay for pluripotency[63], whereas others have produced a 'scorecard' to evaluate the character of both iPS and ES cell lines and predict the quality and utility of any pluripotent cell line in a high-throughput manner[47,69]. Using DNA methylation mapping, gene-expression profiling and a quantitative differentiation assay, Bock and colleagues made a systematic comparison of 20 established ES cell lines and 12 iPS cell lines[47]. They confirmed that, despite overall similarities, transcriptional and epigenetic variation is common between iPS cell lines, between ES cell lines, and between iPS cell and ES cell populations. These data provide a reference for the variation among human pluripotent cell lines, which assists in predicting the functional consequences of these differences. We can conclude from these studies that any given iPS cell line generated by today's technology might not be completely equivalent to the ideal ES cell.

The differences between iPS cells and ES cells, as well as those among iPS cells, clearly affect the utility of these cells in research, disease modelling and therapeutics, providing an impetus for investigators to evaluate their cell populations carefully and precisely. The differences do not diminish the potential of iPS cells, given that iPS cells have considerable advantages over ES cells. Rather than replacing ES cells with iPS cells, it is becoming clear that these two cell types complement one another. Researchers are still in the process of developing the necessary protocols to harness the potential of iPS cells; however, as it becomes clear how to evaluate the genetic, epigenetic and functional status of different iPS cell lines, further applications of these cells will be uncovered, and progress will be made in creating iPS cell lines and designing protocols to accomplish the ambitious goals of the field.

Medical applications of iPS cells

Generating patient-specific stem cells has been a long-standing goal in the field of regenerative medicine. Despite considerable challenges, generating disease-specific and patient-specific iPS cells through reprogramming has become almost routine. These cells provide a unique platform from which to gain mechanistic insight into a variety of diseases, to carry out *in vitro* drug screening, to evaluate potential therapeutics and to explore gene repair coupled with cell-replacement therapy (Fig. 2). In the past few years, the number of reports on applications of iPS cells has steadily increased, testifying to the broad influence of this breakthrough technology (Table 2). Despite the continued presence of substantial hurdles, the pace of this work is such that no review can capture the current state of the field; thus, we point to a few publications that highlight the promising medical applications of iPS cells but also indicate their key limitations.

In 2009, Lee and colleagues harnessed iPS cells to demonstrate disease modelling and drug screening for familial dysautonomia, a rare genetic disorder of the peripheral nervous system[70]. In almost all cases, familial dysautonomia is caused by a single point mutation in the gene encoding the inhibitor of nuclear factor-κB (IκB)-kinase-complex-associated protein (*IKBKAP*) that manifests as an extensive autonomic nervous system deficit and dysfunction in small-fibre sensory neurons. Although many traditional cell-based models have been used to study the pathogenesis of familial dysautonomia and to screen for candidate drugs, none has used symptom-relevant human cell types. With the successful derivation of iPS cells from patients with familial dysautonomia, investigators produced central and peripheral nervous system precursors and subsequently found three disease-related phenotypes, thus providing validation that disease-relevant cell types could accurately reflect disease pathogenesis *in vitro*[70]. After screening with multiple compounds, the authors showed that the disease phenotype could be partially normalized by kinetin, a plant hormone. This initial report demonstrated how iPS cells can facilitate the discovery of therapeutic compounds and described how these cells provided a platform for modelling different severities of familial dysautonomia and for generating predictive tests to determine differences in the clinical manifestation of the disorder.

Such applications of iPS cells in drug screening and discovery are destined to expand to encompass numerous disease conditions. Several research groups have generated models of long QT syndrome, a congenital disease with 12 types, each of which is associated with abnormal ion-channel function, a prolonged QT interval on an electrocardiogram and a high risk of sudden cardiac death due to ventricular tachyarrhythmia. Much work has been carried out in animal models to probe the underlying mechanisms of this syndrome, but cardiomyocytes have distinct and complex electrophysiological properties that differ between species. In addition, the lack of *in vitro* sources of human cardiomyocytes and the inability to model patient-specific variations of this disease has impeded studies.

In a proof-of-principle study for using iPS cells to capture the physiological mechanisms of genetic variation, Moretti and colleagues differentiated iPS cells from individuals with type 1 long QT syndrome into cardiomyocytes and, as predicted, observed prolonged action potentials in the ventricular and atrial cells[71]. Using this model system, these investigators uncovered a dominant-negative trafficking defect associated with the particular mutation that causes this variant of long QT syndrome. Further investigation of long QT syndrome iPS-cell-derived cardiomyocytes showed that these cells had an increased susceptibility to catecholamine-induced tachyarrhythmia, and compounds that exacerbated the condition (including isoprenaline) were identified[71]. Treatment of these cardiomyocytes with β-adrenergic receptor blockers attenuated the long QT phenotype.

Type 2 long QT syndrome has also been modelled in cardiomyocytes, by Itzhaki and colleagues[72]. The authors derived type 2 long QT syndrome iPS cells to evaluate the potency of existing and new pharmacological agents that might exacerbate or ameliorate the condition. Their studies show that the long QT syndrome phenotype was aggravated by blockers of ERG-type potassium channels, whereas nifedipine, a calcium-channel blocker, and pinacidil, an agonist of ATP-sensitive potassium channels, both ameliorated the long QT syndrome phenotype, as shown by the decreased duration of action potentials in long QT syndrome cardiomyocytes, as well as the elimination of early after-depolarizations and the abolishment of all triggered arrhythmias. A possible limitation of these beneficial drugs is excessive shortening of the action-potential duration, leading to short QT syndrome.

Importantly, these studies established that the iPS-cell model can be used to identify complex cardiotoxic effects of drugs, as well as to define protective pharmacological agents, including optimal drug dosages.

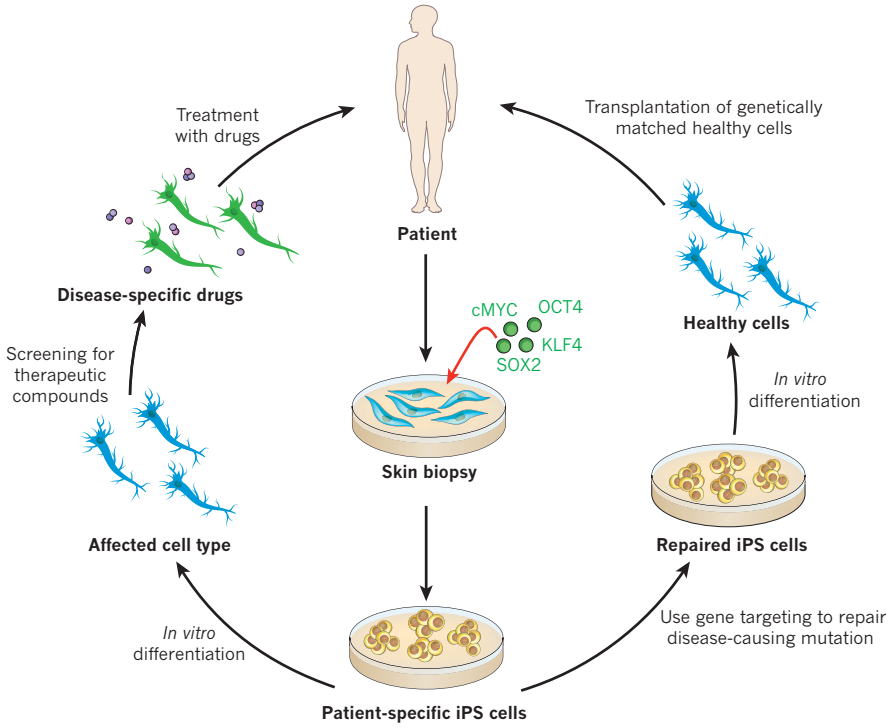


Figure 2 | Medical applications of iPS cells. Reprogramming technology and iPS cells have the potential to be used to model and treat human disease. In this example, the patient has a neurodegenerative disorder. Patient-specific iPS cells — in this case derived by ectopic co-expression of transcription factors in cells isolated from a skin biopsy — can be used in one of two pathways. In cases in which the disease-causing mutation is known (for example, familial Parkinson's disease), gene targeting could be used to repair the DNA sequence (right). The gene-corrected patient-specific iPS cells would then undergo directed differentiation into the affected neuronal subtype (for example, midbrain dopaminergic neurons) and be transplanted into the patient's brain (to engraft the nigrostriatal axis). Alternatively, directed differentiation of the patient-specific iPS cells into the affected neuronal subtype (left) will allow the patient's disease to be modelled *in vitro*, and potential drugs can be screened, aiding in the discovery of novel therapeutic compounds.

Given the number of drugs that have notoriously been withdrawn from the market because of their tendency to induce arrhythmias, it is highly likely that the current inadequate approaches for assessing cardiotoxicity will be complemented by iPS-cell-based assessments of drug effects.

A study from our laboratory explored dyskeratosis congenita, a disorder of telomere maintenance, and provided an unanticipated insight into the basic biology of telomerase that has therapeutic implications[73]. In its most severe form, dyskeratosis congenita is caused by a mutation in the dyskerin gene (*DKC1*), which is X linked, leading to shortened telomeres and premature senescence in cells and ultimately manifesting as the degeneration of multiple tissues. Because the reprogramming of cells to an induced pluripotent state is accompanied by the induction of the gene encoding telomerase reverse transcriptase (*TERT*), we investigated whether the telomerase defect would limit the derivation and maintenance of iPS cells from individuals with dyskeratosis congenita. Although the efficiency of iPS-cell derivation was poor, we were able to successfully reprogram patient fibroblasts. Surprisingly, whereas the mean telomere length immediately after reprogramming was shorter than that of the parental fibroblast population, continued passage of some iPS cell lines led to telomere elongation over time. This process was accompanied by upregulation of the expression of *TERC*, which encodes the RNA subunit of telomerase.

Further analysis established that TERT and TERC, as well as DKC1, were expressed at higher levels in dyskeratosis-congenita-derived iPS cells than in the parental fibroblasts[73]. We determined that the genes encoding these components of the telomerase pathway — including a *cis* element in the 3′ region of the *TERC* locus that is essential for a transcriptionally active chromatin structure — were direct binding targets of the pluripotency-associated transcription factors. Further analysis indicated that transcriptional silencing owing to a 3′ deletion in the *TERC* locus leads to the autosomal dominant form of dyskeratosis congenita by diminishing *TERC* transcription. Although telomere length is restored in dyskeratosis-congenita-derived iPS cells, differentiation into somatic cells is accompanied by a return to pathogenesis with low *TERC* expression and a decay in telomere length. This finding showed that TERC RNA levels are dynamically regulated and that the pluripotent state of the cells is reversible, suggesting that drugs that elevate or stabilize *TERC* expression might rescue defective telomerase activity and provide a therapeutic benefit. Although we set out to understand the pathogenesis of dyskeratosis congenita with this study, we showed that a high expression level of multiple telomerase components was characteristic of the pluripotent state more generally, illustrating how iPS cells can reveal fundamental aspects of cell biology.

An independent study of the reprogramming of cells from patients with dyskeratosis congenita confirmed the general transcriptional upregulation of multiple telomerase components and the maintenance of telomere lengths in clones[74]; however, in this study, no clones with elongated telomeres were identified. The different outcomes of these studies highlight the limitations of iPS-cell-based disease models that are imposed by clonal variation as a result of the inherent technical infidelity of reprogramming[75]. This point also introduces an additional important consideration. Before a given iPS-cell disease model can be claimed to be truly representative of the disease, how many patients must be involved, and how many iPS cell lines must be derived from each patient? Although the answers to these questions are unclear, it is crucial to keep these issues in mind when generating disease models and making claims based on these models.

Although iPS cells are an invaluable tool for modelling diseases *in vitro*, the goal of developing patient-specific stem cells has also been motivated by the prospect of generating a ready supply of immune-compatible cells and tissues for autologous transplantation. At present, the clinical translation of iPS-cell-based cell therapies seems more futuristic than the *in vitro* use of iPS cells for research and drug development, but two groundbreaking studies have provided the proof of principle in mouse models that the dream might one day be realized. Hanna, Jaenisch and colleagues used homologous recombination to repair the genetic defect in iPS cells derived from a humanized mouse model of sickle-cell anaemia[76]. Directed differentiation of the repaired iPS cells into haematopoietic progenitors followed by transplantation of these cells into the affected mice led to the rescue of the disease phenotype. The gene-corrected iPS-cell-derived haematopoietic progenitors showed stable engraftment and correction of the disease phenotype.

In another landmark study from Jaenisch's research group, Wernig and colleagues derived dopaminergic neurons from iPS cells that, when implanted into the brain, became functionally integrated and improved the condition of a rat model of Parkinson's disease[77]. The successful implantation and functional recovery in this model is evidence of the therapeutic value of pluripotent stem cells for cell-replacement therapy in the brain — one of the most promising areas for the future of iPS-cell applications.

Table 2 | Diseases modelled with iPS cells

Disease	Molecular defect of donor cell	Cell type differentiated from iPS cells	Disease phenocopied in differentiated cells	Drug or functional tests
Neurological				
Amyotrophic lateral sclerosis (ALS)	Heterozygous Leu144Phe mutation in SOD1	Motor neurons and glial cells	ND	No
Spinal muscular atrophy (SMA)	Mutations in *SMN1*	Neurons and astrocytes, and mature motor neurons	Yes	Yes
Parkinson's disease	Multifactorial; mutations in *LRRK2* and/or *SNCA*	Dopaminergic neurons	No	Yes
Huntington's disease	72 CAG repeats in the huntingtin gene	None	NA	No
Down's syndrome	Trisomy 21	Teratoma with tissue from each of the three germ layers	Yes	No
Fragile X syndrome	CGG triplet repeat expansion resulting in the silencing of *FMR1*	None	NA	No
Familial dysautonomia	Mutation in *IKBKAP*	Central nervous-system lineage, peripheral neurons, haematopoietic cells, endothelial cells and endodermal cells	Yes	Yes
Rett's syndrome	Heterozygous mutation in *MECP2*	Neural progenitor cells	Yes	Yes
Mucopolysaccharidosis type IIIB (MPS IIIB)	Homozygous mutation in *NAGLU*	Neural stem cells and differentiated neurons	Partially	Yes
Schizophrenia	Complex trait	Neurons	Yes	Yes
X-linked adrenoleukodystrophy (X-ALD), childhood cerebral ALD (CCALD) and adrenomyeloneuropathy (AMN)	Mutation in *ABCD1*	Oligodendrocytes and neurons	Partially	Yes
Haematological				
ADA SCID	Mutation or deletion in *ADA*	None	ND	No
Fanconi's anaemia	*FAA* and *FAD2* corrected	Haematopoietic cells	No (corrected)	No
Schwachman–Bodian–Diamond syndrome	Multifactorial	None	NA	No
Sickle-cell anaemia	Homozygous HbS mutation	None	NA	No
β-Thalassaemia	Homozygous deletion in the β-globin gene	Haematopoietic cells	ND	No
Polycythaemia vera	Heterozygous Val617Phe mutation in *JAK2*	Haematopoietic progenitors ($CD34^+CD35^+$)	Partially	No
Primary myelofibrosis	Heterozygous mutation in *JAK2*	None	NA	No
Metabolic				
Lesch–Nyhan syndrome (carrier)	Heterozygous mutation in *HPRT1*	None	NA	No
Type 1 diabetes	Multifactorial; unknown	β-Cell-like cells (express somatostatin, glucagon and insulin; glucose-responsive)	ND	No
Gaucher's disease, type III	Mutation in *GBA*	None	NA	No
α1-Antitrypsin deficiency (A1ATD)	Homozygous mutation in the α1-antitrypsin gene	Hepatocyte-like cells (fetal)	Yes	No

Together, these findings provide proof of principle for using reprogramming with gene repair and cell-replacement therapy for treating diseases. Using iPS cells in cell-replacement therapy offers the promise of therapeutic intervention that is not compounded by the use of immunosuppressive drugs to prevent tissue rejection, while harnessing targeted gene-repair strategies, such as homologous recombination and zinc-finger nucleases, to repair genetic defects. These strategies provide the opportunity for generating an unlimited population of stem cells that can be differentiated into the desired cell type for studying disease mechanisms, for screening and developing drugs or for developing a suitable cell-replacement therapy. There have been considerable advances and successes to this end; however, selecting an appropriate disease target, directing the differentiation of iPS cells into phenotype-relevant cell populations and identifying disease-relevant phenotypes remain major hurdles. It is unclear whether iPS cells used for cell-replacement therapy would completely evade an immune response when returned to the patient, because a recent study has shown the immune rejection of teratomas formed from iPS cells, even in syngeneic mice[78]. Nevertheless, iPS cells provide a promising model with which to study disease mechanisms, discover new therapies and develop truly personalized treatments.

Predictions for the evolution of the art

Few fields have enjoyed the remarkable upsurge in activity and excitement that followed the initial report of the reprogramming of somatic cells into iPS cells in 2006. Despite heady progress, crucial challenges must be met for the field to realize its full potential. There is as yet no consensus on the most consistent or optimal protocol for deriving the most reliable and, ultimately, the safest iPS cells. Increasing the reprogramming efficiency and effecting reprogramming without genetically modifying the cells are goals that have been achieved. Using more-uniform protocols and more-rigorous controls would facilitate experimental and potentially therapeutic consistency

Disease	Molecular defect of donor cell	Cell type differentiated from iPS cells	Disease phenocopied in differentiated cells	Drug or functional tests
Metabolic cont.				
Glycogen storage disease Ia (GSD1a)	Defect in glucose-6-phosphate gene	Hepatocyte-like cells (fetal)	Yes	No
Familial hypercholesterolaemia	Autosomal dominant mutation in *LDLR*	Hepatocyte-like cells (fetal)	Yes	No
Crigler–Najjar syndrome	Deletion in *UGT1A1*	Hepatocyte-like cells (fetal)	ND	No
Hereditary tyrosinaemia, type 1	Mutation in *FAHD1*	Hepatocyte-like cells (fetal)	ND	No
Pompe disease	Knockout of *GAA*	Skeletal muscle cells	Yes	No
Progressive familial cholestasis	Multifactorial	Hepatocyte-like cells (fetal)	ND	No
Hurler syndrome (MPS IH)	Genetic defect in *IDUA*	Haematopoietic cells	No	No
Cardiovascular				
LEOPARD syndrome	Heterozygous mutation in *PTPN11*	Cardiomyocytes	Yes	No
Type 1 long QT syndrome	Dominant mutation in *KCNQ1*	Cardiomyocytes	Yes	No
Type 2 long QT syndrome	Missense mutation in *KCNH2*	Cardiomyocytes	Yes	Yes
Primary immunodeficiency				
SCID or leaky SCID	Mutation in *RAG1*	None	NA	No
Omenn syndrome (OS)	Mutation in *RAG1*	None	NA	No
Cartilage-hair hypoplasia (CHH)	Mutation in *RMRP*	None	NA	No
Herpes simplex encephalitis (HSE)	Mutation in *STAT1* or *TLR3*	Mature cell types of the central nervous system	No	No
Other category				
Duchenne muscular dystrophy	Deletion in the dystrophin gene	None	NA	No
Becker muscular dystrophy	Unidentified mutation in dystrophin	None	NA	No
Dyskeratosis congenita (DC)	Deletion in *DKC1*	None	NA	No
Cystic fibrosis	Homozygous deletion in *CFTR*	None	NA	No
Friedreich's ataxia (FRDA)	Trinucleotide GAA repeat expansion in *FXN*	Sensory and peripheral neurons, and cardiomyocytes	Partially	No
Retinitis pigmentosa	Heterogeneity in causative genes and mutations: mutations in *RP9, RP1, PRPH2* or *RHO*	Retinal progenitors, photoreceptor precursors, retinal-pigment epithelial cells and rod photoreceptor cells	Yes	Yes
Recessive dystrophic epidermolysis bullosa (RDEB)	Mutation in *COL7A1*	Haematopoietic cells, and epidermis-like keratinocytes that differentiate into cells of all three germ layers *in vivo*	Partially	Yes
Scleroderma	Unknown	None	NA	No
Osteogenesis imperfecta	Mutation in *COL1A2*	None	NA	No

An extended version of this table includes references and more information about drug and functional tests (Supplementary Table 1). *ABCD1*, ATP-binding cassette, sub-family D, member 1; *ADA*, adenine deaminase; *CFTR*, cystic fibrosis transmembrane conductance regulator; *COL1A2*, α2-chain of type I collagen; *COL7A1*, α1-chain of type VII collagen; *DKC1*, dyskerin; *FAA*, Fanconi's anaemia, complementation group A; *FAD2*, Fanconi's anaemia, complementation group D2; *FAHD1*, fumarylacetoacetate hydrolase; *FMR1*, fragile X mental retardation 1; *FXN*, frataxin; *GAA*, acid α-glucosidase; *GBA*, acid β-glucosidase; HbS, sickle haemoglobin; *HPRT1*, hypoxanthine phosphoribosyltransferase 1; *IDUA*, α-L-iduronidase; *JAK2*, Janus kinase 2; *KCNH2*, potassium voltage-gated channel, subfamily H (eag-related), member 2; *KCNQ1*, potassium voltage-gated channel, KQT-like subfamily, member 1; LDLR, low-density lipoprotein receptor; *LRRK2*, leucine-rich repeat kinase 2; *MECP2*, methyl CpG binding protein 2; NA, not applicable; *NAGLU*, α-N-acetylglucosaminidase; ND not determined; *PRPH2*, peripherin 2; *PTPN11*, protein tyrosine phosphatase, non-receptor type 11; *RAG1*, recombination activating gene 1; *RHO*, rhodopsin; *RMRP*, RNA component of mitochondrial-RNA-processing endoribonuclease; *RP*, retinitis pigmentosa; SCID, severe combined immunodeficiency; *SMN1*, survival of motor neuron 1; *SNCA*, α-synuclein; SOD1, superoxide dismutase 1; *STAT1*, signal transducer and activator of transcription 1; *TLR3*, Toll-like receptor 3; *UGT1A1*, UDP glucuronosyltransferase 1 family, polypeptide A1.

between laboratories and would yield standardized cell lines that could be used with confidence in both basic and applied studies.

Barring that, researchers must agree on standards of molecular analysis to ensure that the reprogrammed cells that most closely approximate the generic state of the naive genome can be identified. Because iPS cells are subject to the same type of culture adaptations that affect karyotypic integrity as human ES cells[79], it is important to define protocols that minimize the time in culture. In addition, cell lines used in clinical applications will need to be evaluated frequently for aberrant culture-induced changes at all stages: from the somatic cells to the reprogrammed and differentiated cells[80]. Understanding the genomic alterations that take place during the reprogramming, culture and differentiation of iPS cells will be crucial for designing experiments and ensuring that the derived cells are functional, pure and appropriate for use in research and therapy. Minimizing any aberrations is important, but as long as researchers understand that aberrations will arise — and can describe and control their effects — even imperfect cells can be used, and preferential differentiation can be taken advantage of whenever possible. Characteristics of iPS cells that were initially perceived as flaws, including varying differentiation propensities, might prove useful in clinical settings to generate cell types that have been difficult to obtain thus far.

Generating more stringent markers of pluripotency and assays to distinguish the abilities of a given iPS cell line are key priorities. Building on the progress that has already been made using ES cells[81], researchers must continue to improve the understanding of directed differentiation and to develop new protocols. With refined differentiation protocols, researchers will be able to investigate the pathophysiological basis of genetic diseases and carry out drug screening on affected cell types. These protocols will bring the field a step closer to patient-matched cells and tissues for clinical transplantation, a long-standing ambition of the stem-cell field that might be its ultimate measure of success. ■

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
This breakthrough paper describes the derivation of iPS cells directly from mouse somatic cells through the ectopic co-expression of reprogramming transcription factors, providing an alternative source of pluripotent cells for research.
2. Dreisch, H. Entwicklungsmechanische Studien I. Der Wert der ersten beiden Furchungszellen in der Echinodermenentwickelung. Experimentelle Erzeugung von Teil und Doppelbildungen. *Z. Wiss. Zool.* **53,** 160–183 (1891).
3. Dewey, M. J., Martin, D. W. Jr, Martin, G. R. & Mintz, B. Mosaic mice with teratocarcinoma-derived mutant cells deficient in hypoxanthine phosphoribosyltransferase. *Proc. Natl Acad. Sci. USA* **74,** 5564–5568 (1977).
4. Gardner, R. L. Mouse chimeras obtained by the injection of cells into the blastocyst. *Nature* **220,** 596–597 (1968).
5. Brinster, R. L. The effect of cells transferred into the mouse blastocyst on subsequent development. *J. Exp. Med.* **140,** 1049–1056 (1974).
6. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292,** 154–156 (1981).
7. Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl Acad. Sci. USA* **78,** 7634–7638 (1981).
8. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–1147 (1998).
9. Vogel, G. & Holden, C. Ethics questions add to concerns about NIH lines. *Science* **321,** 756–757 (2008).
10. Mosher, J. T. *et al.* Lack of population diversity in commonly used human embryonic stem-cell lines. *N. Engl. J. Med.* **362,** 183–185 (2010).
11. Tabar, V. *et al.* Therapeutic cloning in individual parkinsonian mice. *Nature Med.* **14,** 379–381 (2008).
12. Noggle, S. *et al.* Human oocytes reprogram somatic cells to a pluripotent state. *Nature* **478,** 70–75 (2011).
13. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–146 (2008).
14. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
15. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–1920 (2007).
16. Stadtfeld, M., Brennand, K. & Hochedlinger, K. Reprogramming of pancreatic β cells into induced pluripotent stem cells. *Curr. Biol.* **18,** 890–894 (2008).
17. Eminli, S., Utikal, J., Arnold, K., Jaenisch, R. & Hochedlinger, K. Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous *Sox2* expression. *Stem Cells* **26,** 2467–2474 (2008).
18. Kim, J. B. *et al.* Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature* **454,** 646–650 (2008).
19. Hanna, J. *et al.* Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* **133,** 250–264 (2008).
20. Aoi, T. *et al.* Generation of pluripotent stem cells from adult mouse liver and stomach cells. *Science* **321,** 699–702 (2008).
21. Utikal, J., Maherali, N., Kulalert, W. & Hochedlinger, K. Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *J. Cell Sci.* **122,** 3502–3510 (2009).
22. Sun, N. *et al.* Feeder-free derivation of induced pluripotent stem cells from adult human adipose stem cells. *Proc. Natl Acad. Sci. USA* **106,** 15720–15725 (2009).
23. Maherali, N. *et al.* A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell* **3,** 340–345 (2008).
24. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–199 (2007).
25. Wray, J., Kalkan, T. & Smith, A. G. The ground state of pluripotency. *Biochem. Soc. Trans.* **38,** 1027–1032 (2010).
26. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143,** 508–525 (2010).
27. Maherali, N. & Hochedlinger, K. Guidelines and techniques for the generation of induced pluripotent stem cells. *Cell Stem Cell* **3,** 595–605 (2008).
28. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2,** 230–240 (2008).
This study enumerated the molecular markers and functional criteria for defining reprogrammed cell populations.
29. Chan, E. M. *et al.* Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nature Biotechnol.* **27,** 1033–1037 (2009).
30. Payer, B., Lee, J. T. & Namekawa, S. H. X-inactivation and X-reactivation: epigenetic hallmarks of mammalian reproduction and pluripotent stem cells. *Hum. Genet.* **130,** 265–280 (2011).
31. Maherali, N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1,** 55–70 (2007).
32. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465,** 175–181 (2010).
33. Boland, M. J. *et al.* Adult mice generated from induced pluripotent stem cells. *Nature* **461,** 91–94 (2009).
34. Zhao, X. Y. *et al.* iPS cells produce viable mice through tetraploid complementation. *Nature* **461,** 86–90 (2009).
35. Eakin, G. S., Hadjantonakis, A. K., Papaioannou, V. E. & Behringer, R. R. Developmental potential and behavior of tetraploid cells in the mouse embryo. *Dev. Biol.* **288,** 150–159 (2005).
36. Eggan, K. & Jaenisch, R. Differentiation of F_1 embryonic stem cells into viable male and female mice by tetraploid embryo complementation. *Methods Enzymol.* **365,** 25–39 (2003).
37. Lensch, M. W., Schlaeger, T. M., Zon, L. I. & Daley, G. Q. Teratoma formation assays with human embryonic stem cells: a rationale for one type of human-animal chimera. *Cell Stem Cell* **1,** 253–258 (2007).
38. Park, I. H. *et al.* Disease-specific induced pluripotent stem cells. *Cell* **134,** 877–886 (2008).
This study derived iPS cells from patients with a range of diseases, demonstrating the applications of iPS cells for disease modelling, pathogenesis studies and drug development.
39. Daley, G. Q. *et al.* Broader implications of defining standards for the pluripotency of iPSCs. *Cell Stem Cell* **4,** 200–202 (2009).
40. Miura, K. *et al.* Variation in the safety of induced pluripotent stem cell lines. *Nature Biotechnol.* **27,** 743–745 (2009).
41. Feng, Q. *et al.* Hemangioblastic derivatives from human induced pluripotent stem cells exhibit limited expansion and early senescence. *Stem Cells* **28,** 704–712 (2010).
42. Hu, B. Y. *et al.* Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl Acad. Sci. USA* **107,** 4335–4340 (2010).
43. Kim, K. *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* **467,** 285–290 (2010).
This report describes an exhaustive comparison of pluripotent stem cells derived from mouse embryos or reprogrammed through nuclear transfer or transcription factor co-expression, by using several *in vitro* differentiation assays and methylation analysis; it revealed that iPS cells manifest molecular and behavioural features of the donor tissue of origin, indicating a 'memory' of the somatic tissue.
44. Hu, Q., Friedrich, A. M., Johnson, L. V. & Clegg, D. O. Memory in induced pluripotent stem cells: reprogrammed human retinal-pigmented epithelial cells show tendency for spontaneous redifferentiation. *Stem Cells* **28,** 1981–1991 (2010).
45. Bar-Nur, O., Russ, H. A., Efrat, S. & Benvenisty, N. Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet Beta cells. *Cell Stem Cell* **9,** 17–23 (2011).
46. Urbach, A., Bar-Nur, O., Daley, G. Q. & Benvenisty, N. Differential modeling of fragile X syndrome by human embryonic stem cells and induced pluripotent stem cells. *Cell Stem Cell* **6,** 407–411 (2010).
47. Bock, C. *et al.* Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell **144,** 439–452 (2011).
This study generated genome-wide reference maps of DNA methylation and gene expression, together with the differentiation potential of each cell line, providing a resource for assessing the similarity of ES cells and iPS cells, as well as for predicting the differentiation efficiency of a particular cell line and creating a scorecard for the comprehensive characterization of any pluripotent cell line.
48. Pick, M. *et al.* Clone- and gene-specific aberrations of parental imprinting in human induced pluripotent stem cells. *Stem Cells* **27,** 2686–2690 (2009).
49. Ben-David, U., Mayshar, Y. & Benvenisty, N. Large-scale analysis reveals acquisition of lineage-specific chromosomal aberrations in human adult stem cells. *Cell Stem Cell* **9,** 97–102 (2011).
50. Hussein, S. M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471,** 58–62 (2011).
This study showed that significantly more CNVs were present in early-passage human iPS cells than in intermediate-passage human iPS cells, fibroblasts or human ES cells; it also provided evidence that CNVs conferred a selective disadvantage.
51. Laurent, L. C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8,** 106–118 (2011).
52. Ghosh, Z. *et al.* Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PLoS ONE* **5,** e8975 (2010).
53. Wernig, M. *et al.* A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nature Biotechnol.* **26,** 916–924 (2008).
54. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).
55. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471,** 68–73 (2011).
This study analysed the methylomes of human iPS cells, ES cells, somatic cells, and differentiated iPS and ES cells, and revealed megabase-scale DMRs in iPS cells, indicating incomplete reprogramming of these cells.
56. Shen, Y. *et al.* X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. *Proc. Natl Acad. Sci. USA* **105,** 4709–4714 (2008).
57. Tchieu, J. *et al.* Female human iPSCs retain an inactive X chromosome. *Cell Stem Cell* **7,** 329–342 (2010).
58. Marchetto, M. C. *et al.* A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143,** 527–539 (2010).
59. Pomp, O. *et al.* Unexpected X chromosome skewing during culture and reprogramming of human somatic cells can be alleviated by exogenous telomerase. *Cell Stem Cell* **9,** 156–165 (2011).
60. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).

61. Newman, A. M. & Cooper, J. B. Lab-specific gene expression signatures in pluripotent stem cells. *Cell Stem Cell* **7,** 258–262 (2010).
62. Humpherys, D. *et al.* Epigenetic instability in ES cells and cloned mice. *Science* **293,** 95–97 (2001).
63. Soldner, F. *et al.* Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* **136,** 964–977 (2009).
64. Carey, B. W. *et al.* Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell Stem Cell* **9,** 588–598 (2011).
65. Ohi, Y. *et al.* Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. *Nature Cell Biol.* **13,** 541–549 (2011).
66. Polo, J. M. *et al.* Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nature Biotechnol.* **28,** 848–855 (2010).
67. Martinez, Y. *et al.* Cellular diversity within embryonic stem cells: pluripotent clonal sublines show distinct differentiation potential. *J. Cell. Mol. Med.* http://dx.doi.org/10.1111/j.1582-4934.2011.01334.x (in the press).
68. Osafune, K. *et al.* Marked differences in differentiation propensity among human embryonic stem cell lines. *Nature Biotechnol.* **26,** 313–315 (2008).
69. Muller, F. J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nature Methods* **8,** 315–317 (2011).
70. Lee, G. *et al.* Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. *Nature* **461,** 402–406 (2009).
71. Moretti, A. *et al.* Patient-specific induced pluripotent stem-cell models for long-QT syndrome. *N. Engl. J. Med.* **363,** 1397–1409 (2010).
72. Itzhaki, I. *et al.* Modelling the long QT syndrome with induced pluripotent stem cells. *Nature* **471,** 225–229 (2011).
This study generated iPS cells from patients with long QT syndrome and stimulated them to differentiate into cardiomyocytes that paralleled the disease phenotype *in vitro*, and these cells were then used to evaluate the potency of existing and new therapeutic agents.
73. Agarwal, S. *et al.* Telomere elongation in induced pluripotent stem cells from dyskeratosis congenita patients. *Nature* **464,** 292–296 (2010).
74. Batista, L. F. *et al.* Telomere shortening and loss of self-renewal in dyskeratosis congenita induced pluripotent stem cells. *Nature* **474,** 399–402 (2011).
75. Agarwal, S. & Daley, G. Q. Telomere dynamics in dyskeratosis congenita: the long and the short of iPS. *Cell Res.* **21,** 1157–1160 (2011).
76. Hanna, J. *et al.* Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* **318,** 1920–1923 (2007).
77. Wernig, M. *et al.* Neurons derived from reprogrammed fibroblasts functionally integrate into the fetal brain and improve symptoms of rats with Parkinson's disease. *Proc. Natl Acad. Sci. USA* **105,** 5856–5861 (2008).
78. Zhao, T., Zhang, Z. N., Rong, Z. & Xu, Y. Immunogenicity of induced pluripotent stem cells. *Nature* **474,** 212–215 (2011).
79. Harrison, N. J., Baker, D. & Andrews, P. W. Culture adaptation of embryonic stem cells echoes germ cell malignancy. *Int. J. Androl.* **30,** 275–281 (2007).
80. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7,** 521–531 (2010).
81. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132,** 661–680 (2008).
82. Lowry, W. E. *et al.* Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl Acad. Sci. USA* **105,** 2883–2888 (2008).
83. Huangfu, D. *et al.* Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nature Biotechnol.* **26,** 1269–1275 (2008).
84. Sommer, C. A. *et al.* Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem Cells* **27,** 543–549 (2009).
85. Anokye-Danso, F. *et al.* Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell* **8,** 376–388 (2011).
86. Woltjen, K. *et al. piggyBac* transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458,** 766–770 (2009).
87. Somers, A. *et al.* Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. *Stem Cells* **28,** 1728–1740 (2010).
88. Zhou, W. & Freed, C. R. Adenoviral gene delivery can reprogram human fibroblasts to induced pluripotent stem cells. *Stem Cells* **27,** 2667–2674 (2009).
89. Stadtfeld, M., Nagaya, M., Utikal, J., Weir, G. & Hochedlinger, K. Induced pluripotent stem cells generated without viral integration. *Science* **322,** 945–949 (2008).
90. Okita, K., Nakagawa, M., Hyenjong, H., Ichisaka, T. & Yamanaka, S. Generation of mouse induced pluripotent stem cells without viral vectors. *Science* **322,** 949–953 (2008).
91. Si-Tayeb, K. *et al.* Generation of human induced pluripotent stem cells by simple transient transfection of plasmid DNA encoding reprogramming factors. *BMC Dev. Biol.* **10,** 81 (2010).
92. Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc. Jpn Acad.* **85,** 348–362 (2009).
93. Kim, D. *et al.* Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell* **4,** 472–476 (2009).
94. Zhou, H. *et al.* Generation of induced pluripotent stem cells using recombinant proteins. *Cell Stem Cell* **4,** 381–384 (2009).
95. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7,** 618–630 (2010).
96. Miyoshi, N. *et al.* Reprogramming of mouse and human cells to pluripotency using mature microRNAs. *Cell Stem Cell* **8,** 633–638 (2011).
97. Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc. Natl Acad. Sci. USA* **107,** 9222–9227 (2010).
98. Buecker, C. *et al.* A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell* **6,** 535–546 (2010).
99. Li, W. *et al.* Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. *Cell Stem Cell* **4,** 16–19 (2009).
100. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4,** 487–492 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements G.Q.D. is supported by grants from the National Institutes of Health (RO1-DK70055, RO1-DK59279 and UO1-HL100001, as well as special funds from the American Recovery and Reinvestment Act of 2009 (RC2-HL102815), the Roche Foundation for Anemia Research, Alex's Lemonade Stand Foundation and the Harvard Stem Cell Institute. G.Q.D. is an affiliate member of the Broad Institute, a recipient of Clinical Scientist Awards in Translational Research from the Burroughs Wellcome Fund and the Leukemia and Lymphoma Society, and an investigator of the Howard Hughes Medical Institute and the Manton Center for Orphan Disease Research. We gratefully acknowledge A. de Los Angeles for providing the images for Fig. 1a, d, f; T. Onder, for Fig. 1b; and A. Cherry, for Fig. 1c, e.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to G.D. (george.daley@childrens.harvard.edu).

REVIEW

doi:10.1038/nature10762

Clonal evolution in cancer

Mel Greaves[1] & Carlo C. Maley[2]

Cancers evolve by a reiterative process of clonal expansion, genetic diversification and clonal selection within the adaptive landscapes of tissue ecosystems. The dynamics are complex, with highly variable patterns of genetic diversity and resulting clonal architecture. Therapeutic intervention may destroy cancer clones and erode their habitats, but it can also inadvertently provide a potent selective pressure for the expansion of resistant variants. The inherently Darwinian character of cancer is the primary reason for this therapeutic failure, but it may also hold the key to more effective control.

Cancer is a major cause of death throughout the world and, despite an extraordinary amount of effort and money spent, the eradication or control of advanced disease has not been achieved[1]. Although we have a much greater understanding of cancer biology and genetics[2], translation into clinical practice needs to allow for the cellular complexity of the disease and its dynamic, evolutionary characteristics. These features provide both barriers to, and opportunities for, successful treatment.

In 1976, Peter Nowell[3] published a landmark perspective on cancer as an evolutionary process that is driven by stepwise, somatic-cell mutations with sequential, subclonal selection. This is a parallel to Darwinian natural selection, with cancer clones as the equivalent of asexually reproducing, unicellular quasi-species. Modern cancer biology and genomics have validated cancer as a complex, Darwinian, adaptive system[4,5] (Box 1 and Supplementary Information).

Cancer-clone evolution takes place within tissue ecosystem habitats. These habitats have evolved over a billion years to optimize multicellular function but restrain clonal expansion of renegade cells. However, the resilience of multicellular and long-lived animals depends on the phenotypic properties that, if not tightly regulated, drive or sustain malignancy: that is, cellular self-renewal and stabilization of telomeres, which allow extensive proliferation, angiogenesis, cell migration and invasion[6].

The long time period usually required for cancer symptoms to emerge and the complexity of the resultant mutations is, in part, a reflection of the sequential and random searches for phenotypic solutions to constraints from the micro-environment. The evolutionary progression of cancer is usually stalled or aborted, as shown by the high frequency of clinically covert premalignant lesions[7–9]. Cancer-suppressive mechanisms relegate most cancers to old age, when they have little effect on the reproductive fitness of their hosts.

Limited resources, environment architecture and other constraints of the micro-environment limit the size of solid tumours at every stage of their progression. Even advanced malignancies can show Gompertzian growth[10] — the cancer cell doubling time (around 1–2 days) is orders of magnitude faster than tumour doubling time (around 60–200 days)[10] — implying that the vast majority of cancer cells either die before they can divide[11] or are kept from dividing by the tumour micro-environment. Thus, natural selection in tumours, in the same way as selection in organisms, takes place through competition for space and resources.

Oncologists change cancer-clone dynamics by introducing a potent source of artificial selection in the form of drugs or radiation, but evolutionary principles still apply. Usually, treatment will result in massive cell death, which provides a selective pressure for the proliferation of variant cells that resist treatment (the mechanisms for this are discussed later). Furthermore, many cancer therapeutics are genotoxic; cells surviving treatment, which could then go on to regenerate the cancer, may have mutated further, resulting in cells with improved fitness and malignant potential.

The tools of and insights from evolutionary biology and ecology can therefore be applied to the dynamics of cancer before and after treatment to explain the modest returns from cancer therapy. We show that cancer is an inherently evolutionary process and suggest alternative strategies for effective control.

Mutational drivers and clonal dynamics

The basic principle of a Darwinian evolutionary system is the purposeless genetic variation of reproductive individuals who are united by common descent, together with natural selection of the fittest variants. Cancer is a clear example of such a system. Most mutational processes have a bias at the DNA sequence level. The particular mutational spectra in a cancer cell can be a reflection of error-prone repair processes or associated with a genotoxic exposure (for example, cigarette carcinogens, ultraviolet light and chemotherapeutic drugs[2]). The patterns of genetic instability (chromosomal or microsatellite) in cancer cells may reflect exposure to, and the selective pressure exerted by, some classes of chemical carcinogens[2]. Nevertheless, for the functions encoded in genes, mutagenic processes are essentially blind or non-purposeful (with the exception of intrinsic mutagenic or recombinatorial enzymes preferentially targeting lymphoid immunoglobin or T-cell receptor genes[12]). The recurrent, mutation-endowed fitness traits in cancer reflect the potent impact clonal selection can have.

Clones evolve through the interaction of selectively advantageous 'driver' lesions, selectively neutral 'passenger' lesions and deleterious lesions (a 'hitchhiker' mutation in evolutionary biology is equivalent to a passenger mutation in cancer biology). In addition, 'mutator' lesions increase the rate of other genetic changes[13,14], and micro-environmental[15] changes alter the fitness effects of those lesions. The identification of driver lesions is supported by the independent observation that these lesions occur more frequently in multiple neoplasms than would be expected in the normal background mutation rate, that they are associated with clonal expansions[16,17] and from the type of mutation seen (missense, nonsense, frameshift, splice site, phosphorylation sites and double deletions)[18–20], particularly if the gene involved has a known role in cellular processes relevant to oncogenesis. The evidence gained from genetic studies

[1]Division of Molecular Pathology, The Institute of Cancer Research, Brookes Lawley Building, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. [2]Center for Evolution and Cancer, Helen Diller Family Comprehensive Cancer Center, Department of Surgery, University of California, 2340 Sutter Street PO Box 1351, San Francisco, California 94115, USA.

of human tumours should be corroborated with functional tests and animal models. Passenger lesion status can also be ambiguous or context-dependent: for example, cases of monoallelic loss that only impact on function when the second allele of the same gene is lost, mutations that only cause a phenotypic effect when another gene locus also mutates, or cases in which the mutants are functionally relevant only in the context of therapeutic responses involving that gene.

Only a few studies have attempted to quantify the selective advantage provided by driver mutations. Bozic *et al.*[21] (using a non-spatial population genetics model of sequential, exponential clonal expansion) derived a formula for the proportion of expected neutral passenger mutations versus the proportion of selectively advantageous driver mutations as a function of the selective advantage of the driver mutations. By fitting this equation to glioblastoma and pancreatic cancer resequencing data, the authors estimated that driver mutations gave an average fitness advantage of only 0.4% (ref. 21). To measure the mutant clone selective advantage directly would require longitudinal samples of a neoplasm and estimation of the clone sizes at each time point.

The dynamics of somatic evolution depend on the interaction of mutation rate and clonal expansion. Mutation rate varies substantially between different genomic regions[22] and between different types of abnormality (for example, single-base sequence changes versus balanced chromosomal rearrangements and gene fusions), and mutation rates will increase by the instigation of genetic instability[23–25]. The rate of epigenetic change has been estimated to be orders of magnitude higher than that of genetic change[26], and could be a major determinant of clonal evolution. Natural selection affects epigenetic variation within neoplasms[27], because epigenetic changes are inherited at cell division and can affect cell phenotypes. Evolutionary biology tools to address many of these mutation rate complexities exist (see Supplementary Information), but these remain under used in cancer biology[28]. The traditional model of clonal evolution suggests that a series of clonal expansions grows to dominate the neoplasm ('selective sweeps')[16,21,29], but this can occur only if the time to the next driver mutation is longer than the time required for a clone to sweep through the neoplasm. In addition, if the second mutation occurs in a competitor clone, the expansion of both clones is restrained by mutual competition (known as clonal interference)[30]. Given the large population size and high mutation rate typical of neoplasms, clonal competition is probably common[31,32]. This issue is best addressed by serial sampling, and limited data suggest that parallel clonal expansions occur before subclones begin to dominate in early cancer development[33–35]. Initial evidence indicates that large clonal expansions after cell transformation are rare[26]. Direct evidence, from serial sampling of oncogenic mutations in advanced disease[36], metastasis[37] or post-chemotherapy relapses (see Supplementary Information), indicates selective sweeps originate from pre-existing genetic variants or subclones.

Punctuated equilibrium versus gradualism

The argument of gradualism versus punctuated equilibrium[38] (a longstanding debate in species evolution) has recently emerged in the consideration of the clonal evolution of neoplasms. It is unknown whether malignant clones, with their markedly altered genomes, evolve gradually through a sequence of genetic alterations and clonal expansions; accumulate many lesions over time in a rare, undetected subclone that finally appears in a clonal expansion; or have a few, large-scale punctuated changes, possibly prompted by an acute insult or a single, catastrophic mitotic event that generates multiple lesions across the genome (or on a single chromosome, known as chromothripsis)[39]. Evidence of tens of non-synonymous mutations in cancers was interpreted under the assumption that they were generated by tens of clonal expansions[29]. Reconstruction of genealogies of neoplastic clones, based on genetic heterogeneity within neoplasms, suggests that clones with ancestral genomes are not driven to extinction by later clonal expansions[31–33], which allows the history of a neoplasm to be revealed. Breast cancer data[32] have shown that clones with intermediate genotypes are difficult to detect; each clone generates a cloud of genetic neutral or non-viable subclone variants around it. A study of B-cell chronic lymphocytic leukaemia[40] suggests that intermediate clones can be detected, but at a frequency of <0.001, which was below the detection threshold of the breast cancer study[32]. Intermediate clones may be rare because they have had limited potential to expand or because they were once common but were outcompeted by more recent clones.

The frequency of premalignant clonal lesions (or carcinoma *in situ*) substantially exceeds clinical cancer rates[7–9]. This, as well as cancer dormancy[41] and genetic reconstitution of clonal histories[37], indicates that cancer clones have long periods of stasis. However, cancer-clone evolution probably passes a point of no return, possibly at the metastatic growth stage. If unlimited proliferative capacity is guaranteed by telomere stabilization[25], then clonal expansion is stopped only when the size threatens the life of the patient. When provided (albeit rarely) with the routes for dissemination and immunoselection, cancer cells can have a parasite-like immortality and can re-establish themselves in other individuals[6,42,43].

BOX 1

Cancer as a complex system

- Cancers exist in a variety of taxonomic quasi-classes, genera, species, characterized by divergent cells of origin and mutational spectra. Each cancer is unique.
- Cancers evolve over a variable time frame (anywhere from 1 to 50 years), and the clonal structure, genotype and phenotype can shift over time in each patient. Each cancer is, in effect, multiple different (subclonal) cancers that occupy overlapping or distinct tissue habitats.
- The number of mutations in a cancer can vary from a handful (10–20) to (the more usual) hundreds or thousands. The great majority are passengers, and a modest, but undefined, number are functionally relevant drivers. The mutational processes are very diverse.
- Cancers acquire, through mutational and epigenetic changes, a variety of phenotypic traits that compound to allow territorial expansion, by proliferative self-renewal, migration and invasion — properties that are cruical to normal developmental, physiological and repair processes.
- Advanced, disseminated or very malignant cancers seem to be almost uniquely competent to evade therapy.
- Most, if not all, of this complexity can be explained by classical evolutionary principles.

The cancer ecosystem

Tissue ecosystems provide the venue and determinants for fitness selection (the adaptive landscape[44]). Tissue micro-environments are complex, dynamic states with multiple components that can influence cancer-clone evolution (Fig. 1). For example, transforming growth factor-β is a cancer-ecosystem regulatory molecule[45]. Other cellular and cytokine components of inflammatory lesions are potent and common modulators of the cancer-cell ecosystem[25].

The interaction between cancer cells and their tissue habitats is reciprocal. Cancer cells can remodel tissue micro-environments and specialized niches to their competitive advantage[46]. Cancer-clone expansion is controlled by architectural constraints or barriers, such as sequestration of stem cells into crypts in the gastrointestinal tract[47]

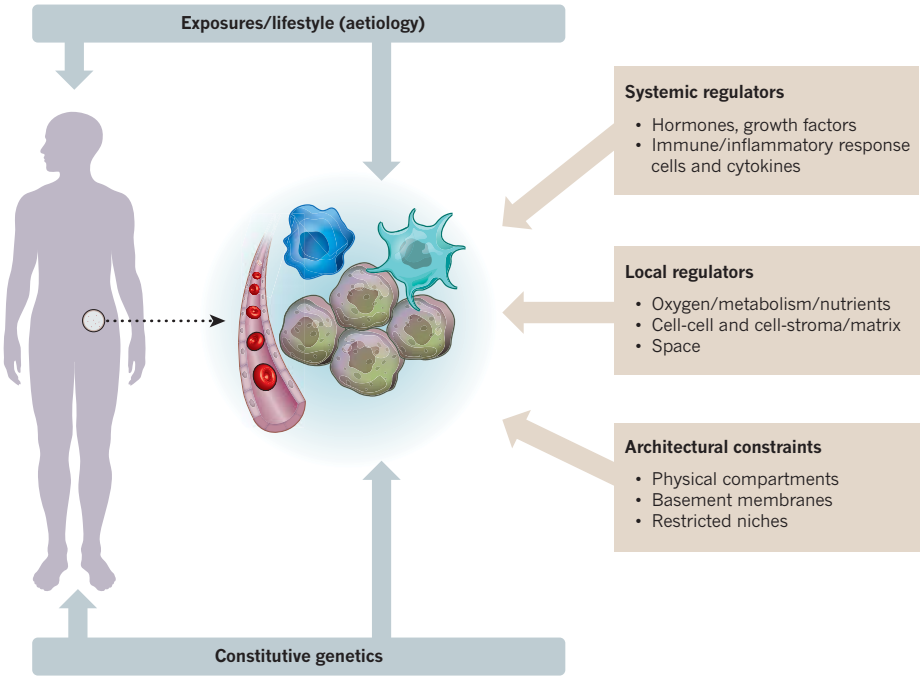


Figure 1 | The complexity of tissue ecosystems. Exposure, the constitutive genetics of the host cells, systemic regulators, local regulators and architectural constraints all impinge on the evolution of somatic cells.

and the need for external signals for proliferation and cell survival. However, some micro-environmental components can promote neoplastic cells; for example, infiltrating macrophages and neovascularization, in response to anoxia, can support neoplastic cell survival and proliferation. Mathematical modelling shows cancer-clone evolutionary selection for more robust or malignant phenotypes is less likely in more stable or homogeneous micro-environments[48]. Spatial heterogeneity of resources in the primary tumour selects for cell migration and emigration, which may explain why there is selection for metastasis[49]. Preclinical models have suggested that normalizing the resources across the primary tumour can suppress metastasis[50]. As clones and subclones expand, migrant cells invade new habitats within and between tissues, in which they experience new selective pressures that can cause further cancer-cell diversity. This malignant feature, and its associated morbidity, characterizes end-stage cancer.

Cancer-cell habitats are not closed systems. The tissue ecosystem, in addition to regulation by systemic factors (such as nutrients and hormones) or invasion by inflammatory or endothelial cells, is modified by external factors. As well as the tissue site, the ecosystem for each cancer includes environmental, lifestyle and associated aetiological exposure of the patient. Genotoxic exposure (such as, cigarette carcinogens or ultraviolet light), infection, and long-term dietary and exercise habits that affect calorie, hormone or inflammation levels can have a profound effect on the tissue micro-environments, as well as directly on cancer cells (Fig. 1). These factors are the aetiological link to the initiation or progression of cancer, and without such modulating exposure, the risk of cancer-clone initiation and evolution would be reduced.

Cancer-tissue ecosystems can be radically altered after chemotherapy or radiotherapy. Most cancer cells may be decimated, but the remodelled landscape creates new selective pressures, resources and opportunities that may allow pre-existing variant cancer cells that survived treatment to emerge. Crucially, stroma or specialized habitat niches may protect cancer cells against the therapy[51].

Cancer genomics and clonal architecture

Cancer-genome sequencing, facilitated by the introduction of second-generation whole-genome sequencing, has provided further insight into the complexity of the genetics and evolutionary biology of cancer cells[2]. In most cases, transformation and metastases are probably clonal[2] because they are derived from single cells; therefore, the identification of the mutations present in all of the cells of a tumour can help to reconstruct the genotype of the founder cell. These founder events limit the genetic and clonal complexity of tumours. We already had a long list of recurring driver mutations (with gain or loss of function) as a result of the fine mapping of chromosomal breaks, candidate gene sequencing and functional screening of bulk samples from tumours. However, the use of genomic screens has demonstrated the scale of cancer-genome complexity. Individual cancers can contain hundreds, or tens of thousands, of mutations and chromosomal alterations[2]. The great majority of these are assumed to be neutral mutations arising from genetic instability. Chromosomal instability (amplifications, deletions, translocations and other structural changes) is a common feature, but it is not clear whether there is an increased rate of simple base-pair mutations in cancer[2,21,23,52]. Evolutionarily neutral alterations are thought to register in the screens because they hitchhike on clonal expansions that are driven by selectively advantageous alterations or by drift. In addition, data have confirmed that each cancer in each patient has an individually unique genomic profile. It is possible that cancer cells need only a modest number of phenotypic traits to deal with all of the constraints and evolve into a fully malignant or metastatic tumour[25], but the genomics data suggest that this can be achieved by an almost infinite variety of evolutionary trajectories and with multiple different combinations of driver mutations[44].

Paradoxically, genome profiles underestimate complexity. So far, they have been mostly one-off snapshots from a single sample at a single diagnostic time point. We know that serial or parallel sampling using more conventional genetic analysis uncovers genetic diversity within a tumour. Whole-genome sequencing of paired primary tumours versus metastatic samples has so far been limited, but it has revealed that individual metastatic lesions are clonal in origin and genetically unique, yet have a clonal ancestry traceable to the primary tumour[2]. 'The genome' description is perhaps also misleading because genetic variants are identified in 5–50% of reads, which suggests subclonal distribution of most mutations[53], but the segregation pattern of mutations within subclones is lost when DNA is extracted from the total cell population. This is important if patient-specific genomic profiles are to provide a platform for selecting therapeutic targets. Arguably, subclonal genetic diversity is key to the success or failure of therapy. This is a considerable challenge, technically and bioinformatically, in cancer genomics and will require deep sequencing[40]

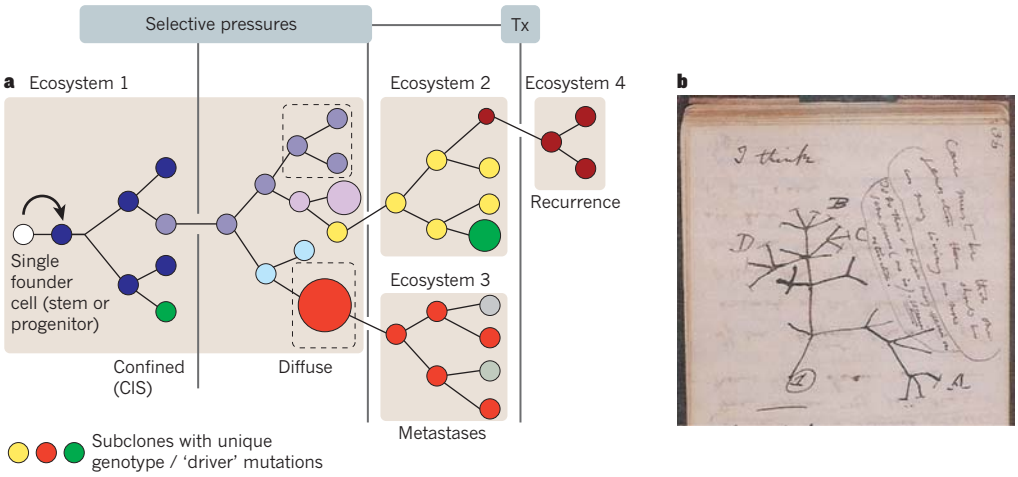


Figure 2 | The branching architecture of evolution. **a**, Cancer clones. Selective pressures allow some mutant subclones to expand while others become extinct or remain dormant. Vertical lines represents restraints or selective pressures. This is a representative pattern for common, solid cancers; as recognized by Nowell[3], leukaemic clones may expand over a shorter time frame (years versus decades), and be subject to fewer restraints and mutational events. Ecosystems 1–4 (boxes) represent the different tissue ecosystems or habitats. Smaller boxes within Ecosystem 1 represent localized habitats or niches. Each differently coloured circle represents a genetically distinct subclone. Metastatic subclones can branch off into different time points in the sequence from either minor or major clones in the primary tumour. Tx, therapy. CIS, carcinoma *in situ*. **b**, Darwin's branching evolutionary tree of speciation from his 1837 notebook.

and investigation of the genomes of single cells for patterns of segregation of mutations to understand the genetic diversity within neoplasms and how this changes in response to interventions.

Subclonal segregation of mutations and clonal architecture

The classic model of clonal evolution suggests there is a sequential acquisition of mutations with concomitant, successive subclonal dominance or selective sweeps. Histopathological evidence of disease progression (adenoma, carcinoma and metastases) supports this model. At each stage of this evolution, individual cells and their progeny (subclones) compete for space and resources. Multiplexed, single-cell mutational analysis (ideally in serial samples) is the most appropriate way to examine clonal architecture. So far, there are only a few examples of this[10,32,33], but they have provided evidence of the complex pattern of subclonal segregation of mutations — consistent with Nowell's model. The large amount of data from tissue sections, small biopsies and, more recently, single-cell analysis[33] is evidence that the evolutionary trajectories are complex and branching, exactly as Nowell proposed and in parallel with Darwin's iconic evolutionary speciation tree (Fig. 2). Attempts to simplify this complex system into a linear sequence of mutational events on the basis of cross-sectional data have probably been misleading[54]. However, by comparing the mutational genomes of the subclones, it is possible to discover their evolutionary or ancestral relationships, as well as the order of events during the development of that neoplasm[32,33,37,53,54]. Clonal evolution from common ancestral cancer cells is demonstrated in identical twins with concordant acute leukaemia[55,56], in metastatic lesions[2,10] and, by inference, in some cases of bilateral testicular cancer[57] (Fig. 3). In this context, divergent cancer-clone genotypes and phenotypes correspond to allopatric speciation in separate natural habitats (for example, Darwin's finches on the Galapagos Islands[58]).

Profiles of subclones within a neoplasm can be used to determine 'molecular clocks' that can then be linked to time events in the history of the neoplasm. For example, DNA methylation changes and base-pair mutations have been used to infer clonal expansion dynamics[26] and the time between initiation, invasion and metastasis[17,37,52]. It is even possible to determine the relative timing of events during progression from a single sample, based on deep sequencing[59].

Subclones may be mixed together within the primary tissue[37,60], but given their single-cell origin and bifurcating pathways, it is not surprising that they can also occupy distinctive territories[35,37,61,62] (Fig. 4a). Cancer-clone evolution involves contemporaneous subclones with distinctive mutational and phenotypic profiles that may be territorially segregated, which has considerable practical implications for diagnosis, prognosis and targeted therapy based on biopsy sampling[63]. It remains unclear whether all subclonal diversification reflects the impact of driver mutations and selective advantage, or is also the result of genetic drift of selectively neutral mutations or even epigenetic alterations. The level of diversity within the subclonal structure can be measured[35,64,65] and has been shown to be a robust biomarker for predicting progression to malignancy in Barrett's oesophagus[65]. It is also associated with the tumour stage and subtype of breast cancer[64].

Units of selection and cancer stem cells

Evolutionary theory suggests that natural selection operates in any system that has components with varying reproductive potential[4]. In

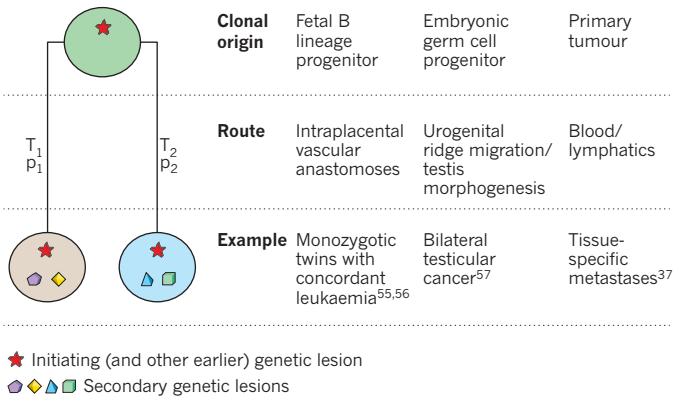


Figure 3 | Divergent (branching) clonal evolution of cancer with topographical separation. In each example, a clonal (single cell) ancestry is indicated by a shared acquired mutation (for example, *ETV6–RUNX1* fusion for leukaemias and *KIT* mutation for testicular cancers). The time at which the two subclones evolve (T_1 and T_2) can be temporarily synchronous or develop several years apart[37,55–57]. The probabilities of subclones emerging as shown are independent and different (p_1 and p_2). In most cases (90% for monozygotic twins), only one twin develops overt leukaemia. The penetrance of bilateral testicular cancer having a common origin[57] is unknown.

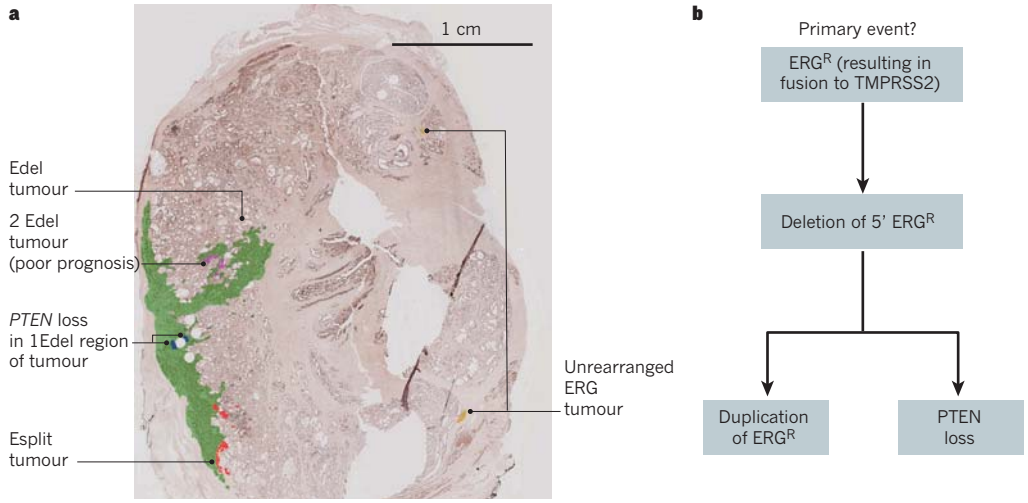


Figure 4 | Topography of cancer subclones. a, Tissue section of prostate to detect genetic events: *TMPRSS2–ERG* fusion (*ERG* via rearrangement (ERG^R)) and *PTEN* loss. **b**, The presumed sequence of clonal events.

the progression of cancer, or its resurgence after therapy, the primary unit of selection is the cell. This cell has to have extensive replicative potential, the so-called cancer stem cell (also known as the cancer-initiating or propagating cell) (Fig. 5).

The cancer stem-cell hypothesis was developed through transplantation experiments with leukaemic cells[66], and although it has been reported to be a general feature of all cancers[67], this idea is contentious. There has been no consensus on whether cancer stem cells are rare or high-frequency cells, or whether they have fixed, hierarchical or variable phenotypes, but considering the evolutionary progression in cancer, cells with extensive propagating activity are unlikely to be fixed entities[68,69]. Cancer stem cells are the cellular drivers of subclonal expansion and so probably vary in frequency and phenotypic features. The only feature they must have is the potential for extensive self-renewal (Fig. 5). Quantitative measures of stem-cell activity or self-renewal (through xenotransplantation or gene-expression signatures) can be used to predict the clinical outcome of several cancer types[70]. The cancer stem cell's ability to self-renew is made stronger by an aberrant genotype and, possibly, other, epigenetic, features. Several testable predictions can be made from this. First, cancer stem cells should evolve and change in genotype and phenotype as the cancer evolves before and after therapy. Some therapies may even provide a strong selection for cancer stem-cell survival and proliferation[71]. Second, as cancers progress, there should be selective pressure for the cells with the most extensive self-renewing capacity, but at the expense of cells with the ability to differentiate. This has been observed in chronic myeloid leukaemia (CML)[72] and mouse models[73,74]. A higher probability of symmetrical self-renewing proliferative cycles would be expected to result in an increased number and frequency of cancer stem cells. It is therefore of some consequence that loss of the *TP53* DNA damage checkpoint, which frequently correlates with cancer progression and clinical intransigence[75], seems to 'release' stem-cell-like transcriptional signatures[76] and leads to enhanced self-renewal in mammosphere culture systems[77]. The frequency of cancer stem cells could then increase from low to very high frequency as the disease progresses[78,79]. Third, for selection to operate through microenvironmental or therapeutic pressures, there should be contemporaneous genetic variation in cancer stem cells, which has been shown in leukaemias[33,80].

These considerations have significant clinical implications. Whatever the frequency and phenotype, if self-renewing cancer stem cells drive and sustain cancer-clone evolution, this suggests they are the repository of functionally relevant mutational events that drive clonal selection before and after therapy. This supports the view that cancer stem-cell restraint or elimination should be the aim of any therapy. However, if cancer stem cells are as genetically (and epigenetically) diverse as evolutionary considerations and initial experiments[33,80,81] indicate, this could be the reason for therapeutic failure. The adaptability of cancer stem cells provided by genetic diversity is added to by what seems to be their intrinsically lowered susceptibility to drugs and irradiation[82]. This may be because of the association with stromal cells[83] and the quiescence of cancer stem-cell subpopulations, as well as the properties of enhanced DNA repair and elevated expression of drug efflux pumps, which may be the evolved contingencies to protect normal stem cells.

Subclonal genetic heterogeneity is a common, if not universal, feature of cancers[84]. However, it cannot be assumed that all subclones are sustained by cancer stem cells; some could be evolutionary deadends generated by cells with only limited propagating potential. It is partly to accommodate this that the *in vivo* assay for cancer stem cells involves sequential transplants[66]. Ideally, the genomes of single cancer stem cells would be interrogated to investigate how they relate to subclones, but this is not currently possible. However, the genetic heterogeneity of cancer stem cells can be inferred by comparing subclonal diversity or clonal architecture before and after transplantation. Quadrant sections of glioblastoma have been shown to have divergent but related genotypes, but all sections contained cells that read-out in the *in vivo* (intracerebral) cancer stem-cell assay[85]. More definitive data come from comparing pre- and post-transplant subclonal genetic profiles that were investigated at the single cell level or by single nucleotide polymorphism arrays in B-cell precursor acute lymphoblastic leukaemia. Multiple subclones from each patient's diagnostic sample registered in the *in vivo* cancer stem-cell transplant assays, albeit with variable competitive potency[33,80,81]. We are still awaiting experimental confirmation that genetic diversity of cancer stem cells is a common feature of cancer, but, assuming that it is, this will have important therapeutic implications.

A Darwinian bypass

Nowell[3] stated in his landmark article "more research should be directed towards understanding and controlling the evolutionary process in tumours before it reaches the late stage seen in clinical cancer". Although cancer therapy has had its successes, in reality very few advanced or metastatic malignancies can be effectively controlled or eradicated. Genetic variation in cancer stem cells, particularly if induced by genetic instability, provides the opportunity for cells to escape and the therapy to fail. Other, non-genetic, mechanisms of positive selection by therapy exist, including signalling plasticity (or oncogene bypass)[86], quiescence[87] and epigenetic changes[88]; however, many of these depend on heritable, and thus selectable, epigenetic variation. Great expectation has been placed on the audit of cancer genomes that, by identifying recurrent and "druggable" mutations, would herald a new phase of highly specific or targeted

small-molecule inhibitors and personalized medicine[89]. Oncogene addiction may be the Achilles heel of cancer in this respect[90]. The success of imatinib and the derivative non-receptor tyrosine (ABL1) kinase inhibitors in CML[90] was very encouraging, but CML is not a typical cancer. It is essentially a premalignant (albeit ultimately lethal) condition, probably driven by a single founder mutation (*BCR–ABL1* fusion), which provides a universal target for therapy. Even in the most favourable of circumstances, escape occurs either by quiescence (and coupled resistance) of cancer stem cells[91] or by mutation of the ABL1 kinase target. Once CML has evolved to an overt malignancy or blast crisis, with increased genetic complexity, ABL1 kinase-directed therapy is often ineffective.

Other small-molecule inhibitors directed at mutant products have produced encouraging results in patients with advanced disease, but the benefits are transitory and cancer clones re-emerge with resistant features. When the targets selected are non-founder mutations, even if they are dominant in the neoplasm, therapy can be predicted to select for subclones lacking the mutant target[70]. Alternatively, subclones can have additional mutations that allow a bypass of the signalling pathway of the drug target, such as the MET proto-oncogene (*MET*) amplification in *EGFR* mutant lung cancer treated with EGFR kinase inhibitors[92].

Supporters of targeted therapy and personalized medicine argue that a combination of drugs that target components of networked signalling and are tailored to the individual patient's cancer genome is the solution to this problem. In this regard, synthetic lethal strategies seem promising[93].

Self-renewing cancer cells are the ultimate target for therapy, so high-throughput screening for selective inhibitors is an encouraging development[71]. Ways to target the components of the self-renewing process itself (independent of specific mutant genotype) deserve exploration, especially if a distinction can be made from normal adult stem cells. In the case of CML, intrinsically resistant (and possibly quiescent) stem cells, have been targeted by combining selective kinase (ABL1) inhibitors with inhibitors of a histone deacetylase[94] or BCL6 (ref. 95). Ultimately, it may prove difficult to thwart the plasticity and adaptability of cancer cells (or cancer stem cells), which are an inherent evolutionary feature of advanced disease, and a 'Darwinian bypass' may be required, for which there are a number of possibilities. An implication of the evolutionary diversity of cancer is that prevention (smoking cessation, avoiding sunburn, prophylactic vaccines, and so on) makes a great deal of sense, as does early detection and intervention (that is, before genetic diversification and dissemination become extensive).

An alternative therapeutic strategy is to focus on the microenvironmental habitat using 'ecological therapy', which aims to change the essential habitat and dependency of the cancer cells[96]. For example, anti-angiogenesis can provide a potent restraint on cancer stem cells[97]. Other examples are the use of bisphosphonates to remodel bone in patients with prostate cancer, the use of aromatase inhibitors in patients with breast cancer, exploiting hypoxia, the use of inhibitors of inflammation or tumour-infiltrating macrophages, and blocking cancer stem-cell interactions with essential stromal or niche components[96,98].

Another alternative is to control the cancer, rather than eradicate it, thereby turning cancer into a chronic disease. Because the speed of evolution is proportional to the fitness differential between the cells, cytotoxic drugs are predicted to select rapidly for resistance[5]. It is thought they cause competitive release[99] by removing all of the competitors of resistant cells. In contrast, cytostatic drugs should delay progression and mortality longer than cytotoxic drugs because sensitive competitor cells remain in the tissue to occupy space and consume resources that would otherwise be used by the resistant clones. In addition, by suppressing cell division, cytostatic drugs also suppress the opportunities for new mutations. A study by Gatenby and colleagues[100] showed that by treating an aggressive ovarian cancer (OVCAR-3) xenograft tumour to maintain a stable size, rather than to eradicate it, host mice could be kept alive much longer. Moreover, the dose of carboplatin necessary to keep the tumour at a manageable size declined over time[100]. Researchers should now focus on what phenotypes can be selected for to make neoplasms less deadly and more clinically manageable.

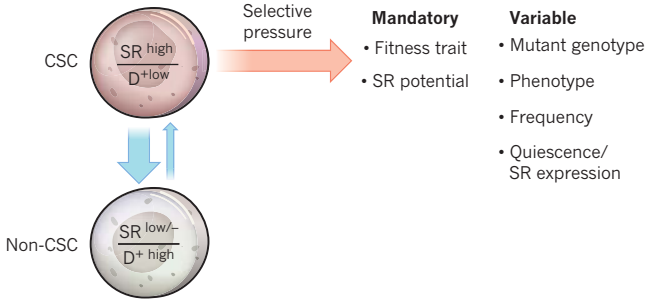


Figure 5 | Selective pressure on cancer stem cells. Selective pressures can include environmentally derived genotoxicity, natural or physiological restraints, cancer therapy, and so on. Mutation in progenitor cells can convert these cells back to a self-renewing population[72]. The small blue arrow represents a mutation; the large blue arrow represents differentiation: in both cases they represent a change in state. In addition to the mandatory trait of self-renewal, cancer stem cells (CSC), can exhibit any phenotypic feature that allows cells to continue to survive and proliferate in the face of a particular constraint. D^+, differentiation; SR, self-renewal.

The evolutionary theory of cancer has survived 35 years of empirical observation and testing, so today it could be considered a bona fide scientific theory. The basic components of somatic evolution are well understood, but the dynamics of somatic evolution remain unclear. Fortunately, there are evolutionary biology tools that may be applied to neoplasms to address many of the fundamental cancer biology questions, such as the order of events in progression, distinguishing driver from passenger mutations, and understanding and preventing therapeutic resistance. The dynamics of clonal diversification and selection are critical to understanding these issues. The challenge now is to use the clinical opportunities to address directly the evolutionary adaptability of neoplasms and design interventions to slow, direct or control cancer-cell evolution to delay or prevent mortality. ■

1. Jemal, A. *et al.* Cancer statistics, 2008. *CA Cancer J. Clin.* **58,** 71–96 (2008).
2. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).
3. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
The foundation paper that established the evolutionary theory of cancer.
4. Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Rev. Cancer* **6,** 924–935 (2006).
5. Pepper, J., Scott Findlay, C., Kassen, R., Spencer, S. & Maley, C. Cancer research meets evolutionary biology. *Evol. Appl.* **2,** 62–70 (2009).
6. Greaves, M. *Cancer: The Evolutionary Legacy* (Oxford Univ. Press, 2000).
7. Sakr, W. A., Haas, G. P., Cassin, B. F., Pontes, J. E. & Crissman, J. D. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *J. Urol.* **150,** 379–385 (1993).
8. Mori, H. *et al.* Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc. Natl Acad. Sci. USA* **99,** 8242–8247 (2002).
9. Reid, B. J., Li, X., Galipeau, P. C. & Vaughan, T. L. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nature Rev. Cancer* **10,** 87–101 (2010).
10. Klein, C. A. Parallel progression of primary tumours and metastases. *Nature Rev. Cancer* **9,** 302–312 (2009).
11. Malaise, E. P., Chavaudra, N. & Tubiana, M. The relationship between growth rate, labelling index and histological type of human solid tumours. *Eur. J. Cancer* **9,** 305–312 (1973).
12. Tsai, A. G. *et al.* Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell* **135,** 1130–1142 (2008).
13. Bardelli, A. *et al.* Carcinogen-specific induction of genetic instability. *Proc. Natl Acad. Sci. USA* **98,** 5770–5775 (2001).
14. Cahill, D. P., Kinzler, K. W., Vogelstein, B. & Lengauer, C. Genetic instability and Darwinian selection in tumors. *Trends Cell Biol.* **9,** M57–M60 (1999).

15. Barcellos-Hoff, M. H., Park, C. & Wright, E. G. Radiation and the microenvironment – tumorigenesis and therapy. *Nature Rev. Cancer* **5,** 867–875 (2005).
16. Maley, C. C. *et al.* Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res.* **64,** 3414–3427 (2004).
17. Tao, Y. *et al.* Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proc. Natl Acad. Sci. USA* **108,** 12042–12047 (2011).
18. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
19. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).
20. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173,** 2187–2198 (2006).
21. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107,** 18545–18550 (2010).
22. Schwartz, M., Zlotorynski, E. & Kerem, B. The molecular basis of common and rare fragile sites. *Cancer Lett.* **232,** 13–26 (2006).
23. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Rev. Cancer* **11,** 450–457 (2011).
24. Weisenberger, D. J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. *Nature Genet* **38,** 787–793 (2006).
25. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
This paper consolidates the common phenotypes that evolve in neoplastic cells of all types.
26. Siegmund, K. D., Marjoram, P., Woo, Y. J., Tavare, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106,** 4828–4833 (2009).
27. Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J. & Mitra, R. D. Intra-tumor heterogeneity of *MLH1* promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res.* **37,** 4603–4612 (2009).
28. Aktipis, C. A., Kwan, V. S. Y., Johnson, K. A., Neuberg, S. L. & Maley, C. C. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. *PLoS ONE* **6,** e261000 (2011).
29. Beerenwinkel, N. *et al.* Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.* **3,** e225 (2007).
30. de Visser, J. A. & Rozen, D. E. Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics* **172,** 2093–2100 (2006).
31. Leedham, S. J. *et al.* Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. *Gut* **57,** 1041–1048 (2008).
32. Navin, N. *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
Single-cell sequencing revealed the clonal structure of two breast cancers.
33. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469,** 356–361 (2011).
Single-cell genetic analyses and xenografts revealed the clonal architecture within acute lymphoblastic leukaemia stem-cell populations and demonstrated repeated independent acquisition of copy number changes within the same neoplasm.
34. Tsao, J. L. *et al.* Colorectal adenoma and cancer divergence. Evidence of multilineage progression. *Am. J. Pathol.* **154,** 1815–1824 (1999).
35. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genet.* **38,** 468–473 (2006).
36. Sidransky, D. *et al.* Clonal expansion of p53 mutant cells is associated with brain tumour progression. *Nature* **355,** 846–847 (1992).
37. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
38. Gould, S. J. & Eldredge, N. Punctuated equilibrium comes of age. *Nature* **366,** 223–227 (1993).
39. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
40. Campbell, P. J. *et al*. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105,** 13081–13086 (2008).
Deep sequencing revealed rare (frequency <0.001) intermediate genotypes between the common clones in leukaemias (using immunoglobulin rearrangements as surrogate mutations).
41. Aguirre-Ghiso, J. A. Models, mechanisms and clinical evidence for cancer dormancy. *Nature Rev. Cancer* **7,** 834–846 (2007).
42. Isoda, T. *et al.* Immunologically silent cancer clone transmission from mother to offspring. *Proc. Natl Acad. Sci. USA* **106,** 17882–17885 (2009).
43. Welsh, J. S. Contagious cancer. *Oncologist* **16,** 1–4 (2011).
44. Gatenby, R. A. & Gillies, R. J. A microenvironmental model of carcinogenesis. *Nature Rev. Cancer* **8,** 56–61 (2008).
45. Bierie, B. & Moses, H. L. Tumour microenvironment: TGFβ: the molecular Jekyll and Hyde of cancer. *Nature Rev. Cancer* **6,** 506–520 (2006).
46. Lathia, J. D., Heddleston, J. M., Venere, M. & Rich, J. N. Deadly teamwork: neural cancer stem cells and the tumor microenvironment. *Cell Stem Cell* **8,** 482–485 (2011).
47. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).
This paper identified natural selection as a driving force in carcinogenesis and identified tissue architecture as a cancer suppressor, and posited an immortal strand of DNA in tissue stem cells.
48. Anderson, A. R., Weaver, A. M., Cummings, P. T. & Quaranta, V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell* **127,** 905–915 (2006).
49. Chen, J., Sprouffske, K., Huang, Q. & Maley, C. C. Solving the puzzle of metastasis: the evolution of cell migration in neoplasms. *PLoS ONE* **6,** e17933 (2011).
50. Mazzone, M. *et al.* Heterozygous deficiency of *PHD2* restores tumor oxygenation and inhibits metastasis via endothelial normalization. *Cell* **136,** 839–851 (2009).
51. Gilbert, L. A. & Hemann, M. T. DNA damage-mediated induction of a chemoresistant niche. *Cell* **143,** 355–366 (2010).
52. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105,** 4283–4288 (2008).
53. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
54. Sprouffske, K., Pepper, J. W. & Maley, C. C. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prev. Res.* **4,** 1135–1144 (2011).
55. Greaves, M. F., Maia, A. T., Wiemels, J. L. & Ford, A. M. Leukemia in twins: lessons in natural history. *Blood* **102,** 2321–2333 (2003).
56. Bateman, C. M. *et al.* Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115,** 3553–3558 (2010).
57. Oosterhuis, J. W. & Looijenga, L. H. Testicular germ-cell tumours in a broader perspective. *Nature Rev. Cancer* **5,** 210–222 (2005).
58. Grant, P. R. & Grant, B. R. *How and Why Species Multiply* (Princeton Univ. Press, 2008).
59. Durinck, S. *et al.* Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1,** 137–143 (2011).
60. Gonzalez-Garcia, I., Sole, R. V. & Costa, J. Metapopulation dynamics and spatial heterogeneity in cancer. *Proc. Natl Acad. Sci. USA* **99,** 13085–13089 (2002).
61. Clark, J. *et al.* Complex patterns of *ETS* gene alteration arise during cancer development in the human prostate. *Oncogene* **27,** 1993–2003 (2008).
62. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
63. Allred, D. C. *et al.* Ductal carcinoma *in situ* and the emergence of diversity during breast cancer evolution. *Clin. Cancer Res.* **14,** 370–378 (2008).
64. Park, S. Y., Gonen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
65. Merlo, L. M. *et al.* A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev. Res.* **3,** 1388–1397 (2010).
66. Dick, J. E. Stem cell concepts renew cancer research. *Blood* **112,** 4793–4807 (2008).
67. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
68. Greaves, M. Cancer stem cells renew their impact. *Nature Med.* **17,** 1046–1048 (2011).
69. Rosen, J. M. & Jordan, C. T. The increasing complexity of the cancer stem cell paradigm. *Science* **324,** 1670–1673 (2009).
70. Greaves, M. Cancer stem cells: back to Darwin? *Semin. Cancer Biol.* **20,** 65–70 (2010).
71. Gupta, P. B. *et al.* Identification of selective inhibitors of cancer stem cells by high-throughput screening. *Cell* **138,** 645–659 (2009).
72. Jamieson, C. H. *et al.* Granulocyte–macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.* **351,** 657–667 (2004).
73. Akala, O. O. *et al.* Long-term haematopoietic reconstitution by $Trp53^{-/-}$ $p16^{Ink4a-/-}p19^{Arf-/-}$ multipotent progenitors. *Nature* **453,** 228–232 (2008).
74. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. *Nature* **442,** 818–822 (2006).
75. Olivier, M. & Taniere, P. Somatic mutations in cancer prognosis and prediction: lessons from *TP53* and *EGFR* genes. *Curr. Opin. Oncol.* **23,** 88–92 (2011).
76. Mizuno, H., Spike, B. T., Wahl, G. M. & Levine, A. J. Inactivation of *p53* in breast cancers correlates with stem cell transcriptional signatures. *Proc. Natl Acad. Sci. USA* **107,** 22745–22750 (2010).
77. Cicalese, A. *et al.* The tumor suppressor *p53* regulates polarity of self-renewing divisions in mammary stem cells. *Cell* **138,** 1083–1095 (2009).
78. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456,** 593–598 (2008).
New xenograft methods revealed that cancer stem cells are common cell types in melanoma.
79. Pece, S. *et al.* Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell* **140,** 62–73 (2010).
80. Notta, F. *et al.* Evolution of human *BCR–ABL1* lymphoblastic leukaemia-initiating cells. *Nature* **469,** 362–367 (2011).
81. Clappier, E. *et al.* Clonal selection in xenografted human T cell acute

lymphoblastic leukemia recapitulates gain of malignancy at relapse. *J. Exp. Med.* **208,** 653–661 (2011).
82. Frank, N. Y., Schatton, T. & Frank, M. H. The therapeutic promise of the cancer stem cell concept. *J. Clin. Invest.* **120,** 41–50 (2010).
83. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. *Nature Biotechnol.* **25,** 1315–1321 (2007).
84. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
85. Piccirillo, S. G. M. *et al.* Distinct pools of cancer stem-like cells coexist within human glioblastomas and display different tumorigenicity and independent genomic evolution. *Oncogene* **28,** 1807–1811 (2009).
86. Solit, D. & Sawyers, C. L. How melanomas bypass new therapy. *Nature* **468,** 902–903 (2010).
87. Goff, D. & Jamieson, C. Cycling toward elimination of leukemic stem cells. *Cell Stem Cell* **6,** 296–297 (2010).
88. Sharma, S. V. *et al.* A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141,** 69–80 (2010).
89. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Med.* **17,** 297–303 (2011).
90. Sawyers, C. L. Shifting paradigms: the seeds of oncogene addiction. *Nature Med.* **15,** 1158–1161 (2009).
91. Graham, S. M. *et al.* Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 *in vitro*. *Blood* **99,** 319–325 (2002).
92. Turke, A. B. *et al.* Preexistence and clonal selection of *MET* amplification in *EGFR* mutant NSCLC. *Cancer Cell* **17,** 77–88 (2010).
93. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145,** 30–38 (2011).
94. Zhang, B. *et al.* Effective targeting of quiescent chronic myelogenous leukemia stem cells by histone deacetylase inhibitors in combination with imatinib mesylate. *Cancer Cell* **17,** 427–442 (2010).
95. Duy, C. *et al. BCL6* enables Ph^+ acute lymphoblastic leukaemia cells to survive *BCR–ABL1* kinase inhibition. *Nature* **473,** 384–388 (2011).
96. Pienta, K. J., McGregor, N., Axelrod, R. & Axelrod, D. E. Ecological therapy for cancer: defining tumors using an ecosystem paradigm suggests new opportunities for novel cancer treatments. *Trans. Oncol.* **1,** 158–164 (2008).
97. Calabrese, C. *et al.* A perivascular niche for brain tumor stem cells. *Cancer Cell* **11,** 69–82 (2007).
98. Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nature Med.* **17,** 320–329 (2011).
99. Wargo, A. R., Huijben, S., de Roode, J. C., Shepherd, J. & Read, A. F. Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. *Proc. Natl Acad. Sci. USA* **104,** 19914–19919 (2007).
100. Gatenby, R. A., Silva, A. S., Gillies, R. J. & Frieden, B. R. Adaptive therapy. *Cancer Res.* **69,** 4894–4903 (2009).
Dosing to maintain tumour size prolonged survival far longer than high-dose therapy in a mouse xenograft model.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The research of M.G. is supported by Leukaemia & Lymphoma Research UK and The Kay Kendall Leukaemia Fund. The research of C.M. is supported by Research Scholar Grant #117209-RSG-09-163-01-CNE from the American Cancer Society and NIH grants P01 CA91955, U54 CA143803, R01 CA149566 and R01 CA140657. The authors thank C.Cooper and J.Clark for the use of the image in Fig. 4.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to M.G. (mel.greaves@icr.ac.uk).

REVIEW

doi:10.1038/nature10763

The contribution of bone to whole-organism physiology

Gérard Karsenty[1] & Mathieu Ferron[1]

The mouse genetic revolution has shown repeatedly that most organs have more functions than expected. This has led to the realization that, in addition to a molecular and cellular approach, there is a need for a whole-organism study of physiology. The skeleton is an example of how a whole-organism approach to physiology can broaden the functions of a given organ, reveal connections of this organ with others such as the brain, pancreas and gut, and shed new light on the pathogenesis of degenerative diseases affecting multiple organs.

There are two definitions of physiology. In its most modern incarnation, physiology focuses on signalling and gene expression events occurring within cells as they relate to a given function and is often referred to as molecular and cellular physiology. An older physiology deals with events occurring and molecules acting outside the cell. It can be called whole-organism physiology and aims at identifying interactions taking place between organs through what Claude Bernard called the "*milieu intérieur*"[1]. Its study relies on three basic principles. First, the premise of whole-organism physiology is that no function is determined by one organ alone; this hypothesis has been verified multiple times through the use of model organisms[2–4]. Second, homeostasis, a principle positing that different organs exert opposite influences on the same function to regulate it tightly, applies to most physiological functions[5]. Closely related to homeostasis is the cardinal rule of endocrinology, that is feedback regulation: a regulated organ talks back to a regulating one to limit its influence[6]. Third is the principle in which whole-organism physiology resembles any other aspect of biology: regulatory molecules appear during evolution with the functions they regulate, not millions of years afterwards.

In vertebrates the modern study of whole-organism physiology uses two main tools. The experimental one, which favoured the renaissance of this physiology, is mouse genetics. This has allowed us to decipher, one gene at a time, how organs influence each other. For the focus of this Review, namely skeleton physiology, mouse genetics has proved to be an extremely reliable tool, possibly because bone is one of the last tissues to appear during evolution, and consequently most genes involved in skeletal biology have conserved functions from mice to humans. The second tool is the observation of internal medicine: mirror images of many physiological processes may be found in disease symptoms or drug side effects. Both tools are equally useful in extending the range of functions of the skeleton and our understanding of their molecular bases.

In this Review, we show how a whole-organism approach to physiology has modified our view of the skeleton. To that end, we describe work pointing towards a coordinated regulation of bone mass, energy metabolism and fertility; a central control of bone mass; and the roles of leptin, serotonin, insulin and osteocalcin in these pathways.

The skeleton and whole-organism physiology

A strategy that could be used to determine whether the skeleton is influenced by organs that are not classically associated with it and whether it influences other organs is to confront features of bone that are unique to clinical and experimental observations. This approach may indicate physiological functions that could affect or be influenced by bone. The reality of such interactions could then be tested genetically.

A striking feature of bone is that it is the only tissue that contains a cell type, the osteoclast, whose unique function is to resorb (destroy) the host tissue[7]. This does not occur at random but rather in the context of a true homeostatic function, which is called bone modelling during childhood and remodelling during adulthood. This function, hereafter referred to as bone (re)modelling, is characterized by alternating phases of destruction by osteoclasts and bone formation by osteoblasts[8]. Bone modelling allows longitudinal growth, without which most vertebrates could not ambulate and, therefore, could not live. It is, by definition, a survival function. Bone (re)modelling occurs daily in multiple locations in an organ covering a very large surface area. Both the cellular events it entails and the surface area of the organ in which they occur suggest there is a high energetic cost. Clinical observations add credibility to this view of bone (re) modelling as an energy-demanding process. Specifically, the absence of food — that is, energy — intake, as in patients with anorexia nervosa, causes a near-total arrest of growth in children and low bone mass in adults[9,10]. Moreover, in a manner unrelated to food intake, gonadal failure leads to low bone mass in both sexes, thereby suggesting a link between bone mass accrual and fertility[11,12]. Although compelling, such clinical evidence remained correlative. It became an incentive for laboratory investigation because of an experimental observation that was striking because it was unexpected. Osteocalcin was, at the time the encoding gene was inactivated in the mouse, the only osteoblast-specific secreted protein[13]. This by itself justified the study of its function *in vivo* with the hope of learning more about bone biology. Surprisingly, osteocalcin deletion resulted in mice that were abnormally obese and that bred poorly[14,15]. These phenotypes suggested that, in ways that still need to be explained, bone was affecting fat accumulation and possibly other aspects of energy metabolism and reproduction.

Taken together, this view of bone (re)modelling, clinical observations and the phenotypes of mice deficient in the osteocalcin gene suggested that there might have been a coordinated regulation of bone mass or growth, energy metabolism and reproduction. Because this hypothesis was triggered by the energy cost of bone (re)modelling, one would expect that the hormones orchestrating it would appear with the skeleton during evolution. Several laboratories have tested this hypothesis in the past 10 years.

[1]Department of Genetics and Development, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA.

A whole-organism physiology view of leptin

For a bone biologist in the late 1990s, it seemed easier to explore the aforementioned assumption by studying a hormone for which the biology has been characterized and the receptor identified. This hormone was leptin, an adipocyte-specific molecule identified for its ability to limit appetite and to favour energy expenditure and reproduction, two tenets of the overarching hypothesis[16–18]. Remarkably, and this is important in view of its function described here, leptin does not appear during evolution with any aspect of energy metabolism or reproduction, but is instead associated with bone (re)modelling. This would be mere coincidence if *ob*/*ob* (obese) or *db*/*db* (diabetic) mice, which lack leptin or its receptor, respectively, did not demonstrate a high bone mass because of a massive increase in bone formation[19]. This is a biological tour de force because these mice have no functional gonads, a situation that would otherwise increase bone resorption and decrease bone mass. Subsequently, leptin regulation of bone mass was verified in sheep and humans[20–22]. The high bone mass (despite hypogonadism) observed in the absence of leptin signalling and the fact that leptin is a vertebrate invention suggested that bone was a major target of this hormone; a model of partial gain-of-function of leptin signalling allowed this hypothesis to be tested. To mediate its functions, leptin binds to a receptor (Lepr) linked to the tyrosine kinase Jak2. Leptin binding activates Jak2, resulting in the phosphorylation of several residues on Lepr. One of these, Tyr 985, binds Socs3, an event that attenuates signalling through Lepr[23] (Fig. 1). Accordingly, mutation of this residue in Lepr (Y985L) in *l*/*l* mice, which are homozygous for this substitution, results in a partial gain of function of leptin signalling[24]. Because the increase in signalling is only partial, phenotypes displayed by *l*/*l* mice reveal one or more functions of leptin that require its lowest threshold of signalling[24]. Hence, *l*/*l* mice breed normally and have normal appetite when fed on normal chow, but are osteoporotic; these observations suggest that the threshold of leptin signalling necessary to affect bone mass is lower than that needed to affect appetite and reproduction[25].

Several genetic pieces of evidence indicate that leptin acts centrally to inhibit the accrual of bone mass. The most convincing argument, although not the only one, is that a neuron-specific deletion of *Lepr* recapitulates the bone phenotype of *ob*/*ob* mice, whereas an osteoblast-specific one does not[25]. This is consistent with what has been shown for other functions of leptin that also occur through a central relay. In broader terms this revealed for the first time a central control of bone mass; its existence has now been verified by other laboratories studying how neuropeptide Y or neuromedin U regulate bone mass[26,27]. Over the years it has been proposed that, in contrast to most of its functions, leptin could also regulate bone mass through local means. This hypothesis is based on studies that injected large amounts (micrograms per day) of leptin into wild-type mice[28] or infused leptin into the third ventricle of *ob*/*ob* mice[29]. That leptin did not decrease the percentage of fat in the latter study, in contrast to expectations, raises questions about the efficacy of these infusions.

Let us concentrate here on the central mode of action of leptin. We needed to know where it signals in the brain to fulfil this function and the regulation of energy metabolism (we viewed these two functions as co-regulated). Initially, it seemed that the hypothalamus was where all the action was taking place. The leptin receptor is highly expressed in ventromedial (VMH) and arcuate neurons of the hypothalamus, and chemical lesioning of these neurons had demonstrated their involvement in the regulation of appetite and bone mass; moreover, leptin infusions in the third ventricle of *ob*/*ob* mice decreased bone mass and appetite only if these hypothalamic neurons were intact[30]. This simple view was shattered by landmark studies showing that the selective inactivation of *Lepr* in VMH or arcuate neurons did not affect appetite or the accrual of bone mass in mice fed on a normal diet, the diet on which *ob*/*ob* mice display hyperphagia[31–33]. One interpretation of these seemingly contradictory results is that they are instead complementary in suggesting that leptin requires the integrity of hypothalamic neurons to regulate appetite and bone mass, but it need not bind to them. In other words, leptin might signal elsewhere in the brain to regulate the synthesis of neurotransmitters that will then act in the hypothalamus. Admittedly, this is a shift in the view of leptin signalling in the brain, but a novel approach was needed because cell-specific deletion of *Lepr* in hypothalamic neurons did not reproduce phenotypes seen in *ob*/*ob* mice fed on normal chow. The fact that patients chronically treated with serotonin reuptake inhibitors can develop low bone mass suggested that serotonin and leptin signalling intersect in the brain[34–36].

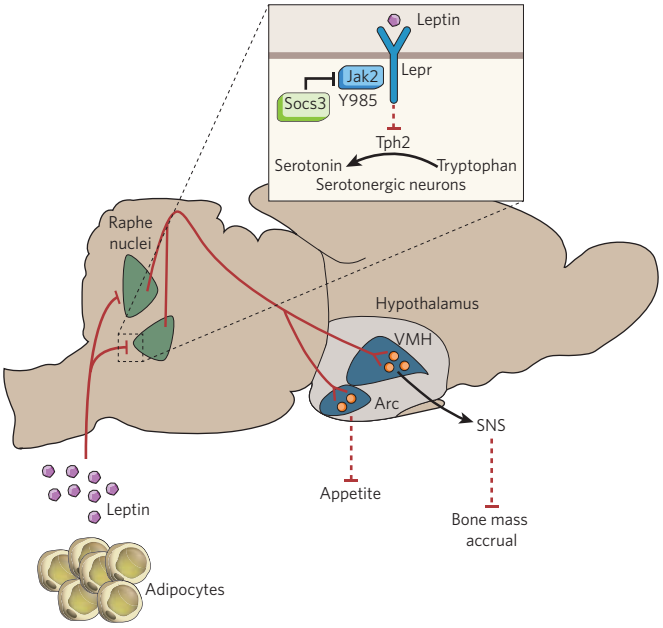


Figure 1 | Leptin co-regulates appetite and bone mass. Leptin signals through its receptor (Lepr), expressed in serotonin-producing neurons of the raphe nuclei in the brainstem, decreasing the expression of tryptophan hydrxylase 2 (*Tph2*), the gene encoding the initial enzyme for serotonin biosynthesis. Signalling by Lepr requires Jak2 and is negatively regulated by Socs3. Serotonin-producing neurons of the raphe nuclei project to the ventromedial (VMH) and the arcuate (Arc) hypothalamic nuclei. Serotonin signalling in VMH neurons decreases the activity of the sympathetic nervous system (SNS), an inhibitor of bone mass accrual. In Arc neurons, serotonin signalling has anti-anorexigenic effects.

In some instances cell-specific gene inactivation can be achieved without using sophisticated techniques. These are favourable circumstances because cell-specific gene deletion techniques are not without risks. This situation presented itself for leptin and for brain-derived serotonin, which is made only in neurons of the raphe nuclei and does not cross the blood–brain barrier[37]. It could therefore be eliminated from the brain by conventional inactivation of tryptophan hydroxylase 2 (Tph2), the rate-limiting enzyme of serotonin synthesis[38]. The fact that *Tph2*$^{-/-}$ mice were distinctly osteoporotic and anorectic established that brain-derived serotonin favours bone mass accrual and appetite and, because it does not cross the blood–brain barrier, identified serotonin as the first neurotransmitter regulating bone mass accrual[31]. Axonal connections between serotonin-producing neurons and hypothalamic neurons can exist and are probably functionally important, because ablations of serotonin receptors in specific neurons of the hypothalamus result in osteoporosis or anorexia in the same way as *Tph2* deletion does[31,39,40]. This work, suggesting a direct brainstem–hypothalamus axis, does not contradict or exclude a more recent study proposing that serotonin signalling in the hypothalamus occurs through an interneuron[41]. This interneuron will need to be identified for the proper evaluation of the respective importance of these two modes of action.

Leptin inhibits *Tph2* expression in, and serotonin release from, brainstem neurons; consequently, serotonin content in *ob*/*ob* hypothalami is abnormally high. Removing one allele of *Tph2* from *ob*/*ob* mice sufficed to normalize their brain serotonin content, with remarkable consequences[31]. Indeed, *ob*/*ob*;*Tph2*$^{+/-}$ mice, despite having no leptin, had normal appetite, energy expenditure, body weight and bone mass. Subsequently, the molecular bases of this leptin–serotonin–hypothalamus axis were deciphered. They include calmodulin signalling in VMH neurons activating the transcription factor cyclic-AMP-response-element-binding protein (CREB), which induces the expression of genes involved in catecholamine synthesis[39]. This is important because the mediator acting as a bridge between leptin signalling in the brain and in bone cells is the sympathetic nervous system acting on osteoblasts through the β_2 adrenergic receptor (Adrb2)[30,42] (Fig. 1) and because β-blockers can limit the risk of osteoporotic fractures[43].

The fact that leptin inhibits both bone mass accrual and appetite is consistent with the hypothesis that bone acquisition, an energy-expensive process, must be linked to energy (food) intake, otherwise the risk of organ failure anywhere else in the body, at the time of a growth spurt for instance, could be high. This broader view of leptin proposes a simple explanation of why this hormone appeared during evolution with bone and when food was scarce.

Osteocalcin and the endocrine nature of bone

Implicit in the hypothesis that there is coordinated regulation of bone mass, energy metabolism and fertility is the notion that bone is not only a recipient of hormonal inputs but also an endocrine organ affecting the other two functions. It was already known that, through fibroblast growth factor 23 (FGF 23), bone acts as an endocrine organ, but the function regulated by FGF23, namely mineral metabolism, is intimately linked to bone health[44,45]. The question asked at this point was different: does bone regulate physiological functions that, a priori, have nothing to do with bone health?

Although this hypothesis was initially formulated on inspection of mice deficient in the osteocalcin gene, it was the study of another gene and of another mouse model that shaped the concept into the form of a genetic pathway. Embryonic stem-cell phosphatase (ESP) is an obscure protein tyrosine phosphatase without any known substrate that is expressed in only three cell types: the embryonic stem cell, the Sertoli cell of the testes and the osteoblast[15]. This restricted pattern of expression justified the study of its function *in vivo*, and for that purpose *Esp* was deleted either in all cells or only in osteoblasts. The observation that both mouse models developed the same phenotypes implied that the biological functions they reveal take place in osteoblasts[15].

The absence of *Esp* in osteoblasts prompted perinatal death of mice, resulting from severe hypoglycaemia. This happened because the functions of this intracellular tyrosine phosphatase are, through its expression in osteoblasts, to inhibit the expression and secretion of insulin by pancreatic β-cells, to limit insulin sensitivity in liver, muscle and white adipose tissue, and to decrease energy expenditure[15]. All of the phenotypes of *Esp*-deficient mice mirrored those observed in mice deficient in the osteocalcin gene, and genetically *Esp* inhibits osteocalcin's functions[15]. The fact that wild-type, but not osteocalcin-gene-deficient, osteoblasts induce insulin secretion from isolated islets established that osteocalcin is a hormone promoting β-cell proliferation and insulin expression and secretion. It also increases insulin sensitivity and energy expenditure[15] (Fig. 2).

Before considering other aspects of osteocalcin biology one needs to address the relevance of these findings to human physiology. To the best of our knowledge no molecule identified as a hormone in the mouse has lost this function in humans. Although no mutation in osteocalcin or its receptor has yet been reported, this assumption seems to extend to osteocalcin, which has become an increasingly accepted biomarker of insulin resistance in human studies[46,47].

Because most hormones have several functions, the next question was whether this was the case for osteocalcin. This was even more relevant because osteocalcin-gene-deficient mice breed poorly. A study of this phenotype revealed that osteocalcin promotes testosterone synthesis by Leydig cells of the testis and fertility in male mice[14]. For that purpose osteocalcin upregulates, in testes but not in ovaries, the synthesis of enzymes necessary for testosterone biosynthesis, without affecting the expression of the gene encoding Cyp19 (the enzyme converting testosterone to oestradiol). This male-specific function of osteocalcin implied that its putative receptor (OSTR) may be expressed only in testes. This sexual dichotomy allowed the identification of GPRC6A, a G-protein-coupled receptor, as an osteocalcin receptor, that is expressed in mice and humans in Leydig cells of the testis, but not in follicular cells of the ovary[14] (Fig. 2). A pioneering study had previously proposed that GPR6CA could be an osteocalcin receptor, although no binding experiments were done[48].

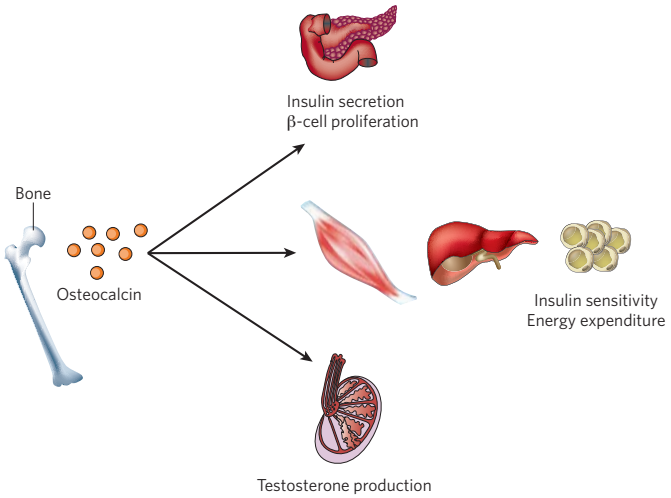


Figure 2 | Osteocalcin, a bone-derived multifunctional hormone. Undercarboxylated osteocalcin stimulates insulin secretion and β-cell proliferation in the pancreas, energy expenditure by muscle, and insulin sensitivity in adipose tissue, muscle and liver. In addition, it promotes male fertility by stimulating testosterone synthesis in Leydig cells of the testis through the activation of its receptor, GPRC6A, in these cells.

Osteocalcin puts bone in the thick of things

If the notion that bone affects energy metabolism and reproduction was implicit in the overall hypothesis, what emerged next — that bone is an endocrine hub on which several hormones converge to recruit osteocalcin as their ultimate relay — was not expected. To appreciate this notion, which was an outcome of the study of the relationship between ESP and osteocalcin, one needs to understand how osteocalcin is modified post-translationally.

Osteocalcin is carboxylated on three glutamic acid residues, a modification that confers on proteins a high affinity for minerals such as the hydroxyapatite crystal present in the mineralized bone matrix[13]. Nevertheless, a small but measurable proportion of undercarboxylated osteocalcin is found in the serum[49], indicating that either some osteocalcin is secreted in an incompletely carboxylated form by osteoblasts or that osteocalcin becomes decarboxylated outside the cell. This was an important question to address because cell-based and *in vivo* investigations had shown that it is the undercarboxylated form of osteocalcin (the one in which Glu 13 is not carboxylated) that is active on β-cells and Leydig cells[14,50].

The search for substrates of ESP identified the insulin receptor as such a molecule[51]. The *Esp*-null mouse, therefore, is a gain-of-function model of insulin signalling in osteoblasts. Subsequently, two groups working independently showed that mice lacking the insulin receptor in osteoblasts ($InsR_{osb}^{-/-}$ mice) were glucose intolerant and

insulin insensitive when fed on normal chow; that is, they were a phenocopy of the osteocalcin-gene-deficient mice[51,52]. Because mice lacking the insulin receptor in skeletal muscle or white adipose tissue do not display glucose intolerance when fed on a normal diet[53,54], insulin must act in additional tissues to achieve glucose homeostasis. The fact that bone is such a tissue legitimizes the notion that this tissue is necessary for glucose homeostasis. In addition, $InsR_{osb}^{-/-}$ mice had significantly less biologically active (undercarboxylated) osteocalcin in their sera, revealing that insulin signalling in osteoblasts is a determinant of osteocalcin bioactivity[53,54]. In a manner that is both elegant and economical, insulin uses the interplay between osteoblasts and osteoclasts for that purpose. Specifically, insulin inhibits the expression in osteoblasts of the gene encoding osteoprotegerin (*Opg*)[51], which hampers osteoclast differentiation. In other words, insulin signalling in osteoblasts favours bone resorption, a process that occurs at pH 4.5 (ref. 55). Acidic pH is the only mechanism known to achieve decarboxylation of proteins[56], therefore, bone resorption decarboxylates and activates osteocalcin[51]. Thus, in a feedforward loop, insulin signalling in osteoblasts promotes its own secretion by activating osteocalcin (Fig. 3), and mice and humans in which bone resorption is genetically impaired show a decrease in the undercarboxylated form of osteocalcin, resulting in glucose intolerance[51]. The functional equivalent of ESP in human osteoblasts is protein tyrosine phosphatase 1B (PTP-1B), a tyrosine phosphatase previously known for its ability to inactivate the insulin receptor in other cell types[51,57–59].

If insulin enhances osteocalcin activity, which in turn favours insulin secretion, how do vertebrates prevent hypoglycaemia? The answer is leptin[60]. Young *ob*/*ob* mice, before they become obese, have no molecular or metabolic evidence of insulin resistance, yet they are markedly hyperinsulinaemic and hypoglycaemic, indicating that a function of leptin is to inhibit insulin secretion[3]. As is true of its other functions, leptin by and large does not act locally on islets to affect insulin secretion, but through a neuronal relay[60]. Because the sympathetic nervous system acting through Adrb2 expressed in osteoblasts mediates the inhibition of bone mass accrual by leptin, it was tempting to ask whether the same pathway might inhibit insulin secretion by modulating osteocalcin expression or activity. Molecular evidence showed that the sympathetic tone mediates the upregulation of *Esp* expression by leptin and that *ob*/*ob* mice have more active osteocalcin in their circulation[60]. Three pieces of genetic evidence demonstrate that leptin regulation of osteocalcin activity occurs through sympathetic signalling in osteoblasts. First, mice lacking *Adrb2* in osteoblasts are hypoglycaemic and hyperinsulinaemic. Second, mice lacking one copy of the leptin gene and one copy of *Adrb2* in osteoblasts are also hyperinsulinaemic and hypoglycaemic. Third, *ob*/*ob* mice that lack the osteocalcin gene have normal expression of the insulin genes and are normo-insulinaemic for significantly longer than *ob*/*ob* mice[60].

Altogether, the metabolic functions of osteocalcin and their regulation by leptin and insulin reveal how intertwined energy metabolism and bone (re)modelling are. They also underscore the importance of the interplay between osteoblasts and osteoclasts for glucose metabolism and indicate that the regulation of osteocalcin activity occurs at both the transcriptional and post-translational levels.

Energy metabolism, bone mass and bone diseases

Energy metabolism is a multistep process initiated by food absorption along the gastrointestinal tract. Thus, in considering whether there is coordinated regulation of bone mass and energy metabolism one needs to determine whether any part or function of the gastrointestinal tract influences the accrual of bone mass. This is important in view of the fact that the gastrointestinal tract is an endocrine organ secreting several hormones whose functions are not all known. Clinically there is reason to ask whether the gastrointestinal tract affects bone mass: osteoporosis is often seen in patients with inflammatory bowel diseases[61,62].

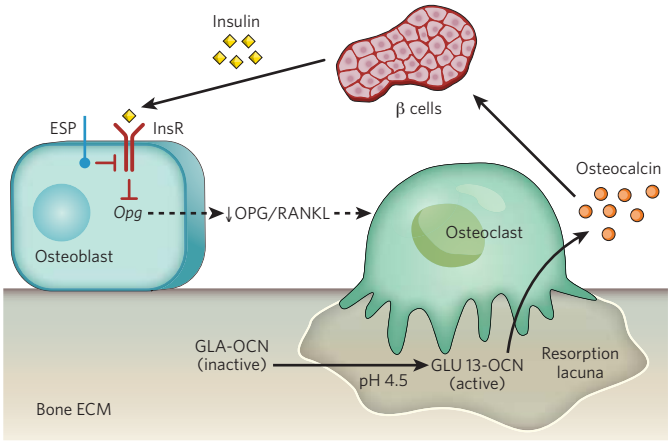


Figure 3 | A feedforward loop links insulin, bone resorption and osteocalcin activity. Insulin signalling in osteoblasts decreases the expression of *Opg*. The decrease in the ratio of osteoprotegerin (OPG) to receptor activator of nuclear factor–κB ligand (RANKL) increases bone resorption by osteoclasts. The acidic pH (4.5) in resorption lacunae decarboxylates (that is, activates) osteocalcin (GLA-OCN) stored in the bone extracellular matrix. Undercarboxylated active osteocalcin (GLU13-OCN) then stimulates insulin secretion by the β-cells of the pancreatic islets and promotes insulin sensitivity in peripheral organs. ECM, extracellular matrix; InsR, insulin receptor.

The study of mouse models of two human genetic diseases, the Coffin–Lowry syndrome (a learning disability syndrome with low bone mass) and the skeletal manifestations of neurofibromatosis 1 (NF1), established the existence of a connection between food absorption and the accrual of bone mass. In these disorders, bone mass abnormalities are due to a decrease or an increase, respectively, in the activity of activating transcription factor 4 (ATF4), a transcription factor that favours bone formation, in part by promoting amino acid import into osteoblasts[63,64]. One could, therefore, put this function of ATF4 to work to correct the skeletal manifestations of Coffin–Lowry syndrome or NF1 in the mouse by modulating protein content in their diets[63] (Fig. 4). Another study looked at the influence of the gastrointestinal tract on mineral metabolism. Given that bone is a mineralized tissue and that the mineral crystal in the bone is made of calcium and phosphate, it is not surprising that significant variations in calcium or phosphate metabolism affect bone mass. One reason for that is that a high extracellular concentration of calcium inhibits secretion by parathyroid glands of parathyroid hormone (PTH), whose physiological role is to increase bone resorption[65]. A study relying on mouse genetics and human pathology showed that the high acidity (low pH) present in the stomach, the entry point of the gastrointestinal tract, is required for proper calcium absorption and, therefore, for normal PTH secretion and bone resorption[66] (Fig. 4). This established a simple connection between the gastrointestinal tract and bone mass that could be exploited for therapeutic purposes. These examples of an influence of the gastrointestinal tract on bone mass raised the possibility that hormones made by the gastrointestinal tract regulate bone mass. A peculiar combination of scientific frustration and luck showed that this was not only the case, but also that this regulation is potentially important for the treatment of the most frequent bone degenerative disease, osteoporosis.

The cell-surface molecule Lrp5 (low-density-lipoprotein receptor related protein 5), despite having no identifiable ligand and no signalling pathway, is remarkably important in bone biology for medical reasons. First, *LRP5* is mutated in two human diseases. In osteoporosis pseudoglioma (OPPG), a disease characterized by the appearance of osteoporosis several years after birth, it is inactivated[67]; in high bone mass syndrome (HBM) there is a missense mutation in *LRP5* that is thought to be a gain-of-function mutation[68,69]. Patients

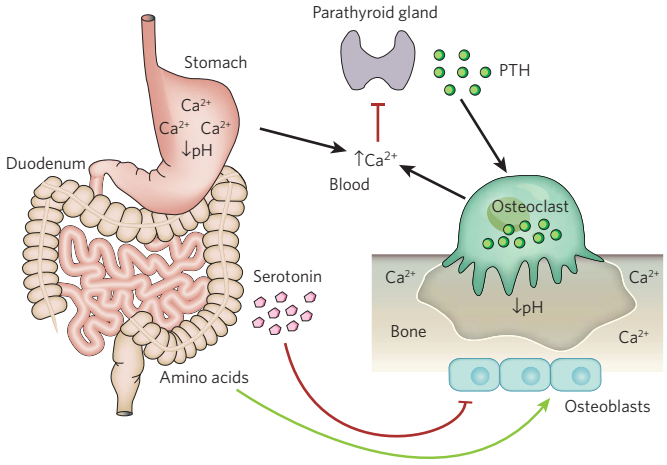


Figure 4 | Interactions between the gastrointestinal tract and bone mass. Stomach acidity (low pH) is required for the proper absorption of calcium (Ca^{2+}) and is, therefore, essential to maintain normal levels of serum calcium. Serum calcium, in turn, negatively regulates secretion from the parathyroid gland of PTH, a hormone that stimulates osteoclast differentiation and bone resorption. Bone resorption by osteoclasts also occurs at low pH and contributes to the maintenance of serum calcium. Peripheral serotonin is produced by the duodenum and inhibits bone formation by osteoblasts, whereas dietary intake of amino acids (proteins) favours collagen synthesis by osteoblasts.

harbouring the latter mutation have no detectable symptoms before adulthood, again arguing against a developmental defect. In fact, the main symptom is that post-menopausal female patients with HBM do not develop osteoporosis[68,69]. Another reason to draw much attention to Lrp5 is that it affects bone mass only by acting on bone formation[70]. In other words, the Lrp5-dependent signalling pathway holds the key to an anabolic treatment for osteoporosis, the ultimate objective for the most frequent bone degenerative disease.

Lrps are known to be able to bind multiple ligands; nevertheless, because of its homology to Arrow[71] (a co-receptor in *Drosophila* for Wingless, a homologue of Wnt) it is as a potential co-receptor for Wnt proteins that Lrp5 has been studied for the longest time. This was a reasonable assumption to test, and cell culture experiments showed repeatedly that Lrp5, like other Lrps, can function as a Wnt co-receptor *in vitro* and in controlling eye vascularization[71–75]. However, when it came to its role in bone, there were discrepancies between expectations and results obtained through the study of animal models. A few examples follow. *Lrp5*-null osteoblasts proliferate poorly *in vivo,* but normally in cell culture[76]. Because Lrp5 is a receptor, this observation is inconsistent with the notion that its signalling is affected in osteoblasts. Furthermore, microarray analyses showed that genes regulated by Lrp5 (cell-cycle regulators) and by the canonical Wnt signalling (*Opg*) in bone specimens were different[77,78]. Accordingly, in mice, inactivation of the canonical Wnt signalling pathway in osteoblasts or osteocytes does not affect bone formation but bone resorption instead[78,79]. In humans, patients with HBM who have been followed up for more than 20 years have not yet developed bone tumours (M. Kassem, personal communication), even though many Wnt family members are aggressive oncoproteins[80].

In an unforeseen turn of events, attempts to understand these discrepancies have brought serotonin back into the picture. While studying the role of brain-derived serotonin through the inactivation of *Tph2*, the function of Tph1, the enzyme responsible for peripheral serotonin synthesis, was investigated as a negative control. *Tph1*$^{-/-}$ mice had already been generated when it was realized that the most overexpressed gene in the absence of *Lrp5* was in fact *Tph1* (ref. 76). From being an internal control in another project *Tph1* became, overnight, the topic of an independent one.

Two groups have generated mouse models of *Lrp5* deletion in a cell-specific manner and have obtained different results[76,81]. The reason for this discrepancy is unknown. Because we ourselves generated one of these models, we believe it would be inappropriate to use the tribune provided by this review to make our case. It is also unnecessary, because, as important as mouse genetics is, it remains a surrogate for human genetics, and so far all patients with OPPG or HBM for whom measurements of circulating serotonin have been published have shown high and low circulating levels of serotonin, respectively[76,82–84]. Similarly, serotonin-producing tumours cause osteoporosis[85], and patients with osteoporosis have high circulating serotonin levels[86] (Fig. 4). As a result, the regulation of bone mass by serotonin has now acquired a (medical) life of its own and deserves to be studied because it holds great promise for the anabolic treatment of osteoporosis. This is even more true given that two proof-of-principle studies have shown that an inhibitor of serotonin synthesis in the gut, shown to be safe in humans in a phase I clinical trial[87,88], cured bone disease in *Lrp5*-null mice and gonadectomy-induced osteoporosis in rodents. In both cases this was achieved through a purely anabolic mode of action, just as Lrp5 affects bone mass[89,90].

Perspective

The most appropriate way to look forward is to formulate the questions that will identify the next frontiers in this aspect of physiology. For instance, are all of the functions of osteocalcin known? What are the genes downstream of osteocalcin in all of its target cells? Is osteocalcin the only bone-derived hormone affecting energy metabolism? Is there a counterpart of osteocalcin in females that affects their reproduction? We will also venture a general and humbling statement: if we have learned so much in so brief a time about the physiology of one organ, it is likely that there are many aspects of whole-organism physiology to be discovered through the use of model organisms. ■

1. Bernard, C. *Introduction à l'Etude de la Médecine Expérimentale* (Flammarion, 1865).
2. Gurney, C. W. Erythropoietin, erythropoiesis, and the kidney. *J. Am. Med. Assoc.* **173,** 1828–1829 (1960).
3. Friedman, J. M. & Halaas, J. L. Leptin and the regulation of body weight in mammals. *Nature* **395,** 763–770 (1998).
4. Chen, S. K. *et al.* Hematopoietic origin of pathological grooming in *Hoxb8* mutant mice. *Cell* **141,** 775–785 (2010).
5. Cannon, W. B. *The Wisdom of the Body* (Norton, 1932).
6. Monod, J. & Jacob, F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26,** 389–401 (1961).
7. Teitelbaum, S. L. Bone resorption by osteoclasts. *Science* **289,** 1504–1508 (2000).
8. Ducy, P., Schinke, T. & Karsenty, G. The osteoblast: a sophisticated fibroblast under central surveillance. *Science* **289,** 1501–1504 (2000).
9. Legroux-Gerot, I., Vignau, J., Collier, F. & Cortet, B. Bone loss associated with anorexia nervosa. *Joint Bone Spine* **72,** 489–495 (2005).
10. Misra, M. & Klibanski, A. The neuroendocrine basis of anorexia nervosa and its impact on bone metabolism. *Neuroendocrinology* **93,** 65–73 (2011).
11. Riggs, B. L., Khosla, S. & Melton, L. J. A unitary model for involutional osteoporosis: estrogen deficiency causes both type I and type II osteoporosis in postmenopausal women and contributes to bone loss in aging men. *J. Bone Miner. Res.* **13,** 763–773 (1998).
12. Riggs, B. L. & Melton, L. J. Involutional osteoporosis. *N. Engl. J. Med.* **314,** 1676–1686 (1986).
13. Hauschka, P. V., Lian, J. B., Cole, D. E. & Gundberg, C. M. Osteocalcin and matrix Gla protein: vitamin K-dependent proteins in bone. *Physiol. Rev.* **69,** 990–1047 (1989).
14. Oury, F. *et al.* Endocrine regulation of male fertility by the skeleton. *Cell* **144,** 796–809 (2011).
15. Lee, N. K. *et al.* Endocrine regulation of energy metabolism by the skeleton. *Cell* **130,** 456–469 (2007).
This paper describes for the first time the endocrine function of bone.
16. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372,** 425–432 (1994).
This report describes the identification of the leptin gene.
17. Halaas, J. L. *et al.* Weight-reducing effects of the plasma protein encoded by the obese gene. *Science* **269,** 543–546 (1995).
18. Chehab, F. F., Lim, M. E. & Lu, R. Correction of the sterility defect in homozygous obese female mice by treatment with the human recombinant leptin. *Nature Genet.* **12,** 318–320 (1996).
19. Ducy, P. *et al.* Leptin inhibits bone formation through a hypothalamic relay:

a central control of bone mass. *Cell* **100,** 197–207 (2000).
20. Pogoda, P. *et al.* Leptin inhibits bone formation not only in rodents, but also in sheep. *J. Bone Miner. Res.* **21,** 1591–1599 (2006).
21. Elefteriou, F. *et al.* Serum leptin level is a regulator of bone mass. *Proc. Natl Acad. Sci. USA* **101,** 3258–3263 (2004).
22. Gibson, W. T. *et al.* Congenital leptin deficiency due to homozygosity for the Δ133G mutation: report of another case and evaluation of response to four years of leptin therapy. *J. Clin. Endocrinol. Metab.* **89,** 4821–4826 (2004).
23. Bjorbaek, C., Elmquist, J. K., Frantz, J. D., Shoelson, S. E. & Flier, J. S. Identification of SOCS-3 as a potential mediator of central leptin resistance. *Mol. Cell* **1,** 619–625 (1998).
24. Bjornholm, M. *et al.* Mice lacking inhibitory leptin receptor signals are lean with normal endocrine function. *J. Clin. Invest.* **117,** 1354–1360 (2007).
25. Shi, Y. *et al.* Dissociation of the neuronal regulation of bone mass and energy metabolism by leptin *in vivo*. *Proc. Natl Acad. Sci. USA* **105,** 20529–20533 (2008).
26. Baldock, P. A. *et al.* Neuropeptide Y knockout mice reveal a central role of NPY in the coordination of bone mass to body weight. *PLoS ONE* **4**, e8415 (2009).
27. Sato, S. *et al.* Central control of bone remodeling by neuromedin U. *Nature Med.* **13,** 1234–1240 (2007).
28. Cornish, J. *et al.* Leptin directly regulates bone cell function *in vitro* and reduces bone fragility *in vivo*. *J. Endocrinol.* **175,** 405–415 (2002).
29. Bartell, S. M. *et al.* Central (ICV) leptin injection increases bone formation, bone mineral density, muscle mass, serum IGF-1, and the expression of osteogenic genes in leptin-deficient *ob/ob* mice. *J. Bone Miner. Res.* **26,** 1710–1720 (2011).
30. Takeda, S. *et al.* Leptin regulates bone formation via the sympathetic nervous system. *Cell* **111,** 305–317 (2002).
This paper reports that leptin regulation of bone mass requires the sympathetic nervous system.
31. Yadav, V. K. *et al.* A serotonin-dependent mechanism explains the leptin regulation of bone mass, appetite, and energy expenditure. *Cell* **138,** 976–989 (2009).
This study identifies brain serotonin as a critical mediator of the central action of leptin.
32. Balthasar, N. *et al.* Leptin receptor signaling in POMC neurons is required for normal body weight homeostasis. *Neuron* **42,** 983–991 (2004).
33. Dhillon, H. *et al.* Leptin directly activates SF1 neurons in the VMH, and this action by leptin is required for normal body-weight homeostasis. *Neuron* **49,** 191–203 (2006).
34. Warden, S. J. *et al.* Psychotropic drugs have contrasting skeletal effects that are independent of their effects on physical activity levels. *Bone* **46,** 985–992 (2010).
35. Bliziotes, M. Update in serotonin and bone. *J. Clin. Endocrinol. Metab.* **95,** 4124–4132 (2010).
36. Kaye, W., Gendall, K. & Strober, M. Serotonin neuronal function and selective serotonin reuptake inhibitor treatment in anorexia and bulimia nervosa. *Biol. Psychiatry* **44,** 825–838 (1998).
37. Mann, J. J. *et al.* Relationship between central and peripheral serotonin indexes in depressed and suicidal psychiatric inpatients. *Arch. Gen. Psychiatry* **49,** 442–446 (1992).
38. Walther, D. J. *et al.* Synthesis of serotonin by a second tryptophan hydroxylase isoform. *Science* **299,** 76 (2003).
This report describes the cloning of the *Tph2* gene, which is responsible for brain serotonin synthesis.
39. Oury, F. *et al.* CREB mediates brain serotonin regulation of bone mass through its expression in ventromedial hypothalamic neurons. *Genes Dev.* **24,** 2330–2342 (2010).
40. Yadav, V. K. *et al.* Leptin-dependent serotonin control of appetite: temporal specificity, transcriptional regulation, and therapeutic implications. *J. Exp. Med.* **208,** 41–52 (2011).
41. Lam, D. D. *et al.* Leptin does not directly affect CNS serotonin neurons to influence appetite. *Cell Metab.* **13,** 584–591 (2011).
42. Elefteriou, F. *et al.* Leptin regulation of bone resorption by the sympathetic nervous system and CART. *Nature* **434,** 514–520 (2005).
43. Bonnet, N. *et al.* Protective effect of beta blockers in postmenopausal women: influence on fractures, bone density, micro and macroarchitecture. *Bone* **40,** 1209–1216 (2007).
44. ADHR Consortium. Autosomal dominant hypophosphataemic rickets is associated with mutations in *FGF23*. *Nature Genet.* **26,** 345–348 (2000).
45. Hori, M., Shimizu, Y. & Fukumoto, S. Fibroblast growth factor 23 in phosphate homeostasis and bone metabolism. *Endocrinology* **152,** 4–10 (2011).
46. Motyl, K. J., McCabe, L. R. & Schwartz, A. V. Bone and glucose metabolism: a two-way street. *Arch. Biochem. Biophys.* **503,** 2–10 (2010).
47. Pollock, N. K. *et al.* Lower uncarboxylated osteocalcin concentrations in children with prediabetes is associated with beta-cell function. *J. Clin. Endocrinol. Metab.* **96,** E1092–E1099 (2011).
48. Pi, M. *et al.* Identification of a novel extracellular cation-sensing G-protein-coupled receptor. *J. Biol. Chem.* **280,** 40201–40209 (2005).
49. Merle, B. & Delmas, P. D. Normal carboxylation of circulating osteocalcin (bone Gla-protein) in Paget's disease of bone. *Bone Miner.* **11,** 237–245 (1990).
50. Ferron, M., Hinoi, E., Karsenty, G. & Ducy, P. Osteocalcin differentially regulates beta cell and adipocyte gene expression and affects the development of metabolic diseases in wild-type mice. *Proc. Natl Acad. Sci. USA* **105,** 5266–5270 (2008).
51. Ferron, M. *et al.* Insulin signaling in osteoblasts integrates bone remodeling and energy metabolism. *Cell* **142,** 296–308 (2010).
This study identifies a feed-forward loop linking insulin and osteocalcin.
52. Fulzele, K. *et al.* Insulin receptor signaling in osteoblasts regulates postnatal bone acquisition and body composition. *Cell* **142,** 309–319 (2010).
In this report, insulin signalling in osteoblasts is identified as a key regulator of bone mass and of energy metabolism.
53. Bluher, M. *et al.* Adipose tissue selective insulin receptor knockout protects against obesity and obesity-related glucose intolerance. *Dev. Cell* **3,** 25–38 (2002).
54. Bruning, J. C. *et al.* A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM without altering glucose tolerance. *Mol. Cell* **2,** 559–569 (1998).
55. Silver, I. A., Murrills, R. J. & Etherington, D. J. Microelectrode studies on the acid microenvironment beneath adherent macrophages and osteoclasts. *Exp. Cell Res.* **175,** 266–276 (1988).
56. Poser, J. W. & Price, P. A. A method for decarboxylation of γ-carboxyglutamic acid in proteins. Properties of the decarboxylated γ-carboxyglutamic acid protein from calf bone. *J. Biol. Chem.* **254,** 431–436 (1979).
57. Elchebly, M. *et al.* Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* **283,** 1544–1548 (1999).
58. Delibegovic, M. *et al.* Liver-specific deletion of protein-tyrosine phosphatase 1B (PTP1B) improves metabolic syndrome and attenuates diet-induced endoplasmic reticulum stress. *Diabetes* **58,** 590–599 (2009).
59. Delibegovic, M. *et al.* Improved glucose homeostasis in mice with muscle-specific deletion of protein-tyrosine phosphatase 1B. *Mol. Cell. Biol.* **27,** 7727–7734 (2007).
60. Hinoi, E. *et al.* The sympathetic tone mediates leptin's inhibition of insulin secretion by modulating osteocalcin bioactivity. *J. Cell Biol.* **183,** 1235–1242 (2008).
61. Miheller, P., Lorinczy, K. & Lakatos, P. L. Clinical relevance of changes in bone metabolism in inflammatory bowel disease. *World J. Gastroenterol.* **16,** 5536–5542 (2010).
62. Rodriguez-Bores, L., Barahona-Garrido, J. & Yamamoto-Furusho, J. K. Basic and clinical aspects of osteoporosis in inflammatory bowel disease. *World J. Gastroenterol.* **13,** 6156–6165 (2007).
63. Elefteriou, F. *et al.* ATF4 mediation of *NF1* functions in osteoblast reveals a nutritional basis for congenital skeletal dysplasiae. *Cell Metab.* **4,** 441–451 (2006).
64. Yang, X. *et al.* ATF4 is a substrate of RSK2 and an essential regulator of osteoblast biology; implication for Coffin–Lowry syndrome. *Cell* **117,** 387–398 (2004).
65. Raisz, L. G. Stimulation of bone resorption by parathyroid hormone in tissue culture. *Nature* **197,** 1015–1016 (1963).
66. Schinke, T. *et al.* Impaired gastric acidification negatively affects calcium homeostasis and bone mass. *Nature Med.***15,** 674–681 (2009).
67. Gong, Y. *et al.* LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. *Cell* **107,** 513–523 (2001).
This paper identifies *LRP5* as the gene mutated in OPPG.
68. Little, R. D. *et al.* A mutation in the LDL receptor-related protein 5 gene results in the autosomal dominant high-bone-mass trait. *Am. J. Hum. Genet.* **70,** 11–19 (2002).
69. Boyden, L. M. *et al.* High bone density due to a mutation in LDL-receptor-related protein 5. *N. Engl. J. Med.* **346,** 1513–1521 (2002).
This study identifies a mutation in *LRP5* as causing HBM syndrome in humans.
70. Kato, M. *et al. Cbfa*1-independent decrease in osteoblast proliferation, osteopenia, and persistent embryonic eye vascularization in mice deficient in Lrp5, a Wnt coreceptor. *J. Cell Biol.* **157,** 303–314 (2002).
71. Tamai, K. *et al.* LDL-receptor-related proteins in Wnt signal transduction. *Nature* **407,** 530–535 (2000).
72. Hay, E. *et al.* Interaction between LRP5 and Frat1 mediates the activation of the Wnt canonical pathway. *J. Biol. Chem.* **280,** 13616–13623 (2005).
73. Tolwinski, N. S. *et al.* Wg/Wnt signal can be transmitted through Arrow/LRP5,6 and Axin independently of Zw3/Gsk3β activity. *Dev. Cell* **4,** 407–418 (2003).
74. Li, X. *et al.* Sclerostin binds to LRP5/6 and antagonizes canonical Wnt signaling. *J. Biol. Chem.* **280,** 19883–19887 (2005).
75. Ye, X. *et al.* Norrin, frizzled-4, and Lrp5 signaling in endothelial cells controls a genetic program for retinal vascularization. *Cell* **139,** 285–298 (2009).
76. Yadav, V. K. *et al.* Lrp5 controls bone formation by inhibiting serotonin synthesis in the duodenum. *Cell* **135,** 825–837 (2008).
77. Jackson, A. *et al.* Gene array analysis of Wnt-regulated genes in C3H10T1/2 cells. *Bone* **36,** 585–598 (2005).
78. Glass, D. A. *et al.* Canonical Wnt signaling in differentiated osteoblasts controls osteoclast differentiation. *Dev. Cell* **8,** 751–764 (2005).
79. Kramer, I. *et al.* Osteocyte Wnt/β-catenin signaling is required for normal bone homeostasis. *Mol. Cell. Biol.* **30,** 3071–3085 (2010).
80. Reya, T. & Clevers, H. Wnt signalling in stem cells and cancer. *Nature* **434,** 843–850 (2005).
81. Cui, Y. *et al.* Lrp5 functions in bone to regulate bone mass. *Nature Med.* **17,** 684–691 (2011).
82. Saarinen, A. *et al.* Low density lipoprotein receptor-related protein 5

(*LRP5*) mutations and osteoporosis, impaired glucose metabolism and hypercholesterolaemia. *Clin. Endocrinol.* **72,** 481–488 (2010).
83. Frost, M. *et al.* Patients with high-bone-mass phenotype owing to *Lrp5-T253I* mutation have low plasma levels of serotonin. *J. Bone Miner. Res.* **25,** 673–675 (2010).
84. Frost, M. *et al.* Levels of serotonin, sclerostin, bone turnover markers as well as bone density and microarchitecture in patients with high bone mass phenotype due to a mutation in Lrp5. *J. Bone Miner. Res.* **26,** 1721–1728 (2011).
85. Vilaca, T., Yamamoto, R. M., Carvalho, A. B. & Lazaretti-Castro, M. Neuroendocrine tumor associated with severe osteoporosis in a male patient. *Endocr Rev,* **32,** abstr. P3-123 (2011)
86. Modder, U. I. *et al.* Relation of serum serotonin levels to bone density and structural parameters in women. *J. Bone Miner. Res.* **25,** 415–422 (2010).
87. Shi, Z. C. *et al.* Modulation of peripheral serotonin levels by novel tryptophan hydroxylase inhibitors for the potential treatment of functional gastrointestinal disorders. *J. Med. Chem.* **51,** 3684–3687 (2008).
88. Liu, Q. *et al.* Discovery and characterization of novel tryptophan hydroxylase inhibitors that selectively inhibit serotonin synthesis in the gastrointestinal tract. *J. Pharmacol. Exp. Ther.* **325,** 47–55 (2008).
89. Inose, H. *et al.* Efficacy of serotonin inhibition in mouse models of bone loss. *J. Bone Miner. Res.* **26,** 2002–2011 (2011).
90. Yadav, V. K. *et al.* Pharmacological inhibition of gut-derived serotonin synthesis is a potential bone anabolic treatment for osteoporosis. *Nature Med.* **16,** 308–312 (2010).

Acknowledgements We thank T. Clemens, P. Ducy, M. Gershon, M. Kassem, S. Kousteni and members of the Karsenty laboratory for comments on the manuscript. We apologize to our colleagues whose work is not directly discussed and/or cited in this article because of space constraints. G.K. is supported by the National Institutes of Health, and M.F. by the Canadian Diabetes Association.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to G.K. (gk2172@columbia.edu).

REVIEW

doi:10.1038/nature10717

The Amazon basin in transition

Eric A. Davidson[1], Alessandro C. de Araújo[2,3], Paulo Artaxo[4], Jennifer K. Balch[1,5], I. Foster Brown[1,6], Mercedes M. C. Bustamante[7], Michael T. Coe[1], Ruth S. DeFries[8], Michael Keller[9,10], Marcos Longo[11], J. William Munger[11], Wilfrid Schroeder[12], Britaldo S. Soares-Filho[13], Carlos M. Souza Jr[14] & Steven C. Wofsy[11]

Agricultural expansion and climate variability have become important agents of disturbance in the Amazon basin. Recent studies have demonstrated considerable resilience of Amazonian forests to moderate annual drought, but they also show that interactions between deforestation, fire and drought potentially lead to losses of carbon storage and changes in regional precipitation patterns and river discharge. Although the basin-wide impacts of land use and drought may not yet surpass the magnitude of natural variability of hydrologic and biogeochemical cycles, there are some signs of a transition to a disturbance-dominated regime. These signs include changing energy and water cycles in the southern and eastern portions of the Amazon basin.

Humans have been part of the vast forest–river system of the Amazon basin for many thousands of years, but expansion and intensification of agriculture, logging and urban footprints during the past few decades have been unprecedented. The human population of the Brazilian Amazon region increased from 6 million in 1960 to 25 million in 2010, and the forest cover for this region has declined to about 80% of its original area[1]. Efforts to curb deforestation have led to a steep decline in forest clearing in the Brazilian Amazon, from nearly 28,000 $km^2\,yr^{-1}$ in 2004 to less than 7,000 $km^2\,yr^{-1}$ in 2011[1]. However, this progress remains fragile. The river system produces about 20% of the world's freshwater discharge[2], and the forest biomass holds about 100 billion tonnes of carbon (C; refs 3, 4), which is equivalent to more than 10 years' worth of global fossil-fuel emissions. Maintaining the biotic integrity of the biome and the ecosystem services it provides to local, regional and global communities will require improved understanding of the vulnerability and resilience of Amazonian ecosystems in the face of change.

Here we provide a framework for understanding the linkages between natural variability, drivers of change, responses and feedbacks in the Amazon basin (Fig. 1). Although the basin-wide carbon balance remains uncertain, evidence is emerging for a directional change from a possible sink towards a possible source. Where deforestation is widespread at local and regional scales, the dry season duration is lengthening and wet season discharge is increasing. We show that the forest is resilient to considerable natural climatic variation, but global and regional climate change forcings interact with land-use change, logging and fire in complex ways, generally leading to forest ecosystems that are increasingly vulnerable to degradation.

Natural and anthropogenic climatic variation

Changes in Amazonian ecosystems must be viewed in the context of the natural variation in climate[5,6] and soils[7] across the region, as well as natural cycles of climatic variation and extreme events. A climatic gradient spans the Amazon basin (Fig. 2), from the continuously rainy northwest to the wet/dry climate and long dry season of the southern and eastern regions, including the Cerrado (woodland/savannah) in the southeast. This climatic gradient is largely coincident with a gradient in land-use change, with more conversion to agriculture in the drier eastern and southern regions, indicating the interconnectedness of biophysical and socio-economic processes.

The El Niño/Southern Oscillation (ENSO) profoundly affects rainfall in the Amazon basin[5], especially the eastern portion; there is decreased flow of the Amazon River and some of its major tributaries during El Niño years, and increased flow and increased flooding during La Niña years[6]. The ENSO effect is superimposed over a 28-year cycle of variation in precipitation[5,6] such that the biggest floods occur when La Niña coincides with the wet phase in the 28-year cycle; this coincidence last occurred in the mid-1970s (Fig. 3). The worst droughts occur when El Niño coincides with the dry phase of the longer-term cycle, such as the 1992 drought. The North Atlantic Oscillation (NAO) also affects the region, contributing to, for example, the 2005 drought, which resulted in the lowest river levels recorded until then in southern and western tributaries[8]. Although much has been learned about extreme events and decadal-scale cycles, no discernable long-term trend has yet been identified in the total discharge of the Amazon River[9].

Forests are resistant to seasonal droughts

The ability of roots to access deep soil water[10] and to redistribute it[11] helps to maintain evergreen canopies during dry seasons, demonstrating the adaptation of Amazon forest species to seasonal drought. The combination of access to deep soil water and less cloudiness permits continued plant photosynthesis throughout most of the dry season[12]. However, transitional forests and Cerrado ecosystems, where mean annual precipitation is less than 1,700 mm and the dry season lasts for $\geq$4 months, show clear evidence of dry season declines in evapotranspiration and therefore potential water stress[13]. Many tree species in the Amazon and Cerrado produce a flush of new green leaves near the end of every dry season, which is often detected in satellite images as an

[1]The Woods Hole Research Center, 149 Woods Hole Road, Falmouth, Massachusetts 02540-1644, USA. [2]Embrapa Amazônia Oriental, Travessa Dr. Enéas Pinheiro, s/n, Marco, Caixa Postal 48, Belém, Pará 66095-100, Brazil. [3]Instituto Nacional de Pesquisas da Amazonia (INPA), Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), Avenida André Araújo, 2936, Manaus, Amazonas, 69060-001, Brazil. [4]Instituto de Física, Universidade de São Paulo, Rua do Matão, Travessa R, 187, São Paulo, SP 05508-090, Brazil. [5]National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300, Santa Barbara, California 93101, USA. [6]Universidade Federal do Acre, Mestrado em Ecologia e Manejo de Recursos Naturais, Parque Zoobotânico, Distrito Industrial, Rio Branco, AC 69915-900, Brazil. [7]Universidade de Brasília, Instituto de Ciências Biológicas, Departamento de Ecologia, Campus Universitário Darcy Ribeiro, Asa Norte, Brasília, Distrito Federal 70910-900, Brazil. [8]Columbia University, Department of Ecology, Evolution, and Environmental Biology, 1200 Amsterdam Avenue, New York, New York 10027, USA. [9]USDA Forest Service, International Institute of Tropical Forestry, Jardín Botánico Sur, 1201 Calle Ceiba, San Juan, Puerto Rico 00926-1119, USA. [10]Embrapa Monitoramento por Satélite, Avenida Soldado Passarinho, 303, Fazenda Chapadão, Campinas, São Paulo, Brazil. [11]Harvard University, School of Engineering and Applied Sciences, Department of Earth and Planetary Sciences, 20 Oxford Street, Cambridge, Massachusetts 02138, USA. [12]University of Maryland, Earth System Science Interdisciplinary Center, 5825 University Research Court Suite 4001, College Park, Maryland 20740, USA. [13]Universidade Federal de Minas Gerais, Centro de Sensoriamento Remoto, Avenida Antônio Carlos 6627, Belo Horizonte, Minas Gerais 31270-900, Brazil. [14]Imazon, Centro de Geotecnologia do Imazon, Rua Domingos Marreiros 2020, Belém, Pará 66060-160, Brazil.

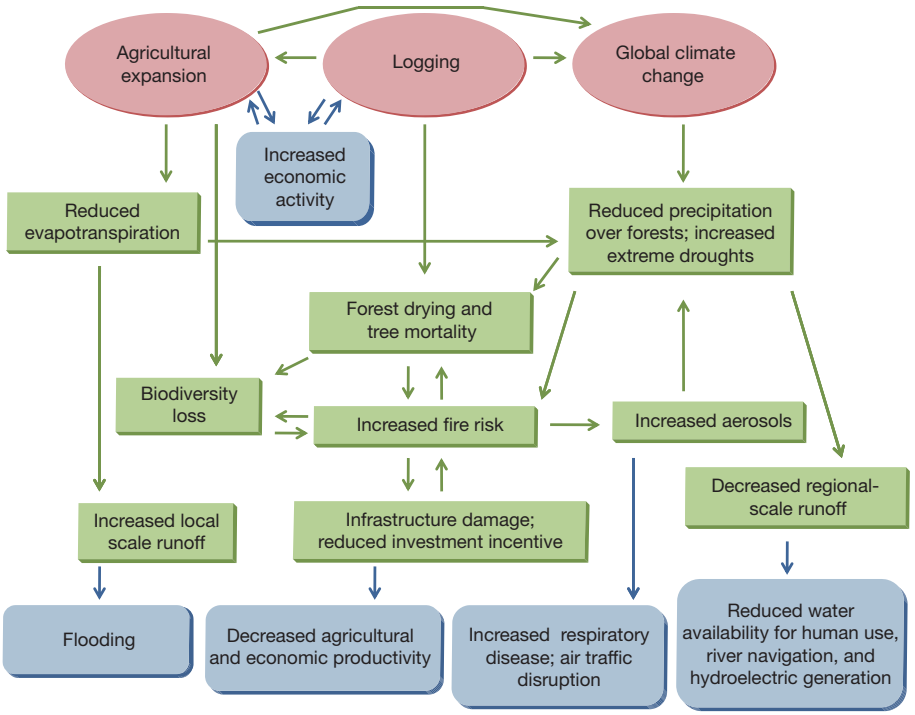


Figure 1 | Interactions between global climate, land use, fire, hydrology, ecology and human dimensions. Forcing factors are indicated with red ovals; processes addressed in this Review are indicated by green boxes and arrows; and consequences for human society are indicated by blue boxes with rounded corners.

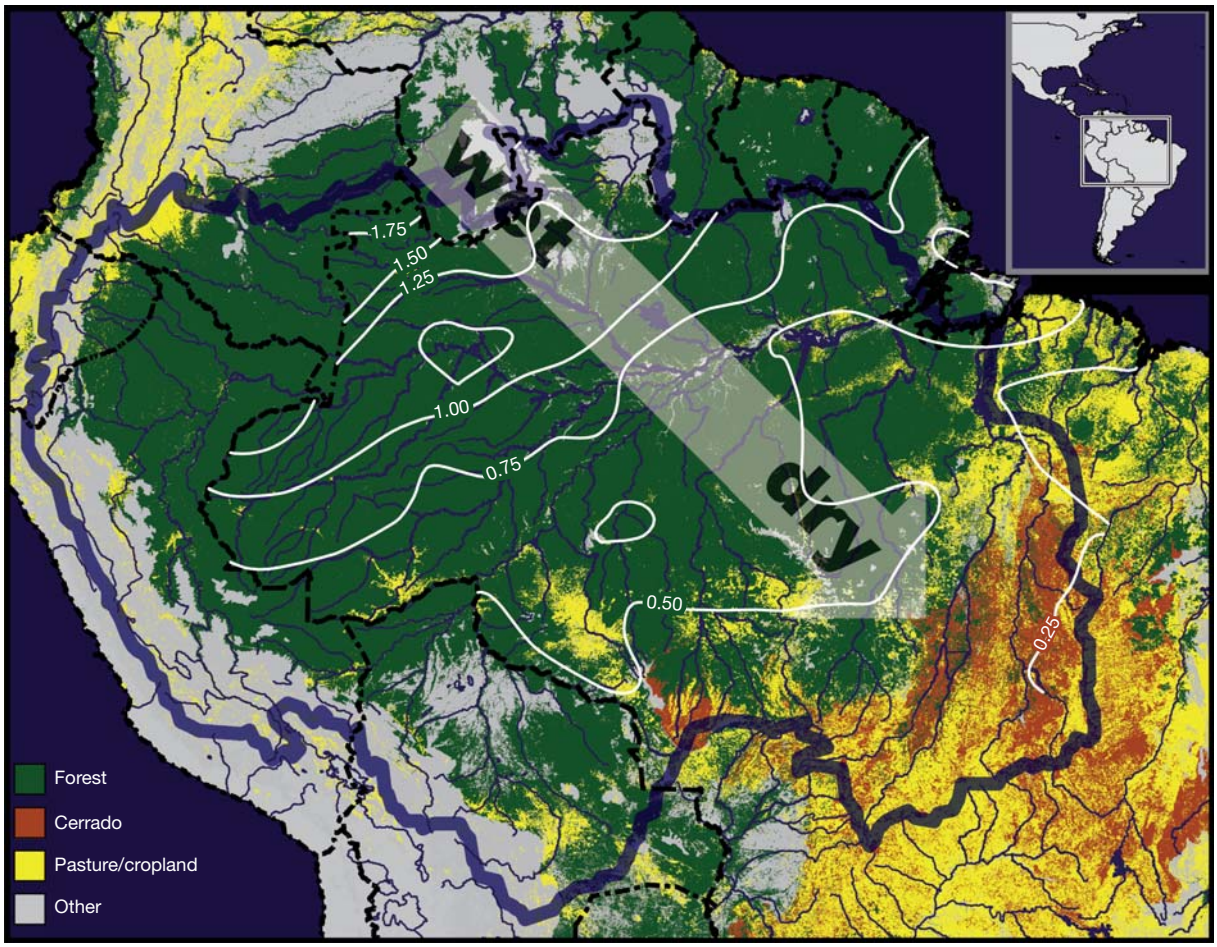


Figure 2 | Climatic gradient across the Amazon basin. Main figure, the hydrologic Amazon basin is demarcated by a thick blue line; isopleths of mean daily precipitation during the three driest months of the year[97] (in mm; white lines) are overlain onto four land-cover classes[98,99] (key at bottom left). These isopleths are presented only for areas within Brazil, because of lack of adequate data elsewhere. The arrow emphasizes the trend from continuously wet conditions in the northwest to long and pronounced dry seasons in the southeast, which includes Cerrado (savannah/woodland) vegetation. National boundaries are demarcated by broken black lines. Inset, map showing area of main figure (boxed).

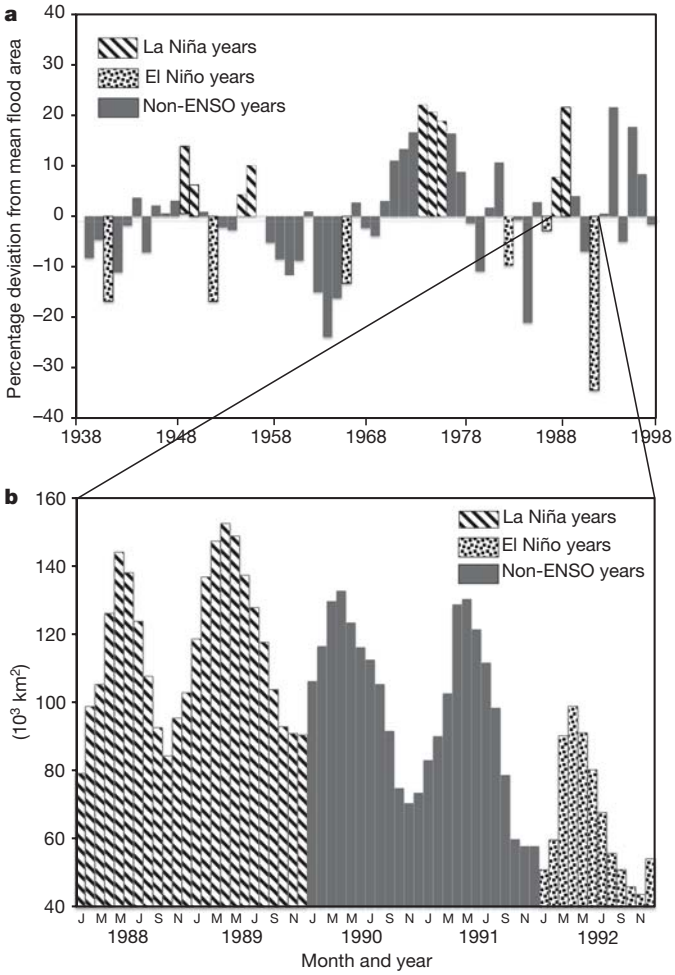


Figure 3 | Decadal and seasonal variation in flood area. The long-term record of simulated interannual variation in percentage deviation from mean flood area (**a**) illustrates how the ENSO events are superimposed over a 28-year cycle of high- and low-precipitation phases[100]. Recurrent seasonal patterns of flooded area (10^3 km^2) for five selected years (**b**) are superimposed over the interannual variation. Striped bars, La Niña years; stippled bars, El Niño years; solid bars, non-ENSO years.

increase in vegetation indices that involve ratios of red and near-infrared reflectance[14,15]. The relation between these satellite-based indices of seasonal greenness and ecosystem productivity remains an unresolved focus of debate in these studies, but in any case, this response represents a short-term phenomenon.

Multi-year or extreme drought

Experimental manipulations and observations of permanent forest plots address responses to multi-year and extreme drought. Two long-term drought experiments have produced remarkably similar results, demonstrating that adaptation to seasonal drought can be overwhelmed by multi-year drought[16,17]. These studies demonstrated a physiological adaptation of the trees, which maintained a relatively constant water tension in the xylem (isohydry) in both wet and dry seasons; but this adaptation may eventually lead to mortality when roots are unable to extract enough soil water during multi-year droughts[18]. After diverting 35–50% of total rainfall for three years using below-canopy panels and gutters, plant-available soil moisture stores became depleted, wood production declined by about 30–60%, tree mortality nearly doubled, and live above-ground biomass decreased by about 18–25% (refs 16, 17). Mortality rates increased to nearly three times that in the control plot during years 4–7 of rainfall exclusion[17].

The severe 2005 drought in the southwestern Amazon—when dry season temperatures were 3–5 °C warmer than normal and rainfall over the Solimões River basin was only 33–65% of average values[8]—may have exceeded the adaptive capacity of many forest species. Analysis of 51 long-term monitoring plots across Amazonia showed that, relative to pre-2005 conditions, most forest plots subjected to increased water deficit in 2005 lost several tons of living tree biomass carbon per hectare, owing to a marginally non-significant decline in growth and a significant increase in mortality of trees[19]. A similarly severe but more extensive drought occurred again in 2010, affecting more than half of the basin and resulting in the lowest discharge ever recorded at Manaus[20,21]. Susceptibility to drought is likely to vary regionally, depending on the climate (total precipitation and its seasonal distribution) and soil water storage properties (texture and depth) to which the existing vegetation types (for example, Cerrado woodlands, tall-statured central Amazon forests, and transition forests) are physiologically adapted. Furthermore, there is evidence that certain taxa are more vulnerable to drought-induced mortality[17,22]. Despite this regional variability, the observations of natural droughts and the drought manipulation experiments indicate similar trends of mortality in response to dry season intensity[23].

Land-use change and regional climate

Land use is changed to capture agricultural and forestry revenues, and results in trade-offs with multiple ecosystem services, such as C storage, climate regulation, hydrologic balance and biodiversity (Fig. 1).

The drivers of deforestation

Road paving is one of the economic activities that stimulates deforestation[24]. Further clearing occurs along networks of 'unofficial roads' that result from the interacting interests of colonist farmers and loggers[25]; loggers minimize their costs by buying the right to log private lands. Although practices vary widely across the region, most small land holders (<200 ha) have kept more than 50% of their land in some combination of mature and secondary forest[26].

International and national demands for cattle and livestock feed are increasingly driving land-use change. Direct conversion of forest to cropland in 2003, mostly by large land holders, represented 23% of the deforestation in forest and Cerrado regions of the state of Mato Grosso[27]. Although cattle pasture remains the dominant use of cleared land, the growing importance of larger and faster conversion to cropland, mostly for soybean export, has defined a trend of forest loss in Amazonia since the early 2000s.

Although selective logging is not an immediate land-use change, it often leads to deforestation. From 1999 to 2003, the area annually logged in the Amazon basin was similar in magnitude to the area deforested[28]. Logged areas are accessible by logging roads and are likely to be cleared within only a few years after initial disturbance[29], and those that are not cleared have a high risk of burning[30]. On the other hand, reduced-impact logging has been demonstrated to be economically viable, while causing only modest and transient effects on carbon storage and water exchange[31]. Expansion of protected areas has also played an important role in reducing deforestation in the Brazilian Amazon (Fig. 4)[32].

Deforestation alters the energy balance

Incoming air from the Atlantic Ocean provides about two-thirds of the moisture that forms precipitation over the Amazon basin[33]. The remainder is supplied through recycling of evapotranspiration, primarily driven by the deep-rooted Amazon trees.

A large number of observational and modelling studies have suggested that deforestation causes two main changes in the energy and water balance of the Amazon basin, as follows. First, partitioning of the net radiation that is absorbed by the land surface changes, with a decrease in the latent heat flux and an increase in the sensible heat flux, primarily because deforestation results in less vegetation being available to transpire water to the atmosphere. Second, replacing the dark rainforest with more reflective pasturelands or crops results in a decrease in solar radiation absorbed by the land surface. Reforestation can reverse these

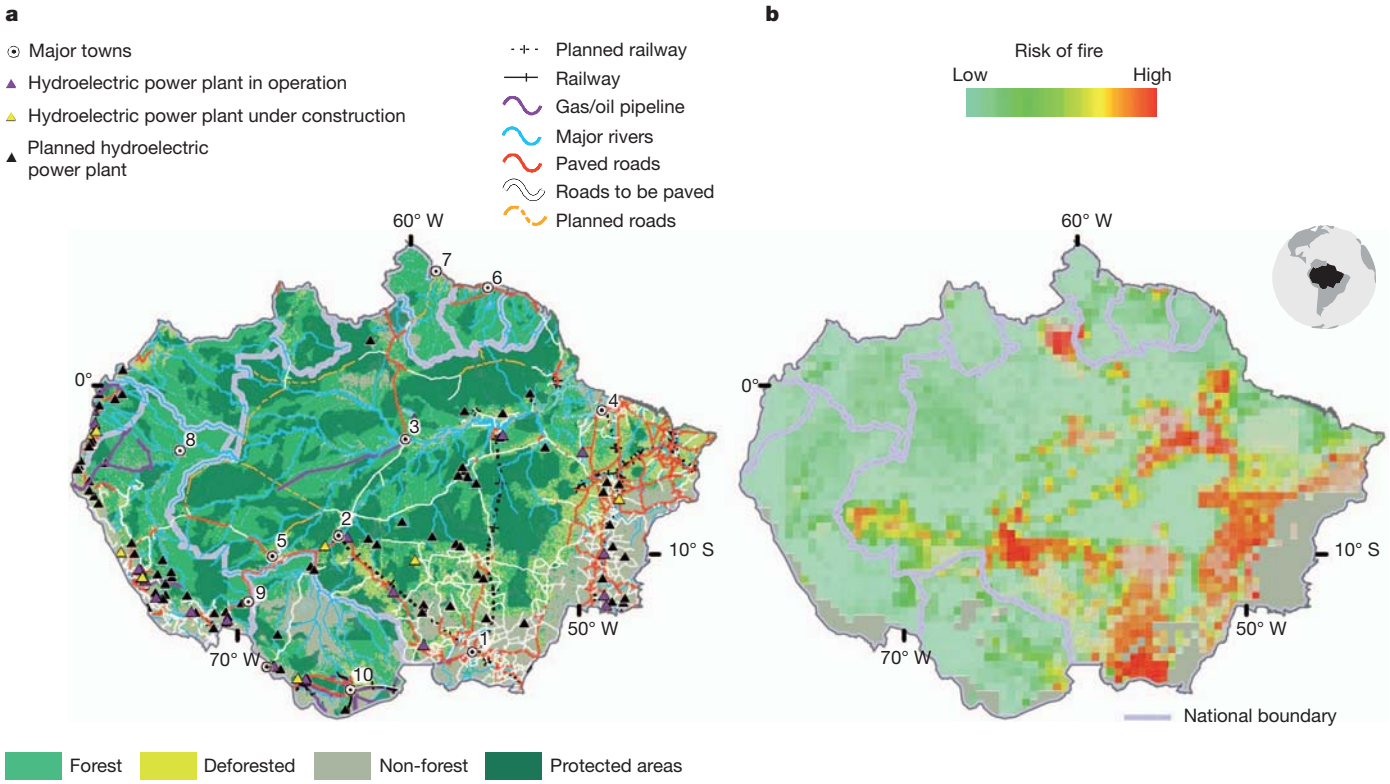


Figure 4 | The Amazon basin today and future fire risks. **a**, Protected areas and major planned infrastructure. **b**, The risk of fire by 2050[96] under business-as-usual deforestation and climate change scenarios[24]. The numbers in **a** indicate the following cities: 1, Cuiabá; 2, Porto Velho; 3, Manaus; 4, Belém; 5, Rio Branco; 6, Paramaribo; 7, Georgetown; 8, Iquitos; 9, Puerto Maldonado; 10, Santa Cruz de la Sierra.

trends. Within a few years of pasture abandonment, regrowing Amazonian forests establish rates of evapotranspiration and reflectivity that are close to those measured in mature forests, even though they have not yet recovered the biomass and species diversity of a mature forest[34,35].

Atmospheric convection and precipitation are driven by the fluxes of energy and water from the land surface. Where clearings for cattle pastures extend tens of kilometres outward from a road, the air above the deforested areas warms up more quickly and tends to rise and draw moist air from the surrounding forest, creating so-called 'vegetation breezes'. This decreases rainfall over the forest while increasing cloudiness, rainfall and thunderstorms over the pasture[36]. Heterogeneous deforestation at large scales (hundreds to thousands of km^2) leads to more complex circulation changes, with suppressed rainfall over core clearings, particularly at the beginning and the end of the wet season, and unchanged or increased rainfall over large remnant forest patches[37,38]. These changes also affect water and light availability, and the C uptake of the remaining forests, but those effects are not yet well quantified.

At deforestation scales greater than 10^5 km^2, numerical models consistently suggest that a significant decrease in basin-wide precipitation will occur[39] due to: (1) a decrease in the evapotranspiration from deforested regions and resultant downwind transport of water vapour; and (2) a decrease in net absorbed solar energy and a consequent general weakening of the continental-scale low-pressure system that drives precipitation over the basin.

Deforestation, climate and river discharge

Taken alone, a decrease in regional precipitation would result in decreased discharge. However, the integrated response of a river system depends on the balance between precipitation and evapotranspiration effects (Fig. 1). Deforestation within a particular watershed would cause reduced evapotranspiration and increased discharge, but deforestation at the continental scale could cause reduced regional precipitation and a tendency towards decreased river discharge[39].

A large disturbance and a long data record are needed to detect unambiguously the effect of deforestation on the discharge from a large river, given the large interannual and decadal variation in precipitation. For most of the major tributaries of the Amazon River, the area deforested is not yet large enough to be able to attribute changes in discharge specifically to deforestation. Similarly, a temporal trend in sediment load could not be distinguished from highly variable interannual and seasonal variation for the Madeira River, which drains the southwestern Amazon basin[40]. However, for the Tocantins River[41] and Araguaia River[42] systems, which drain parts of the Cerrado and rainforest environments in the southeastern Amazon, the relative contributions of climate variability and deforestation have been teased apart. From 1955 to 1995, the area of pasture and cropland in the Tocantins basin increased from about 30% to 50% and annual river discharge increased by about 25%, but changes in precipitation were not statistically significant. Changes of the same magnitude have occurred in the Araguaia River since the 1970s, and sediment load increased by 28% with deforestation. In both rivers, discharge increased mostly during the wet season, when flooding risks are greatest. If deforestation approaches this magnitude in other tributaries, it is likely that land-use change will enhance flooding and sediment transport.

Regional climate change

The IPCC fourth assessment climate change model runs show the highest probability of significant precipitation decrease predicted for southeastern Amazonia, where deforestation is greatest and where the climate and ecosystems transition from short-dry-season rainforest to long-dry-season savannah ecosystems[43,44]. Various global and regional climate modelling approaches have suggested that once deforestation exceeds about 40% of the entire Amazon basin, a 'tipping point' might be passed[45], whereby decreased energy and moisture released to the atmosphere from the largely deforested landscape would result in reduced convection and precipitation, and a shift in the forest–savannah boundary or large-scale dieback of rainforest.

A number of deficiencies in the structure and application of global climate models suggest that the uncertainty of these simulated outcomes and estimated tipping points is very high. A few examples include: (1) many of the models simulate too little precipitation in the Amazon owing to incomplete representation of the role of the Andes in continental circulation and large-scale convection over the core of the western Amazon, and also owing to coarse representation of the land surface with respect to small-scale meteorological processes[43]; (2) inter-annual variation of sea surface temperatures in the tropical Pacific and tropical north Atlantic Ocean are closely associated with extreme flood and drought events in the Amazon[5,8,20], but these teleconnections to the Amazon are not yet adequately represented in global climate models; and (3) the biophysical response of vegetation to increasing atmospheric CO_2, including effects on evapotranspiration, may be one of the largest unknowns for the future of the Amazon forests. The probability of simulated forest dieback due to decreased rainfall is greatly reduced when a strong CO_2 fertilization response is included in a vegetation model[44], but the scale of the actual impact of increasing CO_2 on photosynthetic efficiency remains a large source of uncertainty.

In summary, the changes in precipitation and discharge associated with deforestation already observed in the southern and eastern Amazon demonstrate a potential for significant vegetation shifts and further feedbacks to climate and discharge. Numerical models strongly suggest that potential future deforestation may also cause feedbacks to large-scale climate and vegetation distribution, but the models have deficiencies that prevent confident prediction of the magnitude or spatial distribution of deforestation that would lead to a significant region-wide decrease in precipitation—including whether a threshold, or tipping point, exists whereby the basin could slip into a dry, stable state. Focusing on a theoretical and difficult-to-define tipping point for the entire basin may divert the scientific community from the important large-scale regional changes that are already taking place, such as lengthening of the dry season[37,38] and increases in river discharge[41,42] in ecologically and agriculturally important transition zones of the eastern and southern flanks of the basin.

Fire as cause and consequence of change

The probability of fire is clearly affected by climate and land use, the latter providing the majority of ignition sources today[46]. Fire also affects regional climate through a complex set of biophysical and socio-economic feedback processes (Fig. 1).

Smoke changes cloud physics and rainfall

During the wet season, the air over most of the Amazon region is as pristine as air over the open ocean—only a few hundred aerosol particles per cm^3 of air[47]—inspiring the term 'green ocean'[48]. In stark contrast, burning for land clearing, pasture management and charcoal production, and escaped forest fires during the dry season, increase aerosols to more than 40,000 particles per cm^3 of air in some regions[47]. This smoke and haze affects the microphysical processes within clouds that determine how droplets are formed, making droplets too small to precipitate as rain, thus reducing local rainfall and increasing cloud lifetime[49]. The water vapour remaining in the atmosphere ascends to higher altitudes, where it invigorates thunderstorm formation and lightning strikes, but not necessarily rain. During the dry season, satellite-based measurements of aerosol optical depth were inversely correlated with precipitation[50]. In addition to locally smoke-inhibited rainfall, fires cause further plant stress due to ozone pollution[51] and thick haze that reduces light availability and photosynthesis[52]. Generally, plants are most productive with some scattered light at intermediate levels of aerosol thickness, but conditions during the biomass burning season often exceed this optimum[52].

Drought increases fire susceptibility

The tall, dense tree canopy of central Amazonian forests creates a humid microclimate at ground level, which naturally protects the forest from fire[53]. However, several lines of evidence indicate that this natural resistance may be changing: (1) about 39,000 km^2 of Amazon forest burned during the El Niño drought of 1998[54], including intact, closed-canopy forests; (2) both logging and drought-induced tree mortality allow sunlight to penetrate clearings in the canopy, which dries out the forest floor, rendering it more flammable[53]; (3) after a forest is burned once, it is more likely to burn again, because a burned forest dries out more easily[30]; and (4) ignition sources have also increased owing to pasture management and charcoal making[46]. Although Brazil has made great strides in recent years to reduce rates of deforestation[1], the frequency of fire has not decreased[55], and prospects for continued forest degradation resulting from fires escaping nearby agricultural areas may be a growing risk in many regions.

Fires alter forest characteristics

After fires sweep through Amazonian forests, tree mortality ranges from 8% to 64% of mature stems (≥10 cm diameter at breast height)[56]. More frequent and/or more severe fires tend to increase tree and liana mortality[57]. Big trees are generally better adapted to surviving fire, but tend to be the first to suffer from drought[19,22,57]. Although surviving stems can benefit from the initial pulse of fire-released nutrients and reduced competition, fire-induced mortality reduces overall canopy cover, biomass, and species richness[57,58]. The decline in plant species diversity also reduces the abundance of fruits and invertebrates, thereby changing the food supply of birds and other animals[56]. Frequent fire could change the structure, composition and functioning of vegetation by selecting fire-adapted species and favouring more flammable species (for example, grasses), thus leading to a more savannah-like ecosystem[59].

Multiple fires retard forest regrowth

Fire is used as a tool to help clear land for cattle pasture and to slow the invasion of woody shrubs, but pastures are often abandoned after a few years, when grass productivity declines and weeds can no longer be effectively controlled. Despite tremendous diversity in rates of regrowth among secondary forests from different regions of the basin, the rate of secondary forest regrowth following pasture abandonment was found to be negatively correlated with the number of fires that occurred while in the pasture phase[60]. Nitrogen (N) loss during burning alters the natural patterns of phosphorus limitation on highly weathered soils. In a study of secondary forests growing on abandoned pastures and croplands, several soil and foliage indicators of N limitation were strongest in the youngest forest stands and became less pronounced as the forests aged[61]. After decades of forest regrowth, the N cycle gradually recuperates, establishing a N-rich mature forest, but the rate of recuperation, as well as the rate of forest regrowth, depends, in part, on the legacy of previous land use and fire.

Disturbance effects on greenhouse gases

Changes in greenhouse-gas emissions due to disturbance processes must be placed in the context of natural emissions. Amazonian forests and wetlands are significant natural sources of methane[62–64] and nitrous oxide[65,66] (Fig. 5). Unfortunately, a net carbon balance for the region remains elusive.

Mature forests may be accumulating carbon

Repeated sampling of about 100 permanent plots in the RAINFOR network scattered across nearly all Amazonian countries indicates that mature Amazonian forests have been accumulating carbon at an estimated rate of 0.4 Pg C yr^{-1} (1 Pg = 10^{15} g; 95% confidence interval range of estimate, 0.29–0.57 Pg C yr^{-1}) in the decades before the 2005 drought[19]. The fastest growing trees are in the foothills of the Andes, where the soils are generally younger and more fertile[7], but where the trees are generally smaller and shorter-lived[3,4]. In contrast, the biggest and slowest growing trees occur in the oldest and more nutrient-poor soils of the lowland central and eastern parts of the basin[67]. The soils of mature forests on highly weathered Oxisols and Ultisols are unlikely to

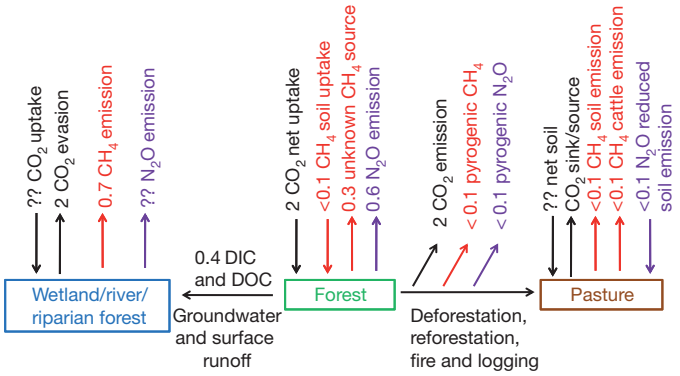


Figure 5 | Estimates of Amazonian greenhouse-gas emissions. Estimates of annual, basin-wide greenhouse-gas fluxes described in the text are presented together here, in a common currency of Pg CO_2-equivalents, using 100-year global warming potentials for CO_2 (black), CH_4 (red) and N_2O (purple). Owing to large uncertainties, all values are rounded to one significant figure, and even these estimates remain subject to debate. Where no estimate is available, "??" is indicated. Note that dissolved inorganic carbon (DIC) and dissolved organic carbon (DOC) may be transported via groundwater and overland flow from upland forests to streamside (riparian) forests, and that CO_2 can be lost (evasion) from river water to the atmosphere.

be significant C sinks[68], although more study is needed on a wider diversity of soils.

The cause of observed biomass increases in mature Amazonian forests remains unknown. Plausible explanations include a rebound from previous human or natural disturbances[69] or a change in resources that limit plant productivity, such as atmospheric CO_2, soil nutrients, or light availability due to changes in radiation, climate and cloudiness[70]. The RAINFOR network is our best indicator of Amazonian above-ground biomass change, but the network is neither a systematic nor a randomized sampling of Amazonian vegetation. Because the network has relatively few small plots covering a vast region, the effects of large-scale natural disturbances over decadal and longer timescales may not be included in the sampling network, leading to an overestimate of a biomass increase[69,71]. Although this challenge to the RAINFOR conclusions has been rebutted[72,73], resolving the issue will require empirical data on the distribution of natural disturbances, which is still poorly known[71]. A recent analysis of satellite images and meteorological data showed that large disturbances (>5 ha) caused by windstorms are rare, with a return interval of about 40,000 years (ref. 74), suggesting that such disturbance effects may not be common enough to undermine extrapolations of carbon uptake rates from the RAINFOR network. However, more work on this topic is needed, including better estimates of the return intervals of smaller disturbances (<5 ha)[71].

Full C accounting should also include exports from forests to aquatic systems. The river water is supersaturated with dissolved CO_2, which is eventually released to the atmosphere at an estimated rate of about 0.5 Pg C yr^{-1} (ref. 75). Estimates of the sources of this C remain poorly constrained—about two-thirds may come from leaf and wood detritus dropped into the river from flooded forests, with about one-third produced by aquatic plants (mats of grasses and other macrophytes) within the river, and a small fraction by algae[76]. Additional possible sources include particulates washed in with soil particles and dissolved organic and inorganic C in ground water[77,78]. We know very little about the C budget of flooded forests and riparian zone forests, which probably contribute significant terrestrially fixed C to streams and rivers[76].

Estimates of CO_2 fluxes based on year-round vertical profiling of atmospheric CO_2 concentrations by aircraft are available now only for part of the eastern Amazon. Fire emissions roughly cancel a modest biological sink during the dry season, so that a wet season source yields an annual net source of C to the atmosphere[79]. This result is consistent with ground-based estimates of slow growing trees and a concentration of land-use change in the eastern part of the basin. It remains to be seen if future aircraft measurements will corroborate the C sink inferred from scattered ground measurements in the more intact forests of the western part of the basin.

Disturbing forests causes net C and N loss

The net effect of Amazonian deforestation and reforestation results in an annual net C source of 0.15–0.35 Pg C (ref. 80). Adding C emissions from fire and logging extends the range to an annual net release of 0.2–0.8 Pg C (ref. 80). The estimated mean annual C emission from deforestation and burning of Cerrado is 0.07 Pg C for 2003 to 2008[81]. These estimates are improving, in part because of advances in the technology for analysing satellite images[82] to combine spatially explicit deforestation rates with regionally specific estimates of forest-C stocks[83].

Pyrogenic CH_4 emissions from conversion of Amazonian and Cerrado native vegetation to pasture are about 1.0 and 0.4 Tg CH_4 yr^{-1}, respectively[81], but this does not include shifting cultivation or wildfire. Annual pyrogenic N_2O emissions from conversion of Amazonian forest to pasture are about 0.01 Tg N as N_2O (ref. 81), but this does not include shifting cultivation or wildfire.

When forests are replaced by cattle pastures, they can either gain[84] or lose[85] soil C. Losses are more common where soil C stocks are initially large, and gains are more common when management inputs (fertilizer, herd rotation, overgrazing avoidance) are greatest[86]. However, changes in soil C stocks are usually dwarfed by much larger losses in tree biomass. In contrast, the sparse and short-statured trees of the Cerrado have less above-ground biomass than an Amazonian forest, but the C stocks in roots and soil organic matter of the Cerrado (100 Mg C ha^{-1} in the top 1 m of soil) can be 2–7 times higher than the above-ground stocks[87]. Well-managed cultivated pastures may provide enough C inputs to maintain soil C[88], but most pastures in the Cerrado region are in advanced stages of degradation, where C inputs are too low to sustain high soil C storage.

Amazonian upland forest soils annually take up about 1–3 Tg of CH_4, and pasture soils are probably a small net annual source of <0.1 Tg CH_4 (ref. 65). A significant net emission of CH_4 in upland forests has been measured, which might include termites or anaerobic respiration in waterlogged wood, soil, bromeliads, or moss patches, but the source remains unknown[64]. Enteric fermentation by cattle is estimated to emit 2.6 and 4.1 Tg CH_4 yr^{-1} in Amazonian and Cerrado regions, respectively[81]. Continuing studies point to major hydroelectric reservoirs as an increasing source of methane[76]. On the basis of chamber flux measurements, upland Amazonian forest soils are estimated to emit 1.3 Tg yr^{-1} of N_2O-N (ref. 65), which is about 15% of global non-anthropogenic emissions. Young cattle pastures have higher N_2O emissions compared to forests, but old pastures have lower emissions, so the net effect of deforestation has been a small annual decrease of <0.1 Tg N_2O-N (ref. 65).

Although secondary forests may be significant carbon sinks in other parts of the world[89], they currently contribute little to the net C balance of the Amazon basin, because they are frequently re-cut before they grow large enough to store much C[90]. Indeed, the area of secondary forests is declining where agriculture continues to expand and intensify, leading to continued loss of biomass-C from those regions[91]. Agroforestry and other alternatives to slash-and-burn agriculture for smallholders have not been widely adopted, but the potential for significant C sequestration per hectare and the techniques of nutrient management in these systems have been demonstrated[92,93].

Emerging evidence for a transition

Are impacts of land-use and climate change in the Amazon basin surpassing the natural variability of climate, greenhouse-gas emissions, and cycles of carbon, nitrogen and water? Thanks to increased research in this area, including the Large-scale Biosphere-Atmosphere (LBA) experiment in Amazonia (see the accompanying World View in this issue for a description of the LBA project), we can answer this question for some, but not all consequences of land-use and climate change. For greenhouse gases, the answer is probably 'not yet' with respect to CH_4

and N_2O, because they remain dominated by large emissions from undisturbed wetlands and soils, respectively, but the answer for CO_2 is more complex (Fig. 5). Although a C budget for the basin remains uncertain, deforestation has moved the net basin-wide budget away from a possible late-twentieth-century net C sink and towards a net source. This directional change is consistent with recent results of inverse modelling based on the TransCom3 network of CO_2 measurements, which reports a shift from a sink in the 1980s to a source in the 2000s for the tropical Americas[94]. Much of the Amazon forest is resilient to seasonal and moderate drought, but this resilience can and has been exceeded with experimental and natural severe droughts, indicating a risk of C loss if drought increases with climate change. The forest is also resilient to initial disturbances, but repeated or prolonged disturbance changes forest structure and nutrient dynamics, potentially leading to a long-term change in vegetation composition and C loss. A combination of regional net flux estimates based on aircraft campaign measurements with ground-based studies that elucidate process-level understanding is needed to narrow uncertainties.

With respect to energy and water cycles, at least two of the large river basins on the southeastern flanks of the Amazon forest that also drain the more heavily deforested Cerrado region—the Tocantins and Araguaia basins—have experienced increases in wet season discharge and sediment load. Evidence for changes in temporal and spatial patterns of precipitation, such as extended length of the dry season, is emerging at local and regional scales. We cannot yet answer the questions of whether total precipitation has changed or whether recent severe droughts and other extreme events are clear indicators of patterns expected to persist. Narrowing uncertainties about the effects of deforestation on regional precipitation, temperature and fire risk will require combining realistic spatial patterns of deforestation and degradation with improved mesoscale circulation models of climate.

The emerging evidence of a system in biophysical transition highlights the need for improved understanding of the trade-offs between land cover, carbon stocks, water resources, habitat conservation, human health and economic development in future scenarios of climate and land-use change[24,32,95,96]. Brazil is poised to become one of the few countries to achieve the transition to a major economic power without destroying most of its forests. However, continued improvements in scientific and technological capacity and human resources will be required in the Amazon region to guide and manage both biophysical and socio-economic transitions.

1. Instituto Nacional de Pesquisas Espaciais & National Institute for Space Research *Projeto Prodes Monitoramento da Florsta Amazonica Brasileira por Satélite Prodes* ⟨http://www.obt.inpe.br/prodes/⟩ (2011).
2. Salati, E. & Vose, R. Amazon basin: a system in equilibrium. *Science* **225,** 129–138 (1984).
   One of the first presentations of the Amazon basin from a systems perspective.
3. Malhi, Y. *et al.* The regional variation of aboveground live biomass in old-growth Amazonian forests. *Glob. Change Biol.* **12,** 1107–1138 (2006).
4. Saatchi, S. S., Houghton, R. A., Dos Santos Alvara, R. C., Soares, J. V. & Yu, Y. Distribution of aboveground live biomass in the Amazon basin. *Glob. Change Biol.* **13,** 816–837 (2007).
   Estimates of regional variation and patterns in forest biomass are presented based on a remote sensing approach.
5. Marengo, J. A. Interdecadal variability and trends of rainfall across the Amazon basin. *Theor. Appl. Climatol.* **78,** 79–96 (2004).
6. Coe, M. T., Costa, M. H., Botta, A. & Birkett, C. Long-term simulations of discharge and floods in the Amazon basin. *J. Geophys. Res.* **107,** 8044, http://dx.doi.org/10.1029/2001JD000740 (2002).
7. Quesada, C. A. *et al.* Variations in chemical and physical properties of Amazon forest soils in relation to their genesis. *Biogeosciences* **7,** 1515–1541 (2010).
8. Marengo, J. A., Nobre, C. A., Tomasella, J., Cardosa, M. F. & Oyama, M. D. Hydro-climate and ecological behaviour of the drought of Amazonia in 2005. *Phil. Trans. R. Soc. B* **363,** 1773–1778 (2008).
9. Marengo, J. in *Tropical Rainforest Responses to Climatic Change* (eds Bush, M. B. & Flenley, J. R.) 236–268 (Springer Praxis Books, 2007).
10. Nepstad, D. C. *et al.* The role of deep roots in the hydrological and carbon cycles of Amazonian forests and pastures. *Nature* **372,** 666–669 (1994).
    First demonstration of the importance of deep rooting for survival of eastern Amazonian trees.
11. Oliveira, R. S. *et al.* Deep root function in soil water dynamics in cerrado savannas of central Brazil. *Funct. Ecol.* **19,** 574–581 (2005).
12. Saleska, S. R. *et al.* Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. *Science* **302,** 1554–1557 (2003).
13. da Rocha, H. R. *et al.* Patterns of water and heat flux across a biome gradient from tropical forest to savanna in Brazil. *J. Geophys. Res.* **114,** G00B12, http://dx.doi.org/10.1029/2007JG000640 (2009).
14. Brando, P., Goetz, S., Baccini, A., Nepstad, D. & Beck, P. Seasonal and interannual variability of climate and vegetation indices across the Amazon. *Proc. Natl Acad. Sci. USA* **107,** 14685–14690 (2010).
15. Huete, A. *et al.* Amazon rainforests green-up with sunlight in dry season. *Geophys. Res. Lett.* **33,** L06405, http://dx.doi.org/10.1029/02005GL025583 (2006).
16. Brando, P. M. *et al.* Drought effects on litterfall, wood production, and belowground carbon cycling in an Amazon forest: results of a throughfall reduction experiment. *Phil. Trans. R. Soc. B* **363,** 1839–1848 (2008).
17. da Costa, A. C. L. *et al.* Effect of 7 yr of experimental drought on vegetation dynamics and biomass storage of an eastern Amazonian rainforest. *New Phytol.* **187,** 579–591 (2010).
18. Fisher, R. A., Williams, M., Lobo do Vale, R., Costa, A. & Meir, P. Evidence from Amazonian forests is consistent with isohydric control of leaf water potential. *Plant Cell Environ.* **29,** 151–165 (2006).
19. Phillips, O. L. *et al.* Drought sensitivity of the Amazon rainforest. *Science* **323,** 1344–1347 (2009).
20. Lewis, S. L., Brando, P. M., Phillips, O. L., van der Heijden, G. M. F. & Nepstad, D. C. The 2010 Amazon drought. *Science* **331,** 554 (2011).
21. Xu, L. *et al.* Widespread decline in greenness of Amazonian vegetation due to the 2010 drought. *Geophys. Res. Lett.* **38,** L07402, http://dx.doi.org/10.1029/2011GL046824 (2011).
22. Nepstad, D. C., Tohver, I. M., Ray, D., Moutinho, P. & Cardinot, G. Mortality of large trees and lianas following experimental drought in an Amazon forest. *Ecology* **88,** 2259–2269 (2007).
23. Phillips, O. L. *et al.* Drought–mortality relationships for tropical forests. *New Phytol.* **187,** 631–646 (2010).
24. Soares-Filho, B. S. *et al.* Modelling conservation in the Amazon basin. *Nature* **440,** 520–523 (2006).
    Landmark presentation of scenarios of development and conservation policies in a spatially explicit simulation model.
25. Arima, E. Y., Walker, R. T., Perz, S. G. & Caldas, M. M. Loggers and forest fragmentation: behavioral models of road building in the Amazon basin. *Ann. Assoc. Am. Geogr.* **95,** 525–541 (2005).
26. Brondízio, E. S. *et al.* in *Amazonia and Global Change* (eds Keller, M., Bustamante, M., Gash, J. & Dias, P. S.) 117–143 (American Geophysical Union, 2009).
27. Morton, D. C. *et al.* Cropland expansion changes deforestation dynamics in the southern Brazilian Amazon. *Proc. Natl Acad. Sci. USA* **103,** 14637–14641 (2006).
28. Asner, G. P. *et al.* Selective logging in the Brazilian Amazon. *Science* **310,** 480–482 (2005).
29. Asner, G. P. *et al.* Condition and fate of logged forests in the Brazilian Amazon. *Proc. Natl Acad. Sci. USA* **103,** 12947–12950 (2006).
30. Nepstad, D. C. *et al.* Road paving, fire regime feedbacks, and the future of Amazon forests. *For. Ecol. Mgmt* **154,** 395–407 (2001).
31. Miller, S. D. *et al.* Reduced impact logging minimally alters tropical rainforest carbon and energy exchange. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1105068108 **108,** 19431–19435 (2011).
32. Soares-Filho, B. S. *et al.* Role of the Brazilian Amazon protected areas in climate change mitigation. *Proc. Natl Acad. Sci. USA* **107,** 10821–10826 (2010).
33. Costa, M. H. & Foley, J. A. Trends in the hydrologic cycle of the Amazon basin. *J. Geophys. Res.* **104,** 14189–14198 (1999).
34. Hölscher, D., Sá, T. D. A., Bastos, T. X., Denich, M. & Fölster, H. Evaporation from young secondary vegetation in eastern Amazonia. *J. Hydrol.* **193,** 293–305 (1997).
35. Vieira, I. C. G. *et al.* Classifying successional forests using Landsat spectral properties and ecological characteristics in eastern Amazonia. *Remote Sens. Environ.* **87,** 470–481 (2003).
36. Avissar, R. & Schmidt, T. An evaluation of the scale at which ground-surface heat flux patchiness affects the convective boundary layer using a large-eddy simulation model. *J. Atmos. Sci.* **55,** 2666–2689 (1998).
37. Butt, N., Oliveira, P. A. & Costa, M. H. Evidence that deforestation affects the onset of the rainy season in Rondonia, Brazil. *J. Geophys. Res.* **116,** D11120, http://dx.doi.org/10.1029/2010JD015174 (2011).
38. Knox, R., Bisht, G., Wang, J. & Bras, R. L. Precipitation variability over the forest to non-forest transition in southwestern Amazonia. *J. Clim.* **24,** 2368–2377 (2011).
39. Coe, M. T., Costa, M. H. & Soares-Filho, B. S. The Influence of historical and potential future deforestation on the stream flow of the Amazon River — land surface processes and atmospheric feedbacks. *J. Hydrol.* **369,** 165–174 (2009).
40. Leite, N. K. *et al.* Intra and interannual variability in the Madeira River water chemistry and sediment load. *Biogeochemistry* **105,** 37–51 (2011).
41. Costa, M. H., Botta, A. & Cardille, J. A. Effects of large-scale changes in land cover on the discharge of the Tocantins River, Southeastern Amazonia. *J. Hydrol.* **283,** 206–217 (2003).
42. Coe, M. T., Latrubesse, E. M., Ferreira, M. E. & Amsler, M. L. The effects of deforestation and climate variability on the streamflow of the Araguaia River, Brazil. *Biogeochemistry* **105,** 119–131 (2011).
43. Malhi, Y. *et al.* Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest. *Proc. Natl Acad. Sci.* **106,** 20610–20615 (2009).
    A review of climate model predictions for the Amazon basin.
44. Rammig, A. *et al.* Estimating the risk of Amazonian forest dieback. *New Phytol.* **187,** 694–706 (2010).
45. Nobre, C. A. & Simone Borma, L. Tipping points' for the Amazon forest. *Curr. Opin. Environ. Sust.* **1,** 28–36 (2009).

46. Alencar, A., Solórzano, L. & Nepstad, D. Modeling forest understory fire in an eastern Amazonian landscape. *Ecol. Appl.* **14,** S139–S149 (2004).
47. Artaxo, P. *et al.* Physical and chemical properties of aerosols in the wet and dry season in Rondônia, Amazonia. *J. Geophys. Res.* **107** (D20), 8081–8095 (2002).
48. Williams, E. *et al.* Contrasting convective regimes over the Amazon: implications for cloud electrification. *J. Geophys. Res.* **107** (D20), 8082–8093 (2002).
49. Andreae, M. O. *et al.* Smoking rain clouds over the Amazon. *Science* **303,** 1337–1342 (2004).
A review of understanding of how smoke from biomass burning affects local and regional climate.
50. Bevan, S. L., North, P. R. J., Grey, W. M. F., Los, S. O. & Plummer, S. E. Impact of atmospheric aerosol from biomass burning on Amazon dry-season drought. *J. Geophys. Res.* **114,** D09204, http://dx.doi.org/10.1029/2008JD011112 (2009).
51. Longo, K. M. *et al.* Correlation between smoke and tropospheric ozone concentration in Cuiabá during SCAR-B. *J. Geophys. Res.* **104** (D10), 12113–12129 (1999).
52. Oliveira, P. H. F. *et al.* The effects of biomass burning aerosols and clouds on the CO_2 flux in Amazonia. *Tellus B* **59,** 338–349 (2007).
53. Ray, D., Nepstad, D. & Moutinho, P. Micrometeorological and canopy controls of flammability in mature and disturbed forests in an east-central Amazon landscape. *Ecol. Appl.* **15,** 1664–1678 (2005).
54. Alencar, A., Nepstad, D. C. & Vera Diaz, M. d. C. Forest understory fire in the Brazilian Amazon in ENSO and non-ENSO Years: area burned and committed carbon emissions. *Earth Interact.* **10,** 6, 1–17 (2006).
55. Aragão, L. E. O. & Shimabukuro, Y. E. The incidence of fire in Amazonian forests with implications for REDD. *Science* **328,** 1275–1278 (2010).
56. Barlow, J. & Peres, C. A. in *Emerging Threats to Tropical Forests* (eds Laurance, W. F. & Peres, C. A.) 225–240 (Univ. Chicago Press, 2006).
57. Balch, J. K. *et al.* Size, species, and fire characteristics predict tree and liana mortality from experimental burns in the Brazilian Amazon. *For. Ecol. Mgmt* **261,** 68–77 (2011).
58. Balch, J. D. *et al.* Negative fire feedback in a transitional forest of southeastern Amazonia. *Glob. Change Biol.* **14,** 2276–2287 (2008).
59. Nepstad, D. C., Stickler, C. M., Soares-Filho, B. & Merry, F. Interactions among Amazon land use, forests and climate: prospects for a near-term forest tipping point. *Phil. Trans. R. Soc. B* **363,** 1737–1746 (2008).
Explores the mechanisms of how land use, fire and climate change interact.
60. Zarin, D. J. *et al.* Legacy of fire slows carbon accumulation in Amazonian forest regrowth. *Front. Ecol. Environ.* **3,** 365–369 (2005).
61. Davidson, E. A. *et al.* Recuperation of nitrogen cycling in Amazonian forests following agricultural abandonment. *Nature* **447,** 995–998 (2007).
Chronosequences of secondary forests were analysed to demonstrate that nitrogen limitation occurs in young Amazonian forests and then gradually declines during secondary succession.
62. Melack, J. M. *et al.* Regionalization of methane emissions in the Amazon basin with microwave remote sensing. *Glob. Change Biol.* **10,** 530–544 (2004).
63. Miller, J. B. *et al.* Airborne measurements indicate large methane emissions from the eastern Amazon basin. *Geophys. Res. Lett.* **34,** L10809, http://dx.doi.org/10.1029/2006GL029213 (2007).
64. do Carmo, J. B., Keller, M., Dias, J. D., de Camargo, P. B. & Crill, P. A source of methane from upland forests in the Brazilian Amazon. *Geophys. Res. Lett.* **33,** 1–4 http://dx.doi.org/10.1029/2005GL025436 (2006).
65. Davidson, E. A. & Artaxo, P. Globally significant changes in biological processes of the Amazon Basin: results of the Large-scale Biosphere-Atmosphere Experiment. *Glob. Change Biol.* **10,** 519–529 (2004).
66. D'Amelio, M. T. S., Gatti, L. V., Miller, J. B. & Tans, P. Regional N_2O fluxes in Amazonia derived from aircraft vertical profiles. *Atmos. Chem. Phys.* **9,** 8785–8797 (2009).
67. ter Steege, H. N. *et al.* Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature* **443,** 444–447 (2006).
68. Telles, E. C. C. *et al.* Influence of soil texture on carbon dynamics and storage potential in tropical forest soils of Amazonia. *Glob. Biogeochem. Cycles* **17**, 1040, http://dx.doi.org/10.1029/2002GB001953 (2003).
69. Fisher, J. I., Hurtt, G. C., Thomas, R. Q. & Chambers, J. Q. Clustered disturbances lead to bias in large-scale estimates based on forest sample plots. *Ecol. Lett.* **11,** 554–563 (2008).
70. Nemani, R. R. *et al.* Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *Science* **300,** 1560–1563 (2003).
71. Chambers, J. Q. *et al.* Lack of intermediate-scale disturbance data prevents robust extrapolation of plot-level tree mortality rates for old-growth tropical forests. *Ecol. Lett.* **12,** E22–E25 (2009).
72. Gloor, M. *et al.* Does the disturbance hypothesis explain the biomass increase in basin-wide Amazon forest plot data? *Glob. Change Biol.* **15,** 2418–2430 (2009).
73. Lloyd, J., Gloor, E. U. & Lewis, S. L. Are the dynamics of tropical forests dominated by large and rare disturbance events? *Ecol. Lett.* **12,** E19–E21 (2009).
74. Espírito-Santo, F. D. B. *et al.* Storm intensity and old growth forest disturbances in the Amazon region. *Geophys. Res. Lett.* **37,** L11403, http://dx.doi.org/10.1029/2010GL043146 (2010).
75. Richey, J. E., Melack, J. M., Aufdenkampe, A. K., Ballester, V. M. & Hess, L. L. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. *Nature* **416,** 617–620 (2002).
Presents first calculations of potential loss of CO_2 to the atmosphere from the Amazon River and its main tributaries.
76. Melack, J. M., Novo, E. M. L. M., Forsberg, B. R., Piedade, M. T. F. & Maurice, L. in *Amazonia and Global Change* (eds Keller, M. *et al.*) 525–542 (American Geophysical Union Books, 2009).
77. Davidson, E. A., Figueiredo, R. O., Markewitz, D. & Aufdenkampe, A. Dissolved CO_2 in small catchment streams of eastern Amazonia: a minor pathway of terrestrial carbon loss. *J. Geophys. Res.* **115,** G04005, http://dx.doi.org/10.1029/2009JG001202 (2010).
78. Johnson, M. S. *et al.* CO_2 efflux from Amazonian headwater streams represents a significant fate for deep soil respiration. *Geophys. Res. Lett.* **35,** L17401, http://dx.doi.org/10.1029/2008GL034619 (2008).
79. Gatti, L. V. *et al.* Vertical profiles of CO_2 above eastern Amazonia suggest a net carbon flux to the atmosphere and balanced biosphere between 2000 and 2009. *Tellus B* http://dx.doi.org/10.1111/j.1600-0889.2010.00484.x (published online, 6 July 2010).
80. Houghton, R. A., Gloor, M., Lloyd, J. & Potter, C. in *Amazonia and Global Change* (eds Keller, M. *et al.*) 409–428 (American Geophysical Union Books, 2009).
The net effect of carbon loss due to deforestation and carbon accumulation from forest regrowth is estimated.
81. Bustamante, M. M. C. *et al.* Estimating greenhouse gas emissions from cattle raising in Brazil. *Clim. Change* (submitted).
82. Morton, D. C. *et al.* Rapid assessment of annual deforestation in the Brazilian Amazon using MODIS data. *Earth Interact.* **9,** 1–22 (2005).
83. Fearnside, P. M. *et al.* Biomass and greenhouse-gas emissions from land-use change in Brazil's Amazonian "arc of deforestation": the states of Mato Grosso and Rondonia. *For. Ecol. Mgmt* **258,** 1968–1978 (2009).
84. Cerri, C. E. P. *et al.* Modelling changes in soil organic matter in Amazon forest to pasture conversion, using the Century model. *Glob. Change Biol.* **10,** 815–832 (2004).
85. Asner, G. P., Townsend, A. R., Bustamante, M. M. C., Nardoto, G. B. & Olander, L. P. Pasture degradation in the Central Amazon: linking changes in carbon and nutrient cycling with remote sensing. *Glob. Change Biol.* **10,** 844–862 (2004).
86. Neill, C. & Davidson, E. A. in *Global Climate Change and Tropical Ecosystems* (eds Lal, R., Kimble, J. M. & Stewart, B. A.) 197–211 (CRC Press, 2000).
87. Grace, J., San Jose, J., Meir, P., Miranda, H. S. & Montes, R. A. Productive and carbon fluxes of tropical savannas. *J. Biogeogr.* **33,** 387–400 (2006).
88. Santos, A. J. B. *et al.* High rates of net ecosystem carbon assimilation by Brachiara pasture in the Brazilian cerrado. *Glob. Change Biol.* **10,** 877–885 (2004).
89. Pan, Y. *et al.* A large and persistent carbon sink in the world's forests. *Science* **333,** 988–993 (2011).
90. Neeff, T., Lucas, R. M., Santos, J. d., Brondizio, E. S. & Freitas, C. C. Area and age of secondary forests in Brazilian Amazonia 1978–2002: an empirical estimate. *Ecosystems* **9,** 609–623 (2006).
91. Almeida, A. S. d., Stone, T. A., Vieira, I. C. G. & Davidson, E. A. Non-frontier deforestation in the eastern Amazon. *Earth Interact.* **14,** 1–15 (2010).
92. Luizão, F., Fearnside, P. M., Cerri, C. E. P. & Lehmann, J. in *Amazonia and Global Change* (eds Keller, M., Bustamante, M., Gash, J. & Dias, P. S.) 311–336 (American Geophysical Union, 2009).
93. Davidson, E. A. *et al.* An integrated greenhouse gas assessment of an alternative to slash-and-burn agriculture in eastern Amazonia. *Glob. Change Biol.* **14,** 998–1007 (2008).
94. Gurney, K. R. & Eckels, W. J. Regional trends in terrestrial carbon exchange and their seasonal signatures. *Tellus B* **63,** 328–339 (2011).
95. da Silva, R. R., Werth, R. D. & Avissar, R. Regional impacts of future land-cover changes on the Amazon Basin wet-season climate. *J. Clim.* **21,** 1153–1170 (2008).
96. Silvestrini, R. A. *et al.* Simulating fire regimes in the Amazon in response to climate change and deforestation. *Ecol. Appl.* **21,** 1573–1590 (2011).
97. Nepstad, D. C. *et al.* Amazon drought and its implications for forest flammability and tree growth: a basin-wide analysis. *Glob. Change Biol.* **10,** 704–717 (2004).
98. Eva, H. D. *et al.* A land cover map of South America. *Glob. Change Biol.* **10,** 731–744 (2004).
99. Sano, E. E., Rosa, R., Brito, J. L. & Ferreira, L. G. *Mapeamento de Cobertura Vegetal do Bioma Cerrado: Estratégias e Resultados* (Embrapa Cerrados, Planaltina, District Federal, Brazil, 2007).
100. Coe, M. T., Costa, M. H. & Howard, E. A. Simulating the surface waters of the Amazon River Basin: impacts of new river geomorphic and dynamic flow parameterizations. *Hydrol. Process.* **21,** 2542–2553 (2007).

Acknowledgements We thank the Brazilian Ministry of Science and Technology (MCT), the National Institute for Space Research (INPE) and the National Institute of Amazonian Research (INPA) for designing, leading and managing the LBA project. We also thank D. Wickland (NASA) for more than a decade of leadership and support for the LBA-Eco project component of LBA. We thank the LBA-Eco team members who contributed to discussions on an early draft of this manuscript at a workshop in Foz do Iguaçu in August 2010, and S. Saleska for comments on the manuscript. We thank P. Lefebvre and W. Kingerlee for assistance with figure and manuscript preparation. Development of this manuscript was supported by NASA grants NNX08AF63A and NNX11AF20G.

Author Contributions E.A.D. wrote an initial rough draft and edited the final draft of the paper. M.K. and M.T.C. contributed significant final edits. All of the other co-authors participated in an LBA-Eco team meeting at Foz do Iguaçu on August, 13, 2010, where this manuscript was designed, and either were co-leaders of breakout groups or made significant subsequent contributions of subsections of text or figures. All co-authors also provided edits throughout.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to E.A.D. (edavidson@whrc.org).

ARTICLE

doi:10.1038/nature10733

A novel retinoblastoma therapy from genomic and epigenetic analyses

Jinghui Zhang[1]*, Claudia A. Benavente[2]*, Justina McEvoy[2]*, Jacqueline Flores-Otero[2]*, Li Ding[3,4], Xiang Chen[1], Anatoly Ulyanov[1], Gang Wu[1], Matthew Wilson[5,6], Jianmin Wang[7], Rachel Brennan[2], Michael Rusch[1], Amity L. Manning[8], Jing Ma[9], John Easton[1], Sheila Shurtleff[9], Charles Mullighan[9], Stanley Pounds[10], Suraj Mukatira[7], Pankaj Gupta[7], Geoff Neale[7], David Zhao[11], Charles Lu[3], Robert S. Fulton[3,4], Lucinda L. Fulton[3,4], Xin Hong[3,4], David J. Dooling[3,4], Kerri Ochoa[3,4], Clayton Naeve[11], Nicholas J. Dyson[8], Elaine R. Mardis[3,4,12], Armita Bahrami[9], David Ellison[9], Richard K. Wilson[3,4,13], James R. Downing[9] & Michael A. Dyer[2,5,14]

Retinoblastoma is an aggressive childhood cancer of the developing retina that is initiated by the biallelic loss of *RB1*. Tumours progress very quickly following *RB1* inactivation but the underlying mechanism is not known. Here we show that the retinoblastoma genome is stable, but that multiple cancer pathways can be epigenetically deregulated. To identify the mutations that cooperate with *RB1* loss, we performed whole-genome sequencing of retinoblastomas. The overall mutational rate was very low; *RB1* was the only known cancer gene mutated. We then evaluated the role of RB1 in genome stability and considered non-genetic mechanisms of cancer pathway deregulation. For example, the proto-oncogene *SYK* is upregulated in retinoblastoma and is required for tumour cell survival. Targeting SYK with a small-molecule inhibitor induced retinoblastoma tumour cell death *in vitro* and *in vivo*. Thus, retinoblastomas may develop quickly as a result of the epigenetic deregulation of key cancer pathways as a direct or indirect result of *RB1* loss.

Retinoblastoma is a rare childhood cancer of the retina that can develop in a sporadic or a heritable form and is fatal if untreated. When the *RB1* gene was first cloned, it was found to undergo biallelic inactivation in virtually all retinoblastoma tumours[1]. Since then, hundreds of genetic lesions have been identified in human cancer. These genetic lesions can be grouped on the basis of the signalling pathways they affect that have direct or indirect mechanistic links to many of the common cellular properties, or hallmarks, of cancer. Thus, the rate of cancer progression is related to the kinetics of acquisition of multiple genetic lesions and/or epigenetic changes that ultimately lead to activation of growth-signalling pathways, evasion of cell death and senescence, acquisition of limitless replicative potential, sustained angiogenesis, and local tissue invasion and metastasis[2].

RB1 inactivation confers limitless replicative potential to retinoblasts and is rate limiting for retinoblastoma tumorigenesis[3]. However, the mechanisms that enable retinoblastoma cells to acquire the additional hallmarks of cancer remain unknown. Evidence from molecular, cellular and cytogenetic studies suggest that RB1 is required for maintaining chromosomal stability[4,5], and that its loss leads to chromosome instability in cells maintained in culture. These data raise the possibility that *RB1* inactivation may underlie the rapid acquisition of cooperating mutations in key cancer pathways through chromosome instability. Alternatively, epigenetic changes may have a more dominant role in cooperating with the loss of *RB1* retinoblastoma tumorigenesis. RB1 has been implicated in regulating most major epigenetic processes, including microRNA regulation, DNA methylation, histone modification and ATP-dependent chromatin reorganization[6–10]. Thus, inactivation of RB1 in retinoblasts may lead to the rapid epigenetic deregulation of cancer genes that contribute to the essential cellular properties of retinoblastoma.

In this study, the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project characterized the genetic landscape of retinoblastoma. Whole-genome sequencing (WGS) of four retinoblastomas and their paired germline DNA samples showed no genetic lesions in known tumour suppressor genes or oncogenes, other than *RB1* and *MYCN*. More importantly, an orthotopic xenograft derived from one of the primary tumours showed no evidence of clonal variation or new coding-region mutations. This finding suggests that the retinoblastoma genome is more stable than previously believed.

Unlike the genetic landscape of retinoblastoma, the epigenetic profile shows profound changes relative to that observed in normal retinoblasts. One of the most striking results was the induction of the expression of the proto-oncogene spleen tyrosine kinase (*SYK*) in human retinoblastoma. SYK is required for tumour cell survival, and inhibition of SYK with a small-molecule inhibitor caused the degradation of MCL1 and caspase-mediated cell death in retinoblastoma cells in culture and *in vivo*. These findings highlight how comprehensive genetic and epigenetic analyses of tumours can be integrated, leading to the discovery of promising new therapeutic approaches and shedding light on the mechanisms underlying the rapid progression of retinoblastoma following *RB1* inactivation.

Retinoblastoma whole-genome sequencing

We performed WGS analysis on four primary human retinoblastoma samples (Supplementary Information, section 1, and Supplementary Table 1) and on matched normal tissue. Local tumour cell invasion,

[1]Department of Computational Biology and Bioinformatics, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [2]Department of Developmental Neurobiology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [3]The Genome Institute, Washington University School of Medicine in St Louis, St Louis, Missouri 63108, USA. [4]Department of Genetics, Washington University School of Medicine in St Louis, St Louis, Missouri 63108, USA. [5]Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, Tennessee 38163, USA. [6]Department of Surgery, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [7]Hartwell Center for Biotechnology & Bioinformatics, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [8]Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [9]Department of Pathology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [10]Department of Biostatistics, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [11]Department of Information Sciences, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [12]Siteman Cancer Center, Washington University School of Medicine in St Louis, St Louis, Missouri 63108, USA. [13]Department of Medicine, Washington University School of Medicine in St Louis, St Louis, Missouri 63108, USA. [14]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.
*These authors contributed equally to this work.

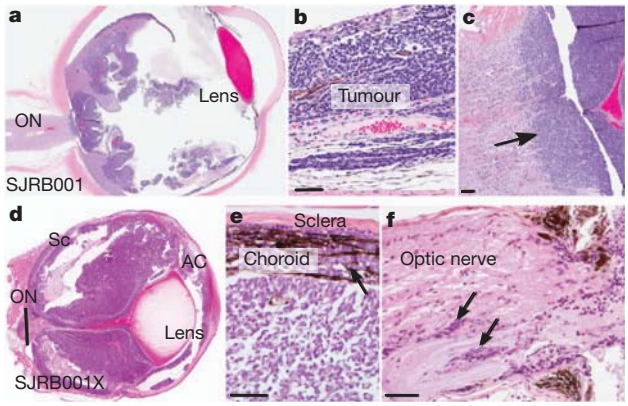


Figure 1 | Characterization of retinoblastomas samples. **a**–**c**, Representative retinoblastoma tumour section (SJRB001) stained with haematoxylin and eosin (H&E), showing choroidal and optic nerve invasion (arrow). **d**–**f**, H&E-stained section of the SJRB001X orthotopic xenograft with choroidal (**e**) and optic nerve (**f**) invasion (arrows). AC, anterior chamber; ON, optic nerve; Sc, sclera. Scale bars, 25 μm.

but not metastasis, was evident in each patient (Fig. 1a–c and Supplementary Fig. 1). We generated an orthotopic xenograft, SJRB001X, of the primary tumour SJRB001 by inoculating primary tumour cells into the vitreous humour of the eyes of immunocompromised mice (Supplementary Information, section 2). SJRB001X had molecular, genetic and histopathologic features similar to those of SJRB001 (Fig. 1d–f; Supplementary Figs 2–4; Supplementary Tables 2–4; and Supplementary Information, section 3).

Using a paired-end sequencing approach, we generated 1,040.9 gigabase pairs (Gb) of sequence data for the samples described; 956.8 Gb (92%) was successfully mapped to the NCBI 36.1 reference genome (Supplementary Information, section 4, and Supplementary Table 5). The average genome coverage was ×28.9, and the average exon coverage was ×23.8 with 98.4% of single nucleotide polymorphisms (SNPs) detected across all nine genomes showing concordance with their corresponding SNP array genotype calls at the same genomic positions (Supplementary Table 5). To provide additional sequence coverage, we performed transcriptome sequencing of all four primary tumours (Supplementary Information, section 5, and Supplementary Table 6).

We identified 668 validated somatic sequence mutations and 40 structural variations across the four cases (Table 1). These included 23 tier-1 mutations in genes, 35 tier-2 mutations in evolutionarily conserved regions of the genome (Supplementary Information, section 6), 309 tier-3 mutations in non-repetitive regions of the genome that are not part of tiers 1 and 2, and 301 tier-4 mutations in repetitive sequences in the genome (Table 1 and Supplementary Table 7). The average number of sequence mutations was 167 per case (range, 56–258), with only 3.25 mutations per case (range, 0–5) resulting in amino-acid changes (Table 1). The estimated mean mutation rate was 6.7×10^{-8} per base (range, 1.03×10^{-7}–2.17×10^{-8}), which is 15-fold less than that in adult tumours analysed by WGS, except for acute myeloid leukaemia[11]. The predominant changes were C→A and G→T transversions (Supplementary Fig. 5), which is consistent with the possibility that some of the transversions result from production of 8-oxoguanine during oxidative stress. Tumour SJRB002 had no somatic sequence variations that resulted in amino-acid changes; the only structural variations were the loss of heterozygosity (LOH) at the *RB1* locus on chromosome 13 and a gain of chromosome 6p. This suggests that very few genetic lesions are required for retinoblastoma progression after *RB1* inactivation.

RB1 inactivation in retinoblastoma

Both *RB1* alleles were inactivated in each sample (Supplementary Figs 6 and 7). Tumours SJRB002 and SJRB003 had mutations in *RB1* combined with copy-number-neutral LOH, and SJRB001 and SJRB004 had somatic sequence mutations combined with *RB1*-promoter hypermethylation (Supplementary Figs 6–8). Deep-sequence analysis of the germline sample from SJRB002 revealed that about 10% of reads contained the R445 nonsense mutation, suggesting the presence of germline chimaerism for the *RB1* mutation (Supplementary Fig. 6). Combining the WGS data with SNP array data of an additional 42 samples, we found that tumours from patients with lower regional nucleotide diversity were much less likely to undergo LOH at the *RB1* locus (Supplementary Tables 8 and 9 and Supplementary Information, section 7). These data show a significant association ($P = 8 \times 10^{-8}$, Fisher's exact test) between a germline genetic variation and mechanism of biallelic *RB1* inactivation in retinoblastoma.

Recurrent lesions in retinoblastoma

To determine whether any of the 11 genes with somatic mutations that caused amino-acid changes or a frameshift in the coding region (Table 1; Fig. 2a, b; Supplementary Figs 9 and 10; Supplementary Information, section 8; and Supplementary Table 10) were recurrently mutated in retinoblastoma, we sequenced all exons from the 11 genes in our recurrent screening cohort of 42 retinoblastomas (Supplementary Information, section 4). Only *BCOR* was recurrently mutated in retinoblastoma (6 of 46, or 13%). Five of the samples had *BCOR* mutations that resulted in truncation of the encoded protein, and one sample had a focal gene deletion (Fig. 2c, Supplementary Table 11 and Supplementary Fig. 11).

We also used the WGS data to identify somatic structural variations including whole-chromosome gains and losses, focal deletions (DEL), insertions (INS), inversions (INV), intrachromosomal rearrangements (ITX), interchromosomal rearrangements (CTX) and regions of LOH (Fig. 2a–c; Supplementary Information, section 9; Supplementary Table 12; and Supplementary Fig. 12). The average

Table 1 | Somatic mutations and structural alterations in retinoblastoma

Sample*	Tier 1†	Non-silent tier 1‡	Genes	Tier 2‖	Tier 3¶	Tier 4#	Total	Mutation rate	Structural variations**
SJRB001 D, G	7	4	*RB1*, *CCNC*, *AGMO*, *RHBG*	16	117	85	225	1.03×10^{-7}	4
SJRB001 X, D§	0	0	NA	8	68	9	85	5.87×10^{-8}	4
SJRB002 D, G	1	0	NA	1	25	29	56	2.17×10^{-8}	0
SJRB003 D, G	7	4	*RB1*, *HNMT*, *LHX8*, *STOML2*	5	67	50	129	5.79×10^{-8}	24
SJRB004 D, G	8	5	*RB1*, *CD300LG*, *SDK1*, *TXK*, *DMWD*	13	100	137	258	8.63×10^{-8}	12

Background mutation rate was calculated on the basis of the ratio of tier-3 mutations to tier-3 bases covered at least tenfold in tumour and germline samples for each pair.
* D, diagnostic tumour sample; G, germline (blood DNA) sample; X, xenograft sample.
† Tier 1 mutations are found in genes and include exons, 5′ and 3′ untranscribed regions, and splice sites. Introns are not included.
‡ Non-silent tier-1 mutations change amino acids in genes.
§ All of the somatic mutations in SJRB001 D, G were identified in SJRB001X. This row highlights the new mutations acquired in the xenograft relative to the primary tumour.
‖ Tier-2 mutations are found in regions of the genome that are conserved between humans and mice.
¶ Tier-3 mutations are found in regions of the genome that are not evolutionarily conserved.
# Tier-4 mutations are found in repetitive regions of the genome.
** Structural variations include focal amplifications and deletions, LOH, and interchromosomal and intrachromosomal translocations.

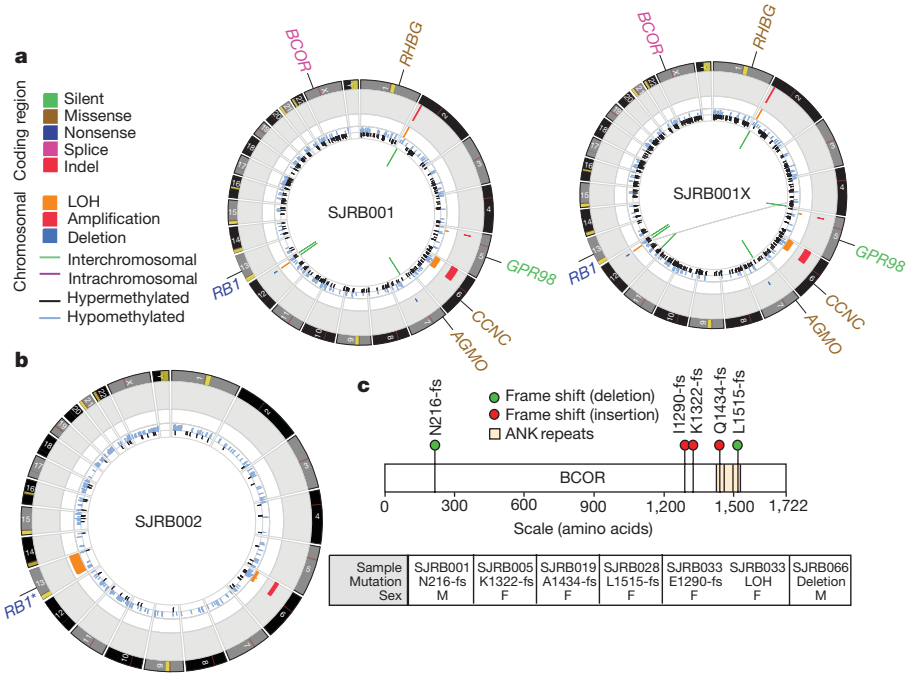


Figure 2 | Genomic profiles of SJRB001, SJRB002 and SJRB001X. **a**, **b**, CIRCOS plots of genetic alterations in two retinoblastomas and the matched orthotopic xenograft. Loss of heterozygosity (orange), amplifications (red) and deletions (blue) are shown. Interchromosomal translocations (green lines) and intrachromosomal translocations (purple lines) are indicated. Sequence mutations in NCBI RefSeq genes included silent single nucleotide variants (SNVs; green), missense SNVs (brown), nonsense SNVs (dark blue), splice-site mutations (pink) and insertion/deletion mutations (indels, red). *Germline mosaic. **c**, *BCOR* mutations identified in the discovery and recurrency cohort.

number of structural variations was ten per case (range, 0–24) (Supplementary Table 1). Tumour SJRB001 had four structural variations (two DEL and ins INS) including a gain of a region of chromosome 2 spanning *MYCN* (Supplementary Table 12), and the only chromosomal lesion in SJRB002 was a gain of chromosome 6p, which occurs in about 40% of human retinoblastomas[12] (Fig. 2b). Only a few genomic regions were affected by the structural variations in SJRB003 and SJRB004 (Supplementary Figs 10, 13 and 14 and Supplementary Table 12).

Orthotopic retinoblastoma xenograft

The genomic landscape of the orthotopic xenograft was remarkably similar to that of the primary tumour, despite continuous growth and multiple passages *in vivo* over nine months (Fig. 2 and Supplementary Table 1). All of the sequence mutations and structural variations detected in SJRB001 were retained in SJRB001X. Only 67 new SNVs and four structural variations were identified in the xenograft, and none affected annotated genes (Fig. 2 and Supplementary Table 1). Moreover, each mutation was identified at a subclonal level (range, 20–30%) and the mutant allele frequency for the lesions in tiers 1–4 in SJRB001 was retained in SJRB001X (Supplementary Fig. 15).

This result was surprising because several studies in mice and cell cultures have linked *RB1* inactivation to defects in chromosome segregation that result in aneuploidy[4,5,13–15] and chromosome instability[4]. We measured the distance between sister chromatids, the distance between kinetochores and the proportion of lagging chromatids in two *RB1*-deficient human retinoblastoma orthotopic xenografts[16] (SJRB001X and SJRB002X). Consistent with results from *RB1*-deficient retinal pigmented epithelium cells[4], the distances between sister chromatids and between kinetochores were increased, and there was evidence of lagging chromosomes (Supplementary Figs 16 and 17). However, less variation in ploidy was observed during spectral karyotype analysis of SJRB001X and SJRB002X, which was more consistent with the ploidy of wild-type cells (Fig. 3a, b and Supplementary Table 13). Moreover, copy number variations were much lower in our cohort of 46 retinoblastomas than in tumours with known genome instability such as ovarian cancer (Fig. 3c). Together, the cytogenetic data and WGS data suggest that the genome is stable and that newly acquired lesions do not provide a selective growth advantage and are thus probably passenger mutations (Supplementary Information, section 9).

Identifying deregulated cancer pathways

There are many examples over the past several decades of epigenomic changes such as DNA methylation contributing to tumorigenesis[17–19]. Indeed, a recent study demonstrated changes in DNA methylation in Wilms' tumours[20], which tend, like retinoblastomas, to have stable genomes. To explore whether epigenetic deregulation of genes or pathways promotes tumorigenesis in retinoblastoma, we carried out an integrative analysis of chromatin immunoprecipitation (ChIP) data, DNA methylation data and gene expression data using order statistics. The SJRB001X sample was used for ChIP assay (Supplementary Figs 18–20), and primary tumour and xenograft samples were used for both DNA methylation and gene expression assays. In all three analyses, experimental results in retinoblastoma tumours were compared with those from human fetal retinae. In this comparison, a total of 104 genes, including 15 known cancer genes (Fig. 4a and Supplementary Tables 14 and 15), were found to have significant differences, indicating that several key cancer genes were epigenetically deregulated.

SYK is a novel therapeutic target

SYK is the fifth most significant gene identified by the integrative analysis and the only upregulated kinase gene (Supplementary Table 15 and Fig. 4a). *SYK* is expressed throughout the haematopoietic system, regulates immunomodulatory signalling and has been implicated in several haematologic malignancies[21–24]. Small-molecule inhibitors of SYK have been developed to treat autoimmune disorders[25], and two of those agents, BAY 61-3606 and R406, have shown efficacy in preclinical leukaemia studies[26–28].

ChIP-on-chip analysis showed increased activating histone modifications (H3K4me3 and K3K9/14Ac) at *SYK*'s promoter, and the repressive histone marker (H3k9me3) was unchanged. Binding of

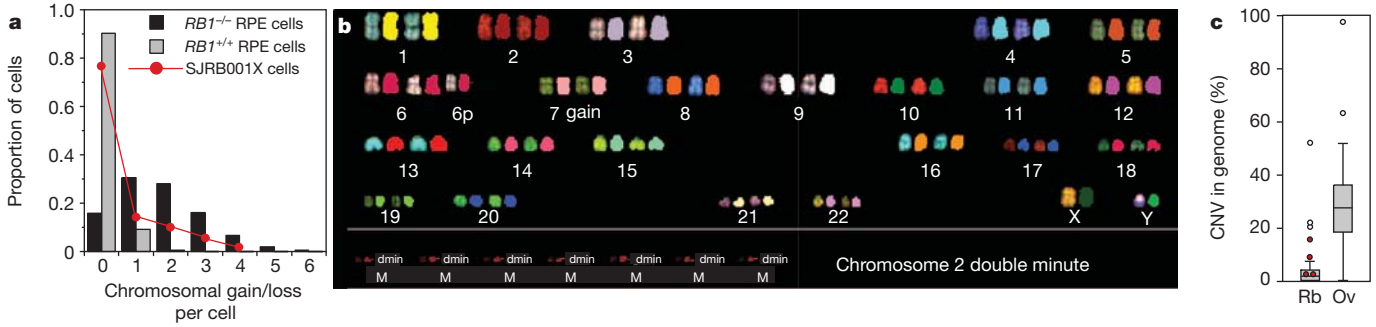


Figure 3 | Analysis of aneuploidy and chromosome instability in retinoblastoma. **a**, Chromosomal mis-segregation of SJRB001X cells after at least 21 rounds of cell division is plotted in red. RPE, retinal pigment epithelium. **b**, Representative spectral karyotype image of SJRB001X after the third passage in mice. **c**, Alterations in the 46 retinoblastoma (Rb) cases relative to 153 high-grade serous ovarian cancer (Ov) cases from The Cancer Genome Atlas. The median percentage of the genome involved in copy number variations (CNVs) was 1.5% for retinoblastoma and 27.7% for ovarian cancer. Red circles, samples used for WGS.

RNA polymerase II to the *SYK* promoter was also increased (Fig. 4b and Supplementary Tables 14 and 15). These ChIP-on-chip results were validated in independent samples by real-time PCR analysis with reverse transcription (Fig. 4c), and we confirmed the increase in *SYK* gene expression (Fig. 4d). SYK protein levels were higher in human retinoblastoma orthotopic xenografts and cell lines than in human fetal retinae (Fig. 4e). To determine whether SYK is expressed in primary human retinoblastomas, we performed immunohistochemistry on a retinoblastoma tissue microarray or whole-eye sections. In total, 100% (82 of 82) showed very strong expression (3+) of SYK in all tumour cells; SYK was not expressed in normal retina (Fig. 4f). SYK's kinase activity is regulated through autophosphorylation at the Tyr 525 and Tyr 526 residues within its catalytic domain. These sites were phosphorylated in retinoblastoma cells, and this phosphorylation was reversed by exposure to BAY 61-3606 or R406 (Fig. 4g and data not shown).

To determine whether *SYK* expression is required for retinoblastoma growth, survival or both, we generated a short hairpin RNA (shRNA)

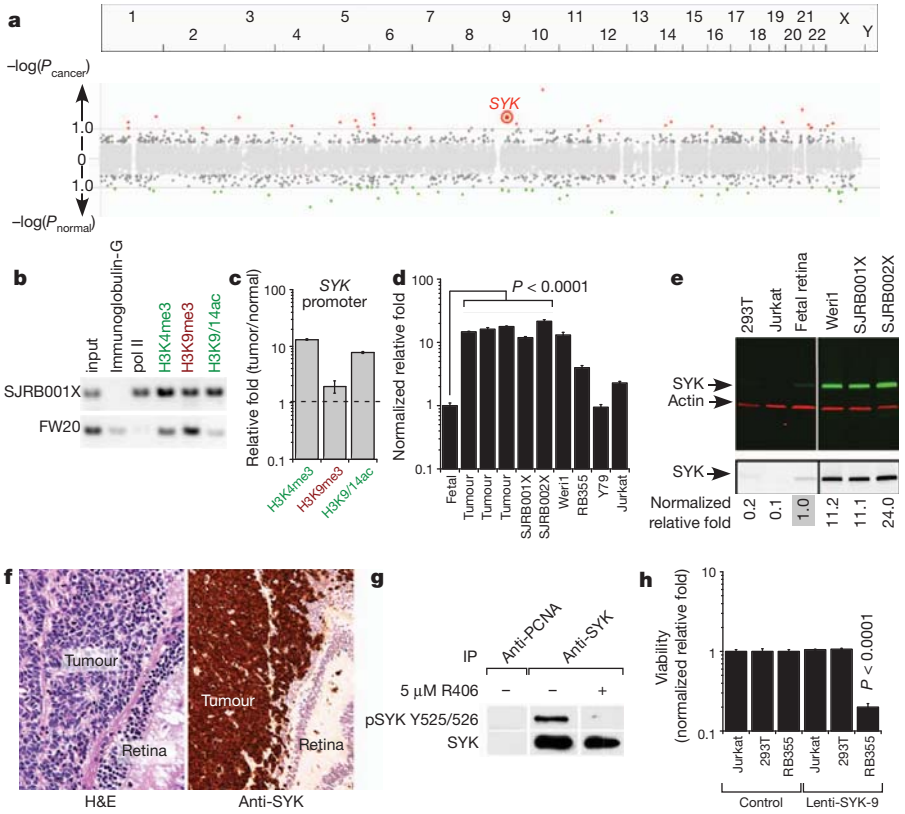


Figure 4 | SYK is expressed in retinoblastoma and is required for survival. **a**, Whole-genome view of the gene ranks based on integrating ChIP-on-chip, methylation and gene expression results. The *y* axis shows $-\log(P)$, where P is the P value of the Q statistic corrected for multiple testing. Significantly (false discovery rate, ≤10%) downregulated (green) or upregulated (red) genes are shown. **b**, **c**, ChIP validation of histone markers of the *SYK* promoter including quantification by quantitative PCR with TaqMan probes. In **c**, each bar shows mean ± s.d. of triplicate samples. **d**, *SYK* expression measured by quantitative PCR in fetal-week-20 retina (fetal), primary retinoblastoma (tumour), orthotopic xenografts (SJRB001X and SJRB002X) and cell lines. Each bar is mean ± s.d. of duplicate samples normalized to GPI1 expression. **e**, Immunoblot of SYK (green) and actin (red) in orthotopic xenografts, human fetal retina and representative cell lines; black and white representation of the SYK immunoblot is in the lower panel. **f**, H&E (purple) and anti-SYK (brown) immunohistochemistry of retinoblastoma tissue. **g**, Immunoprecipitation analysis of SYK and pSYK Y525/526 from Weri1 retinoblastoma cells. **h**, Viability was measured in triplicate cultures 72 h after infection of retinoblastoma cells with a lentivirus vector expressing either a control lentivirus or an shRNA against *SYK*. Scale bars, 10 μm.

against SYK and cloned it into the lentiviral vector Lenti-SYK-9. Lenti-SYK-9 efficiently knocked down *SYK* in retinoblastoma cell lines (Supplementary Fig. 21) and drastically increased apoptosis in retinoblastoma cells (Fig. 4h and Supplementary Fig. 21). Similar results were obtained *in vivo* using SJRB001X (data not shown). We used an empty lentiviral vector and a lentiviral vector encoding an SYK shRNA that less effectively reduced SYK expression (Lenti-SYK-6) as controls. Cell lines that do not express SYK (BJ, 293T and uninduced Jurkat cells) were used as controls and the Lenti-SYK-9 lentivirus had no effect on the growth or apoptosis of the control cells.

We exposed retinoblastoma cell lines that express high levels of *SYK* (Weri1 and RB355) to various concentrations of the SYK inhibitors BAY 61-3606 and R406 for 72 h and then measured cell viability. Jurkat (uninduced) and 293T cells were used as negative controls. Weri1 and RB355 cells were sensitive to both SYK inhibitors, but the Jurkat and 293T cells were unaffected (Fig. 5a). Transmission electron microscopy of retinoblastoma cells treated with the SYK inhibitors showed morphologic features consistent with cell death and mitochondrial defects (Supplementary Fig. 22); this was confirmed by scoring the proportion of activated caspase 3^+ cells (Fig. 5b, c and Supplementary Fig. 22). Jurkat cells showed no increase in activated caspase 3^+ cells after treatment with 10 μM R406 or BAY 61-3606 (data not shown).

The proportions of cells from each line that incorporated EdU were similar, suggesting that retinoblastoma's cell cycle is not affected by SYK inhibition (Fig. 5d, e and Supplementary Fig. 22). The effects of targeting SYK were partly rescued by the pan-caspase inhibitor Q-VD-OPH (Supplementary Fig. 22). Less MitoTracker Red accumulated in retinoblastoma cells exposed to the SYK inhibitors (Supplementary Fig. 22), yet treatment of Jurkat cells with either inhibitor had no effect on MitoTracker Red accumulation (data not shown). Inhibiting SYK in B-CLL cells reduced their autophosphorylation of SYK and destabilized MCL1[26,29], the only anti-apoptotic member of the BCL2 family that is upregulated in retinoblastoma (data not shown). Both inhibitors reduced autophosphorylation of SYK on Y525/526 (Fig. 4g) and reduced MCL1 expression coincident with apoptosis (Fig. 5f).

We tested the efficacy of BAY 61-3606 *in vivo* using our SJRB001X model[30]. The chemotherapy regimen consisted of a single subconjunctival dose of BAY 61-3606 on day 1 and daily doses of topotecan (TPT) on days 1–5 until either six courses (21 days per course) were administered or the tumour had progressed and surgical enucleation was required (Fig. 5g, h). BAY 61-3606 plus TPT significantly improved the outcome ($P = 0.003$; Fig. 5i), and its efficacy was correlated with an increase in activated caspase 3^+ cells in the treated eyes (Fig. 5j). Previous studies using this model have shown that TPT combined with subconjunctival carboplatin had no effect on tumour response or outcome[30], so the improvement seen here can be attributed to targeting SYK. MCL1 expression was also reduced in the treated eyes, which is consistent with increased apoptosis and targeting of SYK *in vivo* (Fig. 5k). Together, these results suggest that SYK is a promising new target for treating retinoblastoma.

Discussion

Genome sequencing reveals that retinoblastomas have a relatively stable genome. The mutational rate and number of structural variations per case that we assessed were among the lowest reported in human cancer to date. Moreover, in one example (SJRB002) the only non-silent mutation found was in *RB1*, and only two structural variations were detected. The minimal increase in passenger mutations in SJRB001X cells, despite prolonged passage, was also consistent with a relatively stable genome. These results are surprising because previous studies have shown that the functional inactivation of RB1 can cause

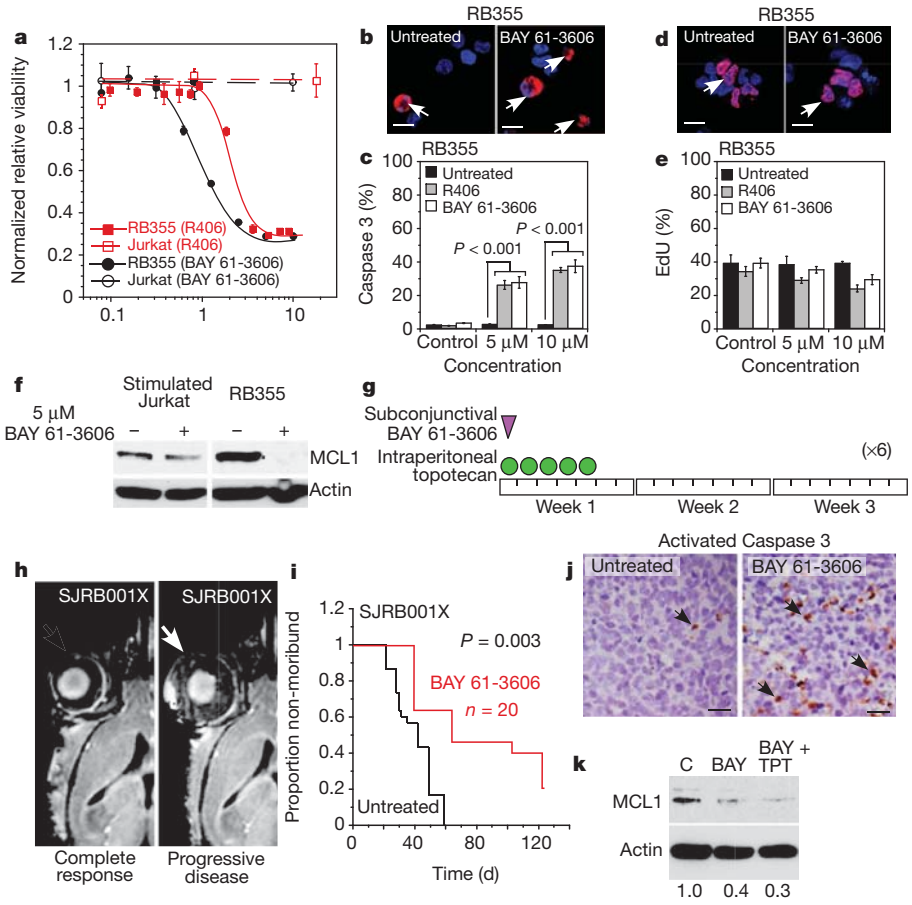


Figure 5 | Retinoblastoma cells are sensitive to SYK inhibitors. **a,** Dose response for SYK inhibitors R406 (red) and BAY 61-3606 (black) in RB355 retinoblastoma cells and a negative control (Jurkat). Each data point is mean ± s.d. of triplicate samples. **b–e**, Immunofluorescence of activated caspase 3 (**b**, **c**) or EdU (**d**, **e**) (red) before and after treatment of RB355 cells with R406 or BAY 61-3606. A total of 250 cells were scored in duplicate for each sample and each treatment condition to derive the mean and s.d. Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI, blue). **f**, Treatment of stimulated Jurkat or RB355 cells with 5 μM BAY 61-3606 for 24 h reduced MCL1 expression. **g**, Schematic of the treatment schedule for mice with SJRB001X tumours. **h**, Representative magnetic resonance images of a mouse whose tumour responded after four courses of treatment with BAY 61-3606 (left) and another whose disease progressed during treatment (right). **i**, Survival curves show that treatment with BAY 61-3606 plus TPT improved outcome. **j**, Immunostaining for activated caspase 3 (arrows) in eyes either untreated or treated with BAY 61-3606. **k**, Immunoblot showing reduction in MCL1 after treatment with BAY 61-3606 (BAY) or BAY 61-3606 plus TPT. C, control. Scale bars, 5 μm (**b**, **d**); 10 μm (**j**).

genomic instability. Although retinoblastoma cells show the mitotic defects that can lead to errors in chromosome segregation, the ploidy of retinoblastoma cells is stable *in vitro* and *in vivo*[31,32]. We do not exclude the possibility that RB1 inactivation promotes change in other contexts, but our data show that genomic instability is not a hallmark of retinoblastoma and do not explain how retinoblastomas progress so rapidly.

We propose that epigenetic mechanisms contribute to retinoblastoma tumorigenesis. We identified several known oncogenes and tumour suppressor genes with histone modifications and altered DNA methylation that correlated with changes in gene expression. Our key discovery was that SYK is important in retinoblastoma. Retinal progenitor cells and retinal neurons express little or no SYK, and SYK has no known function in the developing visual system. Moreover, no recurrent genetic lesions in *SYK* were identified by WGS or SNP array analysis to suggest that this gene drives retinoblastoma tumorigenesis. Only by integrating epigenetic and gene expression analyses did we identify SYK as an important oncogene in retinoblastoma. This is important not only for expanding our understanding of the biology of retinoblastoma but also for advancing immediate therapeutic options that were not previously considered, such as the use of BAY 61-3606 or R406. This study highlights the value of integrating WGS analyses of the genetic and epigenetic features of tumour genomes to finding a cure for cancers such as retinoblastoma.

METHODS SUMMARY

Full details of sample acquisition, molecular and biochemical procedures, informatics and whole genome sequencing, and animal and drug studies are provided in Supplementary Information. The St Jude Children's Research Hospital Institutional Review Board approved experiments involving human subjects and informed consent was obtained from all subjects. For animal studies, all experiments were performed in accordance with federal guidelines and regulations. The St Jude Children's Research Hospital Institutional Animal Care and Use Committee approved all animal experiments. Lentiviral vectors (GIPZ with Lenti-SYK-9 #V3LHS-366147 and Lenti-SYK-6 #V3LHS-366143) encoding shRNAs against SYK were purchased from OpenBiosystems.

Received 27 March; accepted 24 November 2011.
Published online 11 January 2012.

1. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323,** 643–646 (1986).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
3. Knudson, A. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68,** 820–823 (1971).
4. Manning, A. L., Longworth, M. S. & Dyson, N. J. Loss of pRB causes centromere dysfunction and chromosomal instability. *Genes Dev.* **24,** 1364–1376 (2010).
5. Hernando, E. *et al.* Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. *Nature* **430,** 797–802 (2004).
6. Chi, P., Allis, C. D. & Wang, G. G. Covalent histone modifications – miswritten, misinterpreted and mis-erased in human cancers. *Nature Rev. Cancer* **10,** 457–469 (2010).
7. Lu, J., Ruhf, M. L., Perrimon, N. & Leder, P. A genome-wide RNA interference screen identifies putative chromatin regulators essential for E2F repression. *Proc. Natl Acad. Sci. USA* **104,** 9381–9386 (2007).
8. Benetti, R. *et al.* A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. *Nature Struct. Mol. Biol.* **15,** 998 (2008).
9. Wen, H., Andrejka, L., Ashton, J., Karess, R. & Lipsick, J. S. Epigenetic regulation of gene expression by *Drosophila* Myb and E2F2-RBF via the Myb-MuvB/dREAM complex. *Genes Dev.* **22,** 601–614 (2008).
10. Bourgo, R. J. *et al.* SWI/SNF deficiency results in aberrant chromatin organization, mitotic failure, and diminished proliferative capacity. *Mol. Biol. Cell* **20,** 3192–3199 (2009).
11. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
12. Corson, T. W. & Gallie, B. L. One hit, two hits, three hits, more? Genomic changes in the development of retinoblastoma. *Genes Chromosom. Cancer* **46,** 617–634 (2007).
13. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460,** 278–282 (2009).
14. Iovino, F., Lentini, L., Amato, A., &. Di Leonardo, A. RB acute loss induces centrosome amplification and aneuploidy in murine primary fibroblasts. *Mol. Cancer* **5,** 38 (2006).
15. Amato, A., Lentini, L., Schillaci, T., Iovino, F. & Di Leonardo, A. RNAi mediated acute depletion of retinoblastoma protein (pRb) promotes aneuploidy in human primary cells via micronuclei formation. *BMC Cell Biol.* **10,** 79 (2009).
16. McEvoy, J. *et al.* Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. *Cancer Cell* **16,** 260–275 (2011).
17. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nature Rev. Cancer* **4,** 143–153 (2004).
18. Jones, P. A. & Laird, P. W. Cancer epigenetics comes of age. *Nature Genet.* **21,** 163–167 (1999).
19. Laird, P. W. Cancer epigenetics. *Hum. Mol. Genet.* **14** (suppl. 1), R65–R76 (2005).
20. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genet.* **43,** 768–775 (2011).
21. Hahn, C. K. *et al.* Proteomic and genetic approaches identify Syk as an AML target. *Cancer Cell* **16,** 281–294 (2009).
22. Chen, L. *et al.* SYK-dependent tonic B-cell receptor signaling is a rational treatment target in diffuse large B-cell lymphoma. *Blood* **111,** 2230–2237 (2008).
23. Feldman, A. L. *et al.* Overexpression of Syk tyrosine kinase in peripheral T-cell lymphomas. *Leukemia* **22,** 1139–1143 (2008).
24. Young, R. M. *et al.* Mouse models of non-Hodgkin lymphoma reveal Syk as an important therapeutic target. *Blood* **113,** 2508–2516 (2009).
25. Weinblatt, M. E. *et al.* An oral spleen tyrosine kinase (Syk) inhibitor for rheumatoid arthritis. *N. Engl. J. Med.* **363,** 1303–1312 (2010).
26. Baudot, A. D. *et al.* The tyrosine kinase Syk regulates the survival of chronic lymphocytic leukemia B cells through PKCδ and proteasome-dependent regulation of Mcl-1 expression. *Oncogene* **28,** 3261–3273 (2009).
27. Suljagic, M. *et al.* The Syk inhibitor fostamatinib disodium (R788) inhibits tumor growth in the Eμ-TCL1 transgenic mouse model of CLL by blocking antigen-dependent B-cell receptor signaling. *Blood* **116,** 4894–4905 (2010).
28. Buchner, M. *et al.* Spleen tyrosine kinase inhibition prevents chemokine- and integrin-mediated stromal protective effects in chronic lymphocytic leukemia. *Blood* **115,** 4497–4506 (2010).
29. Gobessi, S. *et al.* Inhibition of constitutive and BCR-induced Syk activation downregulates Mcl-1 and induces apoptosis in chronic lymphocytic leukemia B cells. *Leukemia* **23,** 686–697 (2009).
30. Brennan, R. C. *et al.* Targeting the p53 pathway in retinoblastoma with subconjunctival Nutlin-3a. *Cancer Res.* **71,** 4205–4213 (2011).
31. Yan, Y. *et al.* Engraftment and growth of patient-derived retinoblastoma tumour in severe combined immunodeficiency mice. *Eur. J. Cancer* **36,** 221–228 (2000).
32. Squire, J., Gallie, B. L. & Phillips, R. A. A detailed analysis of chromosomal changes in heritable and non-heritable retinoblastoma. *Hum. Genet.* **70,** 291–301 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The WGS was supported as part of the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project. We thank J. Lahti, J. Dalton and M. Valentine for help with FISH analysis, L. Holmfeldt, J. Zhang and M. Barbato for help with sample preparation, and V. Valentine for spectral karyotype analysis. We thank I. Qaddoumi, C. Rodriguez-Galindo and B. Haik for the support of the St Jude Children's Research Hospital retinoblastoma clinical research and W. Lei, D. McGoldrick, D. Alford, S. Espy, J. Obenauer and K. Johnson for assistance with data acquisition, handling and analysis. We thank A. McArthur and C. Guess for editing the manuscript, J. Temirov for help with sister chromatid analysis, J. Thurman for help with histology, F. Krafcik for help with cell culture, J. Wu and C. Billups for statistical analysis, J. Jeon for help with lentiviral preparations and C. Bradley for assistance with preclinical testing. This work was supported, in part, by Cancer Center Support (CA21765) from the NCI; grants to M.A.D. from the NIH (EY014867 and EY018599), the American Cancer Society and the Research to Prevent Blindness Foundation; and the American Lebanese Syrian Associated Charities. M.A.D. is a Howard Hughes Medical Institute Early Career Scientist. This work was also supported by an American Cancer Society Fellowship to A.L.M., the MGH Cancer Center Saltonstall Foundation Scholarship to N.J.D. and funding from AstraZeneca and NIH grants GM81607 and CA64402 to N.J.D.

Author Contributions M.A.D., C.A.B., J. McEvoy, J.F.-O., R.B. and A.L.M. designed the experiments. L.D. and J.Z. led the data analysis. M.A.D., L.D., J.Z., X.C., A.U., G.W., J.W., M.R., J. Ma, S.P., S.M., P.G., G.N., D.Z., C.L., R.S.F., L.L.F., X.H., D.J.D. and K.O. performed the data analysis and provided bioinformatics support. M.A.D., J.Z., A.U., C.A.B., J. McEvoy, J.F.-O., R.B. and A.L.M. prepared the figures and tables. C.A.B, J.M., J.F.-O., R.B. and A.L.M. performed the laboratory experiments. J.F.-O. created the xenograft. M.W. provided samples. J.E., S.S. and C.M. provided assistance with sample processing and handling. A.B. and D.E. provided pathology support. M.A.D., C.N., E.R.M., R.K.W. and J.R.D. supervised the project. M.A.D. wrote the manuscript. J.Z., C.A.B., J. McEvoy, J.F.-O., L.D., M.W., E.R.M., S.S., C.M. and J.R.D. read and commented on the manuscript. J.Z., L.D., X.C., A.U., G.W., J.W., M.R., J. Ma, J.E., S.S., C.M., S.P., S.M., P.G., G.N., D.Z., C.L., R.S.F., L.L.F., X.H., D.J.D., K.O., C.N., E.R.M., A.B., D.E., R.K.W., J.R.D. and M.A.D. are part of the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project.

Author Information The whole-genome sequence data and SNP 6.0 data have been deposited in the NCBI dbGaP under accession number phs000352.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A.D. (michael.dyer@stjude.org), J.R.D. (james.downing@stjude.org) or R.K.W. (rwilson@wustl.edu).

ARTICLE

doi:10.1038/nature10728

Structure of HDAC3 bound to co-repressor and inositol tetraphosphate

Peter J. Watson[1], Louise Fairall[1], Guilherme M. Santos[1] & John W. R. Schwabe[1]

Histone deacetylase enzymes (HDACs) are emerging cancer drug targets. They regulate gene expression by removing acetyl groups from lysine residues in histone tails, resulting in chromatin condensation. The enzymatic activity of most class I HDACs requires recruitment into multi-subunit co-repressor complexes, which are in turn recruited to chromatin by repressive transcription factors. Here we report the structure of a complex between an HDAC and a co-repressor, namely, human HDAC3 with the deacetylase activation domain (DAD) from the human SMRT co-repressor (also known as NCOR2). The structure reveals two remarkable features. First, the SMRT-DAD undergoes a large structural rearrangement on forming the complex. Second, there is an essential inositol tetraphosphate molecule—D-myo-inositol-(1,4,5,6)-tetrakisphosphate (Ins(1,4,5,6)P_4)—acting as an 'intermolecular glue' between the two proteins. Assembly of the complex is clearly dependent on the Ins(1,4,5,6)P_4, which may act as a regulator—potentially explaining why inositol phosphates and their kinases have been found to act as transcriptional regulators. This mechanism for the activation of HDAC3 appears to be conserved in class I HDACs from yeast to humans, and opens the way to novel therapeutic opportunities.

The acetylation of lysine residues in the tails of histone proteins plays an important role in the regulation of gene expression in eukaryotic cells[1,2]. The level of lysine acetylation is controlled through the opposing actions of histone acetyl transferases (HATs) and HDACs. Although chromatin is the best understood substrate for these enzymes, lysine acetylation is emerging as a general regulatory mechanism in a diverse array of cellular processes[3].

There are four classes of HDACs in mammalian cells (reviewed in ref. 4). Class I HDACs are zinc-dependent enzymes, and include HDAC1, 2, 3 and 8. Of these, only HDAC8 is a fully functional enzyme in isolation[5,6]. HDAC1, 2 and 3 require recruitment into large multi-subunit co-repressor complexes for maximal activity[7–12]. These co-repressor complexes bring about the repression of gene expression when recruited to repressive transcription factors, but also contribute to the 'resetting' of chromatin after rounds of transcriptional activation[13–16].

In recent years, HDACs have become important targets for the treatment of a number of cancers[17]. Cancer cell lines treated with HDAC inhibitors undergo terminal differentiation, growth arrest and/or apoptosis. Several HDAC inhibitors are at various stages in clinical trials, and two drugs, vorinostat and romidepsin, have been approved for the treatment of cutaneous T-cell lymphomas[18].

HDAC1 and 2 are found in three co-repressor complexes: NuRD[7,19], CoREST[20,21] and Sin3A[22,23]. In contrast, HDAC3 appears to be uniquely recruited to the SMRT complex (or to the homologous NCoR complex) where it interacts with a conserved DAD within SMRT or NCoR[8,10,12,24–26]. The DAD both recruits and activates HDAC3[10–12]. Recruitment of HDAC3 to the DAD is essential for repression by certain nuclear receptors and for the maintenance of normal circadian physiology[27–29]. It has been proposed that the assembly of the HDAC3 and SMRT-DAD requires a chaperone complex, because when these proteins are expressed in bacteria they do not form a complex[30].

The DAD contains an extended SANT-like domain with an amino-terminal DAD-specific motif. Deletion of this motif results in both loss of binding and failure to activate HDAC3[11].We have previously reported the structure of the isolated DAD from SMRT[31]. This revealed that part of the DAD-specific motif forms an extra helix that is folded against the three helices of the SANT domain to form a four-helix bundle. The N-terminal portion of the DAD-specific motif is unstructured in solution[31].

Here we report the structure of HDAC3 in complex with SMRT-DAD. This structure not only reveals the specificity and mechanism through which SMRT-DAD recruits and activates HDAC3, but also identifies Ins(1,4,5,6)P_4 as a key component of the complex that has the potential to regulate assembly of HDACs with their co-repressors.

Overall architecture of the complex

As HDAC3 and SMRT-DAD do not form a complex when expressed in bacterial cells, full-length HDAC3 and Flag-tagged SMRT-DAD (amino acids 389–480) were expressed in suspension-grown mammalian HEK293 cells. The complex remained tightly associated during a three-step purification, including size exclusion chromatography (Supplementary Fig. 1). Interestingly, at salt concentrations higher than 50 mM, the complex dissociated and the HDAC3 predominately took the form of aggregate or oligomers (Supplementary Fig. 2). Such oligomerization of HDAC3 has also been reported by others[32]. During crystallization, the HDAC3 tail is proteolysed but HDAC3 remains bound to SMRT-DAD and retains deacetylase activity (Supplementary Fig. 3). Small crystals (15 μm; Supplementary Fig. 1) diffracted to 2 Å and the structure was solved by molecular replacement with an HDAC8 structure[33]. Model building and refinement yielded an excellent map to 2.1 Å resolution with clear density for both HDAC3 and SMRT-DAD (Supplementary Fig. 1).

Overall, the HDAC3 structure is similar to the previously determined class I HDAC structures of HDAC8[34] and HDAC2[35], and consists of an eight-stranded parallel β-sheet surrounded by a number of α-helices. The active site lies at the base of a tunnel leading from the surface of the protein. A solvent-exposed tyrosine residue is located on the surface of the enzyme immediately adjacent to the active site tunnel. This tyrosine is unique to HDAC3, and it seems that this

[1]Henry Wellcome Laboratories of Structural Biology, Department of Biochemistry, University of Leicester, Leicester LE1 9HN, UK.

residue might interact with the substrate and possibly contribute to substrate specificity (Supplementary Fig. 4).

Structural rearrangement of SMRT–DAD

On forming a complex with HDAC3, the N-terminal helix of the DAD undergoes a major structural rearrangement (from that seen in the NMR structure), such that it no longer forms part of the core structure, but lies along the surface of HDAC3, making extensive intermolecular interactions (Fig. 1a, b). Along with a further extended region, this DAD-specific motif buries a surface of 1,178 $Å^2$. The remaining three-helix bundle resembles a canonical SANT domain, and buries a further 1,160 $Å^2$ at the interface with HDAC3.

The SMRT-DAD interacts with the N-terminal region of HDAC3 (residues 9–49) that forms helix H1, loop L2, helix H2 and strand S2. This region differs between HDAC8 and HDAC3. Indeed, in HDAC3, helix H1 is distorted and can only be considered to be a pseudo-helix. These differences may explain why HDAC8 is active in the absence of an interacting co-repressor.

An essential role for Ins(1,4,5,6)P_4

At the earliest stages of refinement, the electron density difference map revealed a well-ordered small molecule bound at the interface between HDAC3 and the DAD (Fig. 2a). The electron density was sufficiently well defined that the small molecule could be readily identified as inositol tetraphosphate. During further refinement, it could be unambiguously assigned as D-myo-inositol-1,4,5,6-tetrakisphosphate (based on the axial orientation of the hydroxyl group on carbon 2), referred to here as Ins(1,4,5,6)P_4 (Fig. 2b).

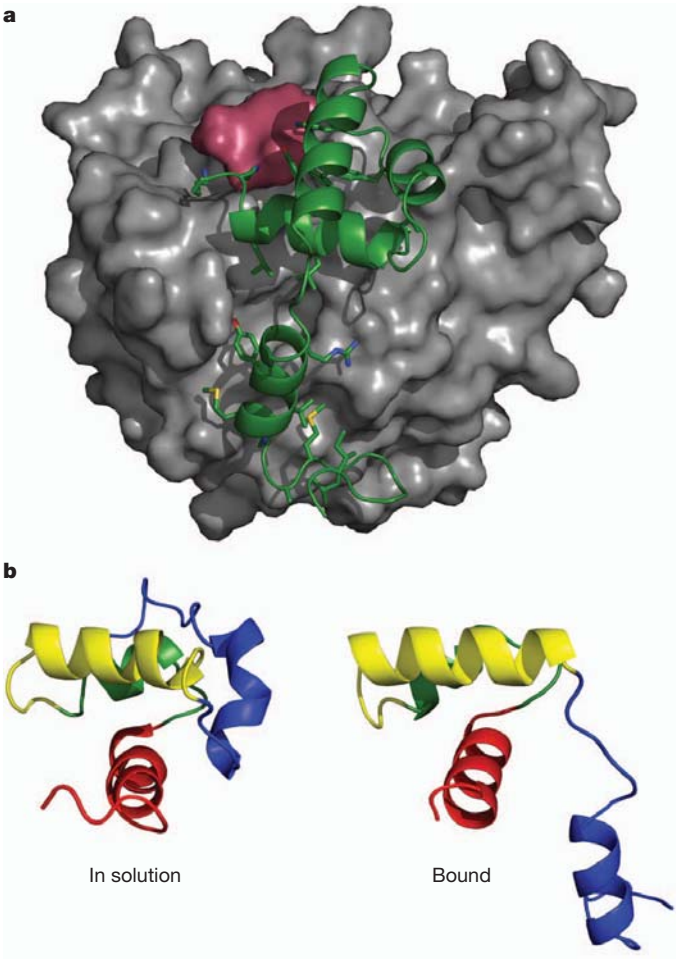


Figure 1 | Structure of the HDAC3–SMRT-DAD complex. **a**, Interaction of SMRT-DAD (green ribbon) with HDAC3 (grey surface). The Ins(1,4,5,6)P_4 at the interface is shown as a raspberry coloured surface. Side chains in the DAD that mediate interaction with HDAC3 and Ins(1,4,5,6)P_4 are shown as sticks. **b**, Structure of the DAD domain in solution (left; PDB code 1XC5) compared with that bound to HDAC3 (right); helices are individually coloured to facilitate comparison.

The Ins(1,4,5,6)P_4 molecule is sandwiched between HDAC3 and the DAD, in a highly basic pocket, making extensive contacts to both proteins, burying a surface of 407 $Å^2$ (Fig. 2c, d). HDAC3 and SMRT-DAD each contribute five hydrogen bonds and salt bridges to the Ins(1,4,5,6)P_4 (His 17, Gly 21, Lys 25, Arg 265, Arg 301 and Lys 449, Tyr 470, Tyr 471, Lys 474 and Lys 475, respectively) (Fig. 2e).

It is notable that the Ins(1,4,5,6)P_4 is sufficiently tightly bound to the complex that it is retained through the entire purification process. It is also striking that the binding is highly specific for Ins(1,4,5,6)P_4, as the electron density shows that the ligand is uniquely Ins(1,4,5,6)P_4 rather than a mixture of inositol phosphates.

A careful examination of the structure suggests that Ins(1,4,5,6)P_4 binding is an essential requirement for the interaction between SMRT and HDAC3, acting as a 'intermolecular glue' that cements the complex together. If the Ins(1,4,5,6)P_4 ligand were not present, then the many basic residues on either side of the binding cleft would probably prevent interaction through charge repulsion. Indeed, at the base of the cleft, the ζ-nitrogen atoms of SMRT Lys 449 and of HDAC3 Lys 25 are just 4.4 Å from each other. Consequently, the assembly of the three-way SMRT–HDAC3–Ins(1,4,5,6)P_4 complex appears to be mutually interdependent, such that both SMRT-DAD and Ins(1,4,5,6)P_4 are required for activation of the HDAC3 enzymatic functionality.

These conclusions are supported by our previous mutagenesis study of SMRT-DAD, in which we looked at the effect of mutations on both the interaction with, and the deacetylase activity of, HDAC3. All mutations that failed to activate HDAC3 also abolished, or significantly impaired, interaction. These include mutations of Lys 449, Tyr 470 and Tyr 471, which play a key role in binding Ins(1,4,5,6)P_4 (ref. 31). The requirement for Ins(1,4,5,6)P_4 to enable complex formation may contribute to the explanation as to why recombinant HDAC3 and SMRT-DAD expressed in bacteria fail to interact, as bacteria probably do not contain sufficient Ins(1,4,5,6)P_4 to support complex formation.

Having discovered that Ins(1,4,5,6)P_4 plays a key role in HDAC3 activation, we asked whether inositol phosphates might contribute to the assembly and activation of other class I HDAC complexes. Significantly, the residues that mediate interaction with Ins(1,4,5,6)P_4 and the SANT domain from SMRT are conserved in both HDAC1 and 2, but not in HDAC8 (Fig. 3a), fitting with the observation that only HDAC8 is fully active in isolation.

Similarly, the co-repressor partners for HDAC1 and 2 (that is, MTA1, 2 and 3 and CoREST1, 2 and 3) contain SANT domains that are analogous to SMRT-DAD. The key Ins(1,4,5,6)P_4 binding residues are almost entirely conserved (Fig. 3b). It is likely that the specificity for the particular HDAC is conferred by the region N-terminal to the SANT domain: that is, the DAD specific motif in SMRT/NCoR and the ELM2 domains in the MTA and CoREST proteins.

Taken together, this suggests that inositol phosphate activation of class I HDACs is a general mechanism and, given that the key residues are also found in the yeast HDAC, Rpd3, and the Snt1 co-repressor, it would appear that this mechanism is also evolutionarily conserved.

Mechanism of activation of HDAC3

The HDAC3–SMRT-DAD structure provides insight into the mechanism through which the DAD and Ins(1,4,5,6)P_4 contribute to the activation of HDAC3. In the crystal, the active site of HDAC3 resembles a product complex (Fig. 4a, b). An acetate molecule (present during purification) is bound at the active site, making hydrogen bonds to the catalytic zinc and side chains of Ty 298 and His 134. Furthermore, a methionine side chain (from an adjacent SMRT-DAD in the crystal lattice) is bound in the active site tunnel, mimicking a lysine residue.

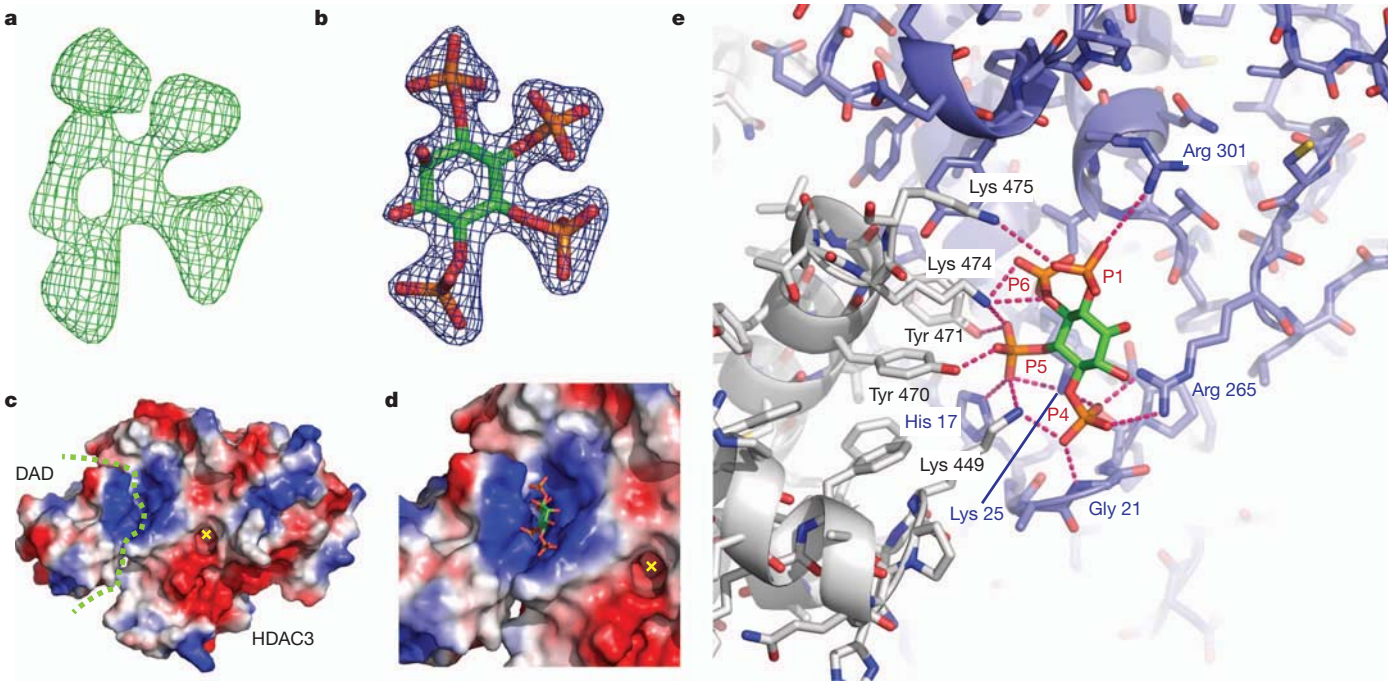


Figure 2 | Ins(1,4,5,6)P_4 binding to the HDAC3–SMRT-DAD complex. **a**, A striking feature in the difference electron density map ($F_o - F_c$ at 3σ, green mesh) observed following molecular replacement. **b**, Electron density corresponding to the Ins(1,4,5,6)P_4 ligand following refinement ($2F_o - F_c$ at 2.25σ, blue mesh). The Ins(1,4,5,6)P_4 is shown as a stick model. **c**, Electrostatic surface representation of the HDAC3–SMRT-DAD complex (red indicates negative potential and blue positive). A strikingly basic pocket is located at the HDAC3–SMRT-DAD interface (indicated by a dashed green line). The active site pocket of HDAC3 is indicated by a yellow cross. The Ins(1,4,5,6)P_4 is not shown for clarity. **d**, Ins(1,4,5,6)P_4 (shown as sticks) binding in the basic pocket at the HDAC3–SMRT-DAD interface. **e**, Detailed interactions of Ins(1,4,5,6)P_4 (sticks) with HDAC3 (blue) and SMRT-DAD (grey). Side chains from SMRT and HDAC3 which make contacts with the Ins(1,4,5,6)P_4 are labelled. The phosphate groups of the Ins(1,4,5,6)P_4 are labelled P1–6. Hydrogen bonds and/or salt bridges are indicated by dashed lines.

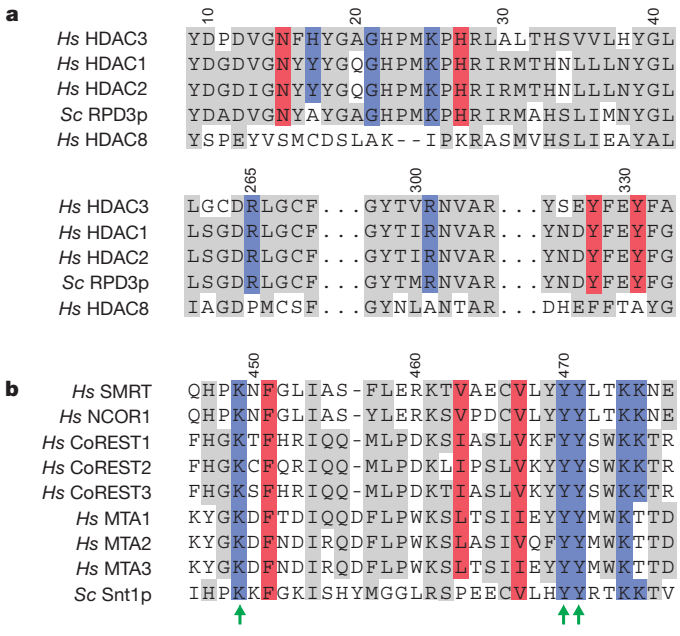


Figure 3 | Sequence conservation suggests that class I HDACs, from yeast to man, require inositol phosphates for assembly and activation. **a**, Alignments of key regions of class I HDACs from *Homo sapiens* (*Hs*) and *Saccharomyces cerevisiae* (*Sc*). Residues that mediate interaction with Ins(1,4,5,6)P_4 and SMRT-DAD are highlighted in blue and red, respectively. Other conserved residues are highlighted in grey. **b**, Alignment of SANT domains from known interaction partners for class I HDACs. Residues that mediate interaction with Ins(1,4,5,6)P_4 and HDAC3 are highlighted in blue and red, respectively. Green arrows indicate residues that impair HDAC3 recruitment and activation when mutated to alanine[31].

The binding surfaces for DAD and Ins(1,4,5,6)P_4 are located to one side of the HDAC3 active site, involving pseudo helix H1 and loops L1 and L6 (Fig. 4a–c). We propose that changes in both conformation and dynamics occur when DAD and Ins(1,4,5,6)P_4 bind to HDAC3, and that these facilitate substrate access to the active site, resulting in enhanced enzyme activity.

There is a key interaction between Ins(1,4,5,6)P_4 and Arg 265 in loop L6 (coloured orange in Fig. 4b). This loop seems to be very important for access to the active site, because Leu 266 forms one wall of the active site tunnel, and in the absence of the Ins(1,4,5,6)P_4 this loop is likely to be relatively mobile.

Comparison of the HDAC3 structure with that of HDAC8, which does not require activation by complex formation, reveals that HDAC8 differs significantly in the region where HDAC3 interacts with SMRT-DAD and Ins(1,4,5,6)P_4 (Fig. 4d). In HDAC8, the equivalent helix to pseudo helix H1 in HDAC3 has a regular helical structure, loop L1 is two amino acids shorter and loop L6 contains a proline residue that partly orientates the loop away from the active site (Fig. 4d). We suggest that together these differences give substrate better access to that active site of HDAC8 than would be possible in the uncomplexed HDAC3. The pattern of crystallographic temperature factors for the various structures supports this interpretation (Supplementary Fig. 5).

To test the importance of Arg 265 as well as loops L1 and L6, we co-expressed mutant HDAC3 constructs with SMRT-DAD in mammalian cells. Strikingly, all the designed mutations resulted not only in total loss of deacetylase activity (Fig. 5a) but also abolished interaction with SMRT-DAD (Fig. 5b). These findings support our interpretation of the importance of these residues in the activation of HDAC3.

To support the conclusions that Ins(1,4,5,6)P_4 is essential for HDAC3 activity and to test the specificity for Ins(1,4,5,6)P_4, we sought to establish an *in vitro* reconstitution assay. We expressed

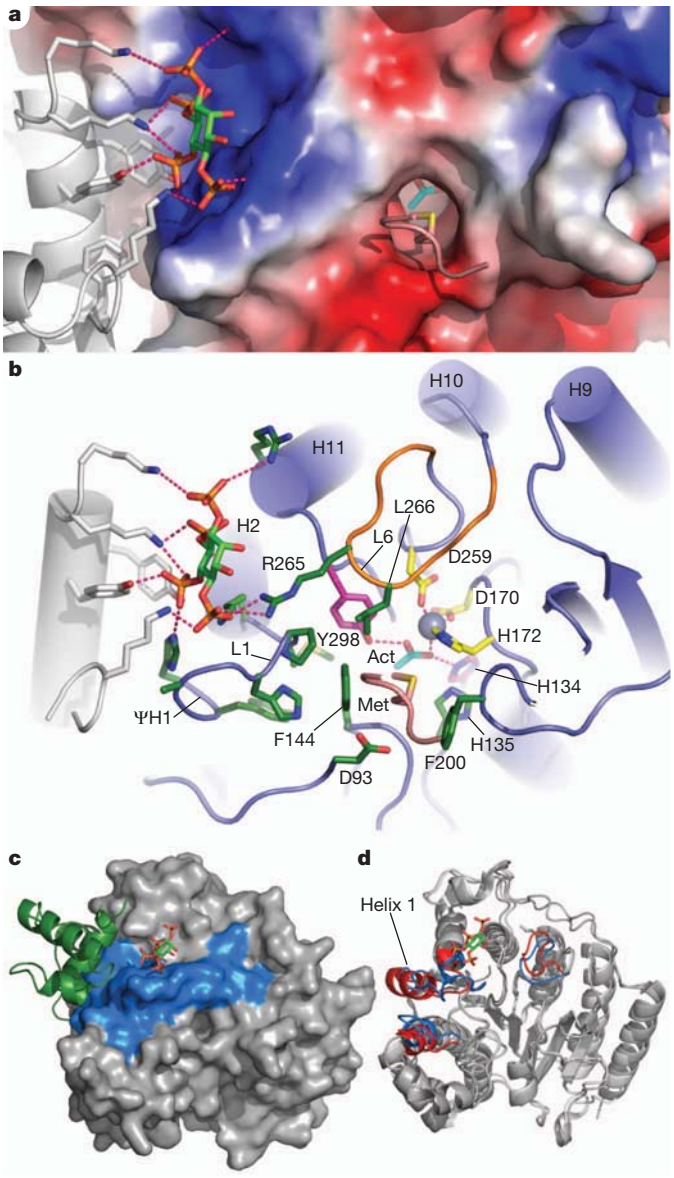


Figure 4 | Mechanism of activation of HDAC3 by binding SMRT-DAD and Ins(1,4,5,6)P_4. **a**, SMRT-DAD (grey cartoon) and Ins(1,4,5,6)P_4 bind adjacent to the HDAC3 (charged surface representation) active site. Acetate and a methionine (lysine mimic) are located in the active site. **b**, Details of the HDAC3 active site. Key residues and loops are labelled—see text for details (hydrogen bonds and/or salt bridges are indicated by dashed lines). ΨH1, pseudo helix H1. **c**, Pseudo helix H1 and loops L1 and L6 are shown in blue on the surface of HDAC3 (grey). These regions are influenced or stabilized by binding of SMRT-DAD (green) and Ins(1,4,5,6)P_4 (shown as sticks). **d**, Comparison of the structures of HDAC3 and HDAC8. Regions of significant difference are coloured blue (HDAC3) and red (HDAC8). Helix 1 is indicated.

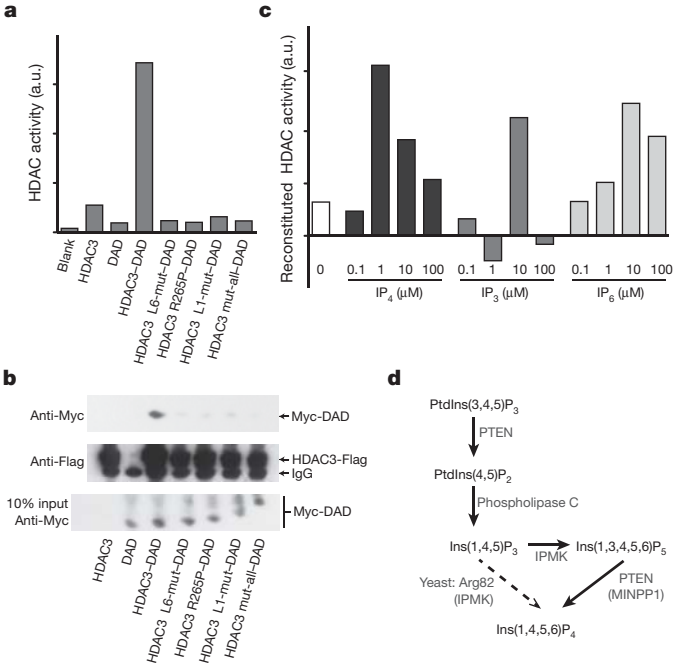


Figure 5 | Exploring the role of Ins(1,4,5,6)P_4 in complex assembly and HDAC3 activation. **a**, Effect of HDAC3 mutations on deacetylase activity of complexes with SMRT-DAD. Residues are mutated to their equivalents in HDAC8. Loop 1 mutation (L1-mut) is H17C/G21A/K25I. Loop 6 mutation (L6-mut) is R264P/L265M. 'Mut-all' is H17C/G21A/K25I/R264P/L265M/R301A. **b**, Immunoblots showing that HDAC3 mutations perturb interaction with SMRT-DAD. Recombinant proteins (Myc-DAD, HDAC3-Flag) and the IgG heavy chain are indicated. **c**, Deacetylase assays of HDAC3-Flag reconstituted with bacterially expressed SMRT-DAD in the presence of various inositol phosphates. IP4, IP3 and IP6 are Ins(1,4,5,6)P_4, Ins(1,4,5)P_3 and Ins(1,2,3,4,5,6)P_6, respectively. **d**, Synthesis pathway for Ins(1,4,5,6)P_4 from phosphatidylinositol-(3,4,5)-trisphosphate (PtdIns(3,4,5)P_3) and phosphatidylinositol-(4,5)-bisphosphate (PtdIns(4,5)P_2). Yeast Arg 82 converts Ins(1,4,5)P_3 to Ins(1,4,5,6)P_4. In mammals, both IPMK and PTEN are required to make Ins(1,4,5,6)P_4.

and purified HDAC3, Flag-tagged at the carboxy terminus, from mammalian cells and incubated the anti-Flag resin with bacterially expressed SMRT-DAD and various concentrations of Ins(1,4,5,6)P_4, Ins(1,4,5)P_3 and Ins(1,2,3,4,5,6)P_6 (Fig. 5c). HDAC activity was very sensitive to inositol phosphate concentration and higher concentrations do not support reconstitution. This can be explained through high levels of free inositol phosphate competing with complex formation. This is analogous to moderate monovalent salt concentrations causing complex dissociation during purification (Supplementary Fig. 2).

A similar pattern of activation was seen using both Ins(1,4,5)P_3 and Ins(1,2,3,4,5,6)P_6, but in both cases a tenfold higher concentration was required to achieve a similar level of HDAC activity to that seen with Ins(1,4,5,6)P_4. This supports the interpretation that Ins(1,4,5,6)P_4 is the physiologically relevant assembly partner.

Does Ins(1,4,5,6)P_4 regulate HDACs?

The finding that Ins(1,4,5,6)P_4 is essential for the assembly of class 1 HDAC repression complexes raises the question as to whether it is a signalling molecule with a role in regulating complex assembly, or if it is simply an essential structural cofactor.

The fact that Ins(1,4,5,6)P_4 remained bound during purification would seem to suggest that it is a structural cofactor. However, to retain an intact complex it was necessary to use low ionic strength buffers. At physiological ionic strengths, the Ins(1,4,5,6)P_4 is likely to dissociate readily and thus could regulate complex assembly.

Importantly, several studies have previously implicated Ins(1,4,5,6)P_4 as a regulator of gene expression and chromatin remodelling—although no molecular mechanism was identified. Yeast Arg 82 is a transcriptional regulator of genes involved in arginine metabolism. It has been shown[36] that Arg 82 is an inositol phosphate kinase that converts Ins(1,4,5)P_3 to Ins(1,4,5,6)P_4 and that kinase activity is required for its role in transcriptional regulation (Fig. 5d). Several further studies have supported the importance of Arg 82 kinase activity in transcriptional regulation and chromatin remodelling[37–39], together suggesting that Ins(1,4,5,6)P_4 has an important regulatory role in yeast.

IPMK, the mammalian homologue of Arg 82, has been reported to phosphorylate Ins(1,4,5)P_3 to form Ins(1,3,4,5,6)P_4 and then

$Ins(1,3,4,5,6)P_5$ (refs 40, 41). Consequently, in mammalian cells, a phosphatase is required to generate $Ins(1,4,5,6)P_4$. The most likely enzyme is the well-known tumour suppressor gene PTEN[42,43], which is known to be active in the nucleus and to play a role in chromosome stability (reviewed in ref. 44). It is tempting to speculate that loss of HDAC complex function might be one of the routes through which PTEN mutations contribute to oncogenesis.

A final important question is whether the levels of $Ins(1,4,5,6)P_4$ are regulated in the cell. This is not easy to establish, as regulation could be temporally and/or spatially compartmentalized. However there is some evidence that both PTEN and $InsP_4$ levels change with progression through the cell cycle[45]. It is also possible that changing levels of $Ins(1,4,5,6)P_4$ may contribute to the circadian regulation of HDAC3 activity[46]. Further research will be needed to clarify these important questions.

Discussion

We present the structure of a histone deacetylase in complex with its activating co-repressor. The structure reveals a striking structural rearrangement of the co-repressor and an unexpected inositol tetraphosphate molecule ($Ins(1,4,5,6)P_4$) acting as an 'intermolecular glue', contributing to the stabilization and activation of HDAC3.

Sequence conservation suggests that $Ins(1,4,5,6)P_4$ plays a key role in co-repressor assembly and activation of class I HDACs from yeast to humans and is likely to act as a regulator of HDAC complex assembly. Although it remains to be established how the availability of $Ins(1,4,5,6)P_4$ might be controlled, the requirement of $Ins(1,4,5,6)P_4$ for co-repressor HDAC assembly presents novel opportunities for therapeutic intervention that may complement existing HDAC inhibitors. It may be possible to develop molecules that target the $Ins(1,4,5,6)P_4$ binding site itself, but it may also be possible to target the enzymes responsible for $Ins(1,4,5,6)P_4$ synthesis.

METHODS SUMMARY

HDAC3 and Flag-tagged SMRT-DAD (residues 389–480) were expressed using transient transfection in suspension grown HEK293F cells. The HDAC3–SMRT-DAD complex was purified by Flag affinity chromatography, followed by TEV protease cleavage and size exclusion chromatography. Crystals of the HDAC3–DAD complex were grown by sitting drop vapour diffusion with a 10% propan-2-ol precipitant. The structure was determined by molecular replacement based on HDAC8 (PDB code 3EW8). Ligands, including $Ins(1,4,5,6)P_4$, zinc, potassium, acetate and glycerol were added during the refinement process.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 July; accepted 23 November 2011.
Published online 9 January 2012.

1. Pogo, B. G., Allfrey, V. G. & Mirsky, A. E. RNA synthesis and histone acetylation during the course of gene activation in lymphocytes. *Proc. Natl Acad. Sci. USA* **55,** 805–812 (1966).
2. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* **12,** 599–606 (1998).
3. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325,** 834–840 (2009).
4. Hildmann, C., Riester, D. & Schwienhorst, A. Histone deacetylases — an important class of cellular regulators with a variety of functions. *Appl. Microbiol. Biotechnol.* **75,** 487–497 (2007).
5. Hu, E. *et al.* Cloning and characterization of a novel human class I histone deacetylase that functions as a transcription repressor. *J. Biol. Chem.* **275,** 15254–15264 (2000).
6. Lee, H., Rezai-Zadeh, N. & Seto, E. Negative regulation of histone deacetylase 8 activity by cyclic AMP-dependent protein kinase A. *Mol. Cell. Biol.* **24,** 765–773 (2004).
7. Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13,** 1924–1935 (1999).
8. Li, J. *et al.* Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. *EMBO J.* **19,** 4342–4350 (2000).
9. Lechner, T. *et al.* Sds3 (suppressor of defective silencing 3) is an integral component of the yeast Sin3·Rpd3 histone deacetylase complex and is required for histone deacetylase activity. *J. Biol. Chem.* **275,** 40961–40966 (2000).
10. Zhang, J., Kalkum, M., Chait, B. T. & Roeder, R. G. The N-CoR-HDAC3 nuclear receptor corepressor complex inhibits the JNK pathway through the integral subunit GPS2. *Mol. Cell* **9,** 611–623 (2002).
11. Guenther, M. G., Barak, O. & Lazar, M. A. The SMRT and N-CoR corepressors are activating cofactors for histone deacetylase 3. *Mol. Cell. Biol.* **21,** 6091–6101 (2001).
12. Wen, Y. D. *et al.* The histone deacetylase-3 complex contains nuclear receptor corepressors. *Proc. Natl Acad. Sci. USA* **97,** 7202–7207 (2000).
13. Métivier, R. *et al.* Estrogen receptor-α directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **115,** 751–763 (2003).
14. Hoberg, J. E., Yeung, F. & Mayo, M. W. SMRT derepression by the IκB kinase α: a prerequisite to NF-κB transcription and survival. *Mol. Cell* **16,** 245–255 (2004).
15. Billin, A. N., Thirlwell, H. & Ayer, D. E. β-catenin-histone deacetylase interactions regulate the transition of LEF1 from a transcriptional repressor to an activator. *Mol. Cell. Biol.* **20,** 6882–6890 (2000).
16. Guan, H.-P., Ishizuka, T., Chui, P. C., Lehrke, M. & Lazar, M. A. Corepressors selectively control the transcriptional activity of PPARγ in adipocytes. *Genes Dev.* **19,** 453–461 (2005).
17. Marks, P. A. & Breslow, R. Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. *Nature Biotechnol.* **25,** 84–90 (2007).
18. Wagner, J. M., Hackanson, B., Lübbert, M. & Jung, M. Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. *Clin. Epigenet.* **1,** 117–136 (2010).
19. Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S. & Reinberg, D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell* **95,** 279–289 (1998).
20. Hakimi, M.-A. *et al.* A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes. *Proc. Natl Acad. Sci. USA* **99,** 7420–7425 (2002).
21. Humphrey, G. W. *et al.* Stable histone deacetylase complexes distinguished by the presence of SANT domain proteins CoREST/kiaa0071 and Mta-L1. *J. Biol. Chem.* **276,** 6817–6824 (2001).
22. Laherty, C. D. *et al.* Histone deacetylases associated with the mSin3 corepressor mediate mad transcriptional repression. *Cell* **89,** 349–356 (1997).
23. Heinzel, T. *et al.* A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature* **387,** 43–48 (1997).
24. Guenther, M. G. *et al.* A core SMRT corepressor complex containing HDAC3 and TBL1, a WD40-repeat protein linked to deafness. *Genes Dev.* **14,** 1048–1057 (2000).
25. Yoon, H.-G. *et al.* Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1. *EMBO J.* **22,** 1336–1346 (2003).
26. Oberoi, J. *et al.* Structural basis for the assembly of the SMRT/NCoR core transcriptional repression machinery. *Nature Struct. Mol. Biol.* **18,** 177–184 (2011).
27. You, S.-H., Liao, X., Weiss, R. E. & Lazar, M. A. The interaction between nuclear receptor corepressor and histone deacetylase 3 regulates both positive and negative thyroid hormone action *in vivo*. *Mol. Endocrinol.* **24,** 1359–1367 (2010).
28. Ishizuka, T. & Lazar, M. A. The nuclear receptor corepressor deacetylase activating domain is essential for repression by thyroid hormone receptor. *Mol. Endocrinol.* **19,** 1443–1451 (2005).
29. Yin, L. *et al.* Rev-erbα, a heme sensor that coordinates metabolic and circadian pathways. *Science* **318,** 1786–1789 (2007).
30. Guenther, M. G. Assembly of the SMRT-histone deacetylase 3 repression complex requires the TCP-1 ring complex. *Genes Dev.* **16,** 3130–3135 (2002).
31. Codina, A. *et al.* Structural insights into the interaction and activation of histone deacetylase 3 by nuclear receptor corepressors. *Proc. Natl Acad. Sci. USA* **102,** 6009–6014 (2005).
32. Yang, W.-M., Tsai, S.-C., Wen, Y.-D., Fejer, G. & Seto, E. Functional domains of histone deacetylase-3. *J. Biol. Chem.* **277,** 9447–9454 (2002).
33. Dowling, D. P., Gantt, S. L., Gattis, S. G., Fierke, C. A. & Christianson, D. W. Structural studies of human histone deacetylase 8 and its site-specific variants complexed with substrate and inhibitors. *Biochemistry* **47,** 13554–13563 (2008).
34. Somoza, J. R. *et al.* Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **12,** 1325–1334 (2004).
35. Bressi, J. C. *et al.* Exploration of the HDAC2 foot pocket: synthesis and SAR of substituted N-(2-aminophenyl)benzamides. *Bioorg. Med. Chem. Lett.* **20,** 3142–3145 (2010).
36. Odom, A. R. A role for nuclear inositol 1,4,5-trisphosphate kinase in transcriptional control. *Science* **287,** 2026–2029 (2000).
37. Steger, D. J. Regulation of chromatin remodeling by inositol polyphosphates. *Science* **299,** 114–116 (2003).
38. El Alami, M., Messenguy, F., Scherens, B. & Dubois, E. Arg82p is a bifunctional protein whose inositol polyphosphate kinase activity is essential for nitrogen and PHO gene expression but not for Mcm1p chaperoning in yeast. *Mol. Microbiol.* **49,** 457–468 (2003).
39. Shen, X., Xiao, H., Ranallo, R., Wu, W.-H. & Wu, C. Modulation of ATP-dependent chromatin-remodeling complexes by inositol polyphosphates. *Science* **299,** 112–114 (2003).
40. Saiardi, A. *et al.* Mammalian inositol polyphosphate multikinase synthesizes inositol 1,4,5-trisphosphate and an inositol pyrophosphate. *Proc. Natl Acad. Sci. USA* **98,** 2306–2311 (2001).

41. Nalaskowski, M. M., Deschermeier, C., Fanick, W. & Mayr, G. W. The human homologue of yeast ArgRIII protein is an inositol phosphate multikinase with predominantly nuclear localization. *Biochem. J.* **366,** 549–556 (2002).
42. Craxton, A., Caffrey, J. J., Burkhart, W., Safrany, S. T. & Shears, S. B. Molecular cloning and expression of a rat hepatic multiple inositol polyphosphate phosphatase. *Biochem. J.* **328,** 75–81 (1997).
43. Caffrey, J. J., Darden, T., Wenk, M. R. & Shears, S. B. Expanding coincident signaling by PTEN through its inositol 1,3,4,5,6-pentakisphosphate 3-phosphatase activity. *FEBS Lett.* **499,** 6–10 (2001).
44. Di Cristofano, A. & Pandolfi, P. P. The multiple roles of PTEN in tumor suppression. *Cell* **100,** 387–390 (2000).
45. Mattingly, R. R., Stephens, L. R., Irvine, R. F. & Garrison, J. C. Effects of transformation with the v-src oncogene on inositol phosphate metabolism in rat-1 fibroblasts. D-myo-inositol 1,4,5,6-tetrakisphosphate is increased in v-src-transformed rat-1 fibroblasts and can be synthesized from D-myo-inositol 1,3,4-trisphosphate in cytosolic extracts. *J. Biol. Chem.* **266,** 15144–15153 (1991).
46. Feng, D. *et al.* A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism. *Science* **331,** 1315–1319 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Owen and the other beamline staff at DIAMOND I24 for help with data collection; J. Goodchild for help with biochemical experiments; and S. Cowley, M. Lazar, P. Moody, L. Nagy, P. Tontonoz and S. Shears for literature and discussions. This work was supported by the Wellcome Trust (grant WT085408).

Author Contributions P.J.W. expressed, purified and crystallized the protein and performed the biochemical studies. P.J.W., L.F. and J.W.R.S. performed the structural determination and wrote the paper. G.M.S. and L.F. performed early expression/purification trials in insect cells. J.W.R.S. conceived the study.

Author Information Atomic coordinates and structure factors are deposited in the Protein Data Bank under accession number 4A69. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.W.R.S. (john.schwabe@le.ac.uk).

METHODS

Protein expression, purification and crystallization. The DAD domain (SMRT 389–480) and full length HDAC3 were cloned into pcDNA3 vector. The DAD domain construct contained an N-terminal 10×His-3×Flag tag and a TEV protease cleavage site. HEK293F cells (Invitrogen) were co-transfected with both constructs using polyethylenimine (PEI) (Sigma). To transfect cells, 0.25 mg DNA total was diluted in 25 ml of PBS (Sigma) and vortexed briefly; 1 ml of 0.5 mg ml^{-1} PEI was added, and the suspension was vortexed briefly, incubated for 20 min at room temperature, then added to the cells. (Final density was 1×10^6 cells per ml in a volume of 250 ml.) For larger volumes, multiple flasks were used. Cells were harvested 48 h after transfection and lysed by sonication in buffer containing 50 mM Tris pH 7.5, 100 mM potassium acetate, 5% v/v glycerol, 0.3% v/v Triton X-100, and Roche complete protease inhibitor (buffer A); the insoluble material was removed by centrifugation. The lysate was pre-cleared using Sepharose 4B (Sigma) and the complex was then bound to Flag resin (Sigma), washed three times with buffer A, three times with buffer B (50 mM Tris pH 7.5, 300 mM potassium acetate, 5% v/v glycerol) and three times with buffer C (50 mM Tris pH 7.5, 50 mM potassium acetate, 5% v/v glycerol, 0.5 mM TCEP). The complex was eluted from the resin by overnight cleavage at 4 °C with TEV protease in buffer C. The eluted protein was further purified by gel filtration on a Superdex S200 column (GE Healthcare) in buffer containing 25 mM Tris/HCl pH 7.5, 50 mM potassium acetate, 0.5 mM TCEP. The purified complex was concentrated to 7.5 mg ml^{-1} for crystallization trials.

Crystals were grown by sitting drop vapour diffusion at 4 °C using 0.1 M HEPES pH 7.5, 0.2 M NaCl and 10% v/v propan-2-ol. Crystals were cubic in nature, grew to a final dimension of $15 \times 15 \times 15$ μm and belong to space group $C222_1$.

Comparison of a fresh protein sample with protein from within the crystallization drops (after 3 months) by SDS–PAGE showed the presence of a truncated form of HDAC3 (Supplementary Fig. 2). Analysis by liquid chromatography-tandem mass spectrometry showed that this HDAC3 was truncated at the C terminus to residue Q376.

Structure determination. Crystals were flash-frozen in mother liquor containing 40% glycerol as a cryoprotectant. Diffraction data were collected on a single crystal in two 45° wedges at the Diamond synchrotron microfocus beamline I24 and processed using XDS[47]. The structure was solved by molecular replacement using HDAC8 (PDB code 3EW8)[33] as a search model in Phaser[48]. Initial model building was performed with ARP/wARP (http://www.embl-hamburg.de/ARP/), which was able to automatically build 95% of the HDAC3 protein chain, and two helices from the DAD. The additional HDAC3 and DAD sequences were fitted following multiple rounds of refinement and building using REFMAC and Coot[49,50]. $F_o - F_c$ density consistent with the ligand, zinc/potassium ions and acetate/glycerol molecules observed during the refinement/model building process were fitted and refined as they became apparent. The final model contains amino acids 2–370 from chain A and 2–370 from chain B of HDAC3, and amino acids 408–476 from chain C and 408–475 from chain D of the DAD. The model also contains two Ins(1,4,5,6)P_4 molecules, two zinc ions, four potassium ions, two acetate molecules and four glycerol molecules. The final model has 97.8% residues in the favoured region, 2.0% in the allowed region and 0.2% in the outlier region of the Ramachandran plot.

HDAC activity assays. C-terminally Flag-tagged HDAC3 and Myc-tagged DAD were co-expressed in HEK 293 cells as described above. Cells were lysed in 50 mM Tris pH 7.5, 50 mM potassium acetate, 5% v/v glycerol, 0.3% v/v Triton X-100 and Roche complete protease inhibitor. In order to standardise the assay, 800 μg total protein was bound to 40 μl Flag resin (Sigma) for 2 h at 4 °C, then washed four times with lysis buffer. HDAC activity was measured using the HDAC Assay Kit (Active Motif) and read using a Victor X5 plate reader (Perkin Elmer).

Reconstitution assays. C-terminally Flag-tagged HDAC3 was expressed in HEK 293 cells and purified as described above. His_6-tagged SMRT-DAD was expressed in *Escherichia coli* strain Rosetta (Novagen) and initial purification was carried out using Nickel NTA agarose (Qiagen) followed by gel filtration chromatography using a Superdex-S200 26/60 column (GE Healthcare) in buffer containing 50 mM Tris/HCl pH 7.5 and 50 mM potassium acetate. 40 μl of HDAC3-Flag resin was mixed with His-DAD, with or without phosphoinositide as required, in a final volume of 1 ml of buffer D (50 mM Tris/HCl pH 7.5, 50 mM potassium acetate, 5% glycerol, 0.3% v/v Triton X-100), and incubated for 2 h at 30 °C. The resin was then washed extensively with buffer D, and HDAC activity was measured using the HDAC Assay Kit (Active Motif) and read on a Victor X5 plate reader (Perkin Elmer).

47. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
48. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
49. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).

LETTER

doi:10.1038/nature10669

Gravitational detection of a low-mass dark satellite galaxy at cosmological distance

S. Vegetti[1], D. J. Lagattuta[2], J. P. McKean[3], M. W. Auger[4], C. D. Fassnacht[2] & L. V. E. Koopmans[5]

The mass function of dwarf satellite galaxies that are observed around Local Group galaxies differs substantially from simulations[1–5] based on cold dark matter: the simulations predict many more dwarf galaxies than are seen. The Local Group, however, may be anomalous in this regard[6,7]. A massive dark satellite in an early-type lens galaxy at a redshift of 0.222 was recently found[8] using a method based on gravitational lensing[9,10], suggesting that the mass fraction contained in substructure could be higher than is predicted from simulations. The lack of very low-mass detections, however, prohibited any constraint on their mass function. Here we report the presence of a $(1.9 \pm 0.1) \times 10^8 M_{\odot}$ dark satellite galaxy in the Einstein ring system JVAS B1938+666 (ref. 11) at a redshift of 0.881, where $M_{\odot}$ denotes the solar mass. This satellite galaxy has a mass similar to that of the Sagittarius[12] galaxy, which is a satellite of the Milky Way. We determine the logarithmic slope of the mass function for substructure beyond the local Universe to be $1.1^{+0.6}_{-0.4}$, with an average mass fraction of $3.3^{+3.6}_{-1.8}$ per cent, by combining data on both of these recently discovered galaxies. Our results are consistent with the predictions from cold dark matter simulations[13–15] at the 95 per cent confidence level, and therefore agree with the view that galaxies formed hierarchically in a Universe composed of cold dark matter.

The gravitational lens system JVAS B1938+666 (ref. 11) has a bright infrared background galaxy at redshift $z = 2.059$ (ref. 16), which is gravitationally lensed into an almost complete Einstein ring of diameter ~0.9 arcsec by a massive elliptical galaxy at redshift $z = 0.881$ (ref. 17). This bright, highly magnified Einstein ring makes this system an excellent candidate in which to search for surface brightness anomalies caused by very low-mass (dark matter) substructure in the halo around the high-redshift elliptical lens galaxy. The presence of a low-mass substructure (for example a luminous or dark satellite galaxy) in the lens galaxy can introduce a localized perturbation of the arc surface brightness distribution. Owing to the multiplicity of the gravitationally lensed images that form these arcs, these surface brightness 'anomalies' can be analysed using a pixelated-lens modelling technique and used to gravitationally detect and quantify the total mass and position of the substructure, down to masses as low as ~0.1% of the mass of the lens inside the Einstein radius[9,10]. The lens system was imaged at 1.6 and 2.2 μm using the Near Infrared Camera 2 on the W. M. Keck 10-m telescope in June 2010. The adaptive optics system was used to correct the incoming wavefront for the blurring induced by the atmosphere, providing a nearly diffraction-limited point spread function with a full-width at half-maximum of ~70 mas. Further details of the data sets and their image processing can be found in Supplementary Information.

A smooth parametric model for the lens potential was constrained by using the surface brightness emission from the Einstein ring of each data set independently, having first removed the smooth light contribution from the lensing galaxy. We represented the mass model using an ellipsoidal power law, $\rho(r) \propto r^{-\gamma}$, where $\rho(r)$ is the combined luminous and dark matter density as a function of the ellipsoidal radius, r. The best-fitting model was then fixed and further refined using local potential corrections defined on a regular grid, which are translated into surface density corrections using the Laplace operator. We found for both the 1.6- and the 2.2-μm adaptive optics data sets that there was a significant positive density correction, which indicated the presence of a mass substructure (Fig. 1 and Supplementary Information). Directly from the pixelated potential correction, we measured a substructure mass of $\sim 1.7 \times 10^8 M_{\odot}$ inside a projected radius of 600 pc around the density peak.

As an independent test, we repeated the analysis of the 2.2-μm data set, which had the highest-significance positive density correction, with different models of the point spread function, different data reduction techniques, different rotations of the lensed images, different models for the lens galaxy surface brightness subtraction and different resolutions for the reconstructed source. We also analysed an independent data set taken at 1.6 μm with the Near Infrared Camera and Multi-Object Spectrograph on board NASA's Hubble Space Telescope. In total, we tested fourteen different models and three different data sets that all independently led to the detection of a positive density correction at the same spatial position, although with varying levels of significance (Supplementary Information). Differential extinction across the gravitational arc could also produce a surface brightness anomaly. However, the colour of the arc was found to be consistent around and at the location of substructure, ruling out the possibly that dust affected our results.

We used an analytic model to determine the mass and the statistical significance of the substructure in the context of a physical model[9,10]. In this analytic approach, a truncated pseudo-Jaffe model was used to parameterize the substructure mass and position, giving three extra free parameters. To obtain a good fit to the observed surface brightness distribution of the lensed background source at 2.2 μm, we found that a substructure was required at a position consistent with the positive density correction detected above. Assuming the substructure to be situated in the plane of the lens galaxy, we objectively compared the smooth and the substructure parametric models in terms of the Bayesian evidence, which is the probability of the data given the model (marginalized over all parameters). Computing the marginalized evidence involved integrating over the multidimensional parameter space within predefined priors. In particular, we assumed that the substructure was equally likely to be located at any point in the lens plane and to have a mass between $4.0 \times 10^6 M_{\odot}$ and $4.0 \times 10^9 M_{\odot}$, the mass range in which comparison with simulations is possible[13,14]. See Supplementary Information for further details on the Bayesian evidence. We found that our nominal model, with a substructure of mass $M_{\rm sub} = (1.9 \pm 0.1) \times 10^8 M_{\odot}$ located at $(0.036 \pm 0.005$ arcsec, 0.576 ± 0.007 arcsec) relative to the lensing galaxy (in sky coordinates defined positive towards the west and the north, respectively), was preferred by a factor of e^{65} over a smooth model. This would heuristically correspond to a 12σ detection of the substructure, if the posterior probability distribution function were Gaussian. This agrees

[1]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Department of Physics, University of California, Davis, California 95616, USA. [3]ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands. [4]Department of Physics, University of California, Santa Barbara, California 93106, USA. [5]Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands.

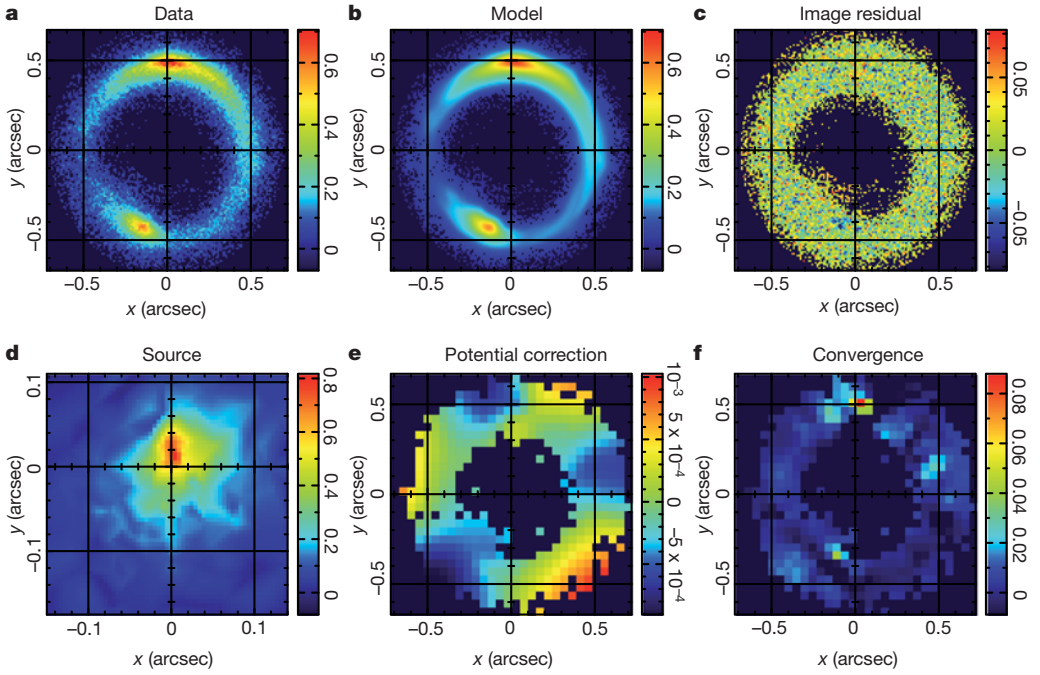


Figure 1 | Detection of a dark-matter-dominated satellite galaxy in the gravitational lens system B1938+666 at $z = 0.881$. The data shown here were measured at 2.2 μm using the W. M. Keck telescope in June 2010. Additional data sets at 1.6 μm, from the Keck telescope and the Hubble Space Telescope, are in Supplementary Information. **a**, Original data set with the lensing galaxy subtracted; **b**, final reconstruction; **c**, image residuals; **d**, source reconstruction; **e**, dimensionless potential correction from a smooth potential required by the model to fit the data; **f**, resulting dimensionless projected density corrections. Colour bars show counts for **a**–**d**. The total lensing potential is defined as the sum of an analytic potential for the host galaxy plus the local pixelated potential corrections defined on a Cartesian grid. The potential corrections are a general correction to the analytical smooth potential and correct for the presence of substructure, for large-scale moments in the density profile of the galaxy and for shear. When the Laplace operator is applied to the potential corrections and translated into surface density corrections, the terms related to the shear and mass sheets become zero and a constant, respectively. A strong positive density correction is found on the top part of the lensed arc. Note that these images are set on an arbitrary regular grid that has the origin shifted relative to the centre of the smooth lens model by $\Delta x = 0.024$ arcsec and $\Delta y = 0.089$ arcsec. When this shift is taken into account, the position of the density correction is consistent with the position of the substructure found in the analytic reconstruction (Supplementary Information).

well with the substructure mass found by the pixelated potential correction method described above, given the systematic and statistical uncertainties (Supplementary information).

We also considered a model containing two substructures. The corresponding Bayesian evidence was significantly lower than the Bayesian evidence of the smooth model and that of the single-substructure model. Above, we have quoted the statistical uncertainties for the substructure mass and position, but the uncertainties will also be related to several sources of systematic error, which are discussed in Supplementary Information. In general, the uncertainty in the substructure mass (under the assumption that it is located in the lens plane) is entirely dominated by the inference in the tidal truncation radius, which requires the substructure to be de-projected, yielding a systematic uncertainty of 0.45 dex.

On the basis of the detection of the substructure, we calculated the joint posterior probability function for the projected mass fraction of the halo that is made up of substructure[18], f, and the slope, α, of the substructure mass function (where $\mathrm{d}N \propto M^{-\alpha}\mathrm{d}M$, where N is the number density of subhaloes per comoving volume and M is the substructure mass; Supplementary Information). For the JVAS B1938+666 lensing galaxy ($M_{\mathrm{lens}} = 2.46 \times 10^{10} M_{\odot}$ within an Einstein radius of 3.39 kpc), we found that $f = 3.9^{+3.6}_{-2.4}\%$ at the 68% confidence level, using a uniform prior probability distribution for α, for a substructure mass range of $4.0 \times 10^{6} M_{\odot}$ to $4.0 \times 10^{9} M_{\odot}$. For a Gaussian prior probability distribution for α centred at 1.9 ± 0.1, which is the slope of the mass function predicted from simulations[13–15], $f = 1.5^{+1.5}_{-0.9}\%$ at the 68% confidence level. The predicted fraction of substructure from simulations within the same mass range and projected distance from the host halo centre is $\sim 0.1^{+0.3}_{-0.1}\%$ (ref. 19), which is marginally smaller than the lower limit implied by our detection of a substructure at the 68% confidence level, independent of the prior set on the slope of the mass function. However, these simulations model the formation of a Milky Way halo at $z = 0$, and simulations of elliptical galaxies out to $z \approx 1$ must be made before a more quantitative conclusion can be drawn.

Whereas flux ratio anomalies of multiply imaged quasars have previously been used to measure statistically the level of substructure in cosmologically distant lens galaxies[20–23], these analyses have degeneracies when the substructures are dark and it is difficult to localize and measure the masses of individual substructures. Hence, the shape of the substructure mass function cannot be constrained and so far it has remained unclear whether the results from quasar flux ratio anomalies are in agreement or disagreement with numerical simulations[19]. Our new low-mass detection also allows us to constrain the slope of the substructure mass function for galaxies other than our own, when combined with the detection of the ~18-fold more massive ($3.5 \times 10^{9} M_{\odot}$) substructure in the previously reported elliptical lens galaxy SDSS J0946+1006 ($M_{\mathrm{lens}} = 2.45 \times 10^{10} M_{\odot}$ within an Einstein radius of 4.60 kpc) at redshift $z = 0.222$ (ref. 8). Combining both detections resulted in a slope of $\alpha = 1.1^{+0.6}_{-0.4}$ for the mass function and an average mass fraction of $\langle f \rangle = 3.3^{+3.6}_{-1.8}\%$ for elliptical galaxies at the 68% confidence level. These results suggest that the slope of the mass function for elliptical galaxies is similar to that observed for Milky Way satellites, but that elliptical galaxies have higher mass fractions.

So far, detailed studies of the mass and luminosity properties of substructures have been limited to the Milky Way[12,24–26] and to some extent also to M31 (ref. 27), which lies at a distance of ~1 Mpc from the Milky Way. About 30 luminous satellite galaxies detected within the Milky Way virial radius (~250 kpc) are considered possible cases of cold dark matter substructure, albeit at a much lower abundance than is predicted from simulations[15]. Twenty-three of the Milky Way satellites

have been found to have masses of $\sim 10^7 M_\odot$ within a 300-pc projected radius, from observations of the dynamics of their stars[28]. This method of determining masses is limited to the Local Group owing to the faintness of the satellite galaxies, and can have a systematic error if the satellite is not in dynamic equilibrium or if there is foreground contamination. The three-dimensional mass of the JVAS B1938+666 substructure within a radius of 300 pc is $M_{300} = (7.2 \pm 0.5) \times 10^7 M_\odot$ ($M_{300} = 3.4 \times 10^7 M_\odot$ for a singular isothermal sphere model), which is comparable to the masses of the Milky Way satellites[28], but is hosted by an elliptical galaxy with a velocity dispersion of $\sigma_{\rm SIS} = 187\ {\rm km\ s^{-1}}$ at a co-moving distance of $\sim$3 Gpc, corresponding approximately to a time when the Universe was half the age it is today. A 3σ upper limit of $L_{\rm V} < 5.4 \times 10^7 L_{\rm V,\odot}$ was found for the luminosity of the substructure in the rest-frame V band within the tidal radius, $r_{\rm t} = 440 \pm 5$ pc ($L_{\rm V,\odot}$, V-band solar luminosity). This is about a factor of four brighter than both the Sagittarius and the Fornax dwarf galaxies[12]. The velocity dispersion of the satellite, based on the Einstein radius found from the best-fitting model, is $\sigma_v \approx 16\ {\rm km\ s^{-1}}$, which corresponds to a circular velocity of $v_{\rm circ} \approx 27\ {\rm km\ s^{-1}}$. Its three-dimensional mass within 600 pc is $M_{600} = (1.13 \pm 0.06) \times 10^8 M_\odot$. The observed properties (mass and circular velocity) are comparable to those of the Sagittarius dwarf galaxy, which is a satellite of the Milky Way[12].

The mass-to-light ratios of low-mass satellites of the Milky Way have also been found to disagree with the expectations from simulations. It has been predicted that satellites with the luminosities of Fornax and Sagittarius should have a velocity dispersion that is a factor of two greater than has been observed ($\sim 16\ {\rm km\ s^{-1}}$), which represents another problem for the cold dark matter paradigm[6]. In the case of the JVAS B1938+666 substructure, the 3σ upper limit on the luminosity is in the range of the Fornax and Sagittarius satellites, and the resulting mass-to-light ratio has a lower limit of $3.5 M_\odot / L_{\rm V,\odot}$. Although this result is consistent with the mass-to-light ratios predicted from simulations, the limit on the luminosity of the substructure will need to be decreased by an order of magnitude before any meaningful comparison can be made. This will only be possible with the next generation of large optical telescopes, which is planned to come into operation by the end of the decade.

Received 26 June; accepted 25 October 2011.

1. Kravtsov, A. Dark matter substructure and dwarf galactic satellites. *Adv. Astron.* **2010,** 281913 (2010).
2. Kauffmann, G., White, S. D. M. & Guiderdoni, B. The formation and evolution of galaxies within merging dark matter haloes. *Mon. Not. R. Astron. Soc.* **264,** 201–218 (1993).
3. Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. Where are the missing galactic satellites? *Astrophys. J.* **522,** 82–92 (1999).
4. Moore, B. *et al.* Dark matter substructure within galactic halos. *Astrophys. J.* **524,** L19–L22 (1999).
5. Boylan-Kolchin, M., Bullock, J. S. & Kaplinghat, M. Too big to fail? The puzzling darkness of massive Milky Way subhaloes. *Mon. Not. R. Astron. Soc.* **415,** L40–L44 (2011).
6. Boylan-Kolchin, M., Springel, V., White, S. D. M. & Jenkins, A. There's no place like home? Statistics of Milky Way-mass dark matter haloes. *Mon. Not. R. Astron. Soc.* **406,** 896–912 (2010).
7. Busha, M. T. *et al.* Statistics of satellite galaxies around Milky Way-like hosts. *Astrophys. J.* (submitted).
8. Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T. & Gavazzi, R. Detection of a dark substructure through gravitational imaging. *Mon. Not. R. Astron. Soc.* **408,** 1969–1981 (2010).
9. Koopmans, L. V. E. Gravitational imaging of cold dark matter substructures. *Mon. Not. R. Astron. Soc.* **363,** 1136–1144 (2005).
10. Vegetti, S. & Koopmans, L. V. E. Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in Galaxies. *Mon. Not. R. Astron. Soc.* **392,** 945–963 (2009).
11. King, L. J. *et al.* A complete infrared Einstein ring in the gravitational lens system B1938 + 666. *Mon. Not. R. Astron. Soc.* **295,** L41–L44 (1998).
12. Strigari, L. E. *et al.* Redefining the missing satellites problem. *Astrophys. J.* **669,** 676–683 (2007).
13. Diemand, J., Kuhlen, M. & Madau, P. Dark matter substructure and gamma-ray annihilation in the Milky Way halo. *Astrophys. J.* **657,** 262–270 (2007).
14. Diemand, J., Kuhlen, M. & Madau, P. Formation and evolution of galaxy dark matter halos and their substructure. *Astrophys. J.* **667,** 859–877 (2007).
15. Springel, V. *et al.* The Aquarius Project: the subhaloes of galactic haloes. *Mon. Not. R. Astron. Soc.* **391,** 1685–1711 (2008).
16. Riechers, D. A. Molecular gas in lensed $z > 2$ quasar host galaxies and the star formation law for galaxies with luminous active galactic nuclei. *Astrophys. J.* **730,** 108–123 (2011).
17. Tonry, J. L. & Kochanek, C. S. Redshifts of the gravitational lenses MG 1131+0456 and B1938+666. *Astrophys. J.* **119,** 1078–1082 (2000).
18. Vegetti, S. & Koopmans, L. V. E. Statistics of mass substructure from strong gravitational lensing: quantifying the mass fraction and mass function. *Mon. Not. R. Astron. Soc.* **400,** 1583–1592 (2009).
19. Xu, D. D. *et al.* Effects of dark matter substructures on gravitational lensing: results from the Aquarius simulations. *Mon. Not. R. Astron. Soc.* **398,** 1235–1253 (2009).
20. Mao, S. & Schneider, P. Evidence for substructure in lens galaxies? *Mon. Not. R. Astron. Soc.* **295,** 587–594 (1998).
21. Dalal, N. & Kochanek, C. S. Direct detection of cold dark matter substructure. *Astrophys. J.* **572,** 25–33 (2002).
22. Metcalf, R. B. & Zhao, H. Flux ratios as a probe of dark substructures in quadruple-image gravitational lenses. *Astrophys. J.* **567,** L5–L8 (2002).
23. Keeton, C. R., Gaudi, B. S. & Petters, A. O. Identifying lenses with small-scale structure. I. Cusp lenses. *Astrophys. J.* **598,** 138–161 (2003).
24. Walker, M. G. *et al.* Velocity dispersion profiles of seven dwarf spheroidal galaxies. *Astrophys. J.* **667,** L53–L56 (2007).
25. Gilmore, G. *et al.* The observed properties of dark matter on small spatial scales. *Astrophys. J.* **663,** 948–959 (2007).
26. Simon, J. D. & Geha, M. The kinematics of the ultra-faint Milky Way satellites: solving the missing satellite problem. *Astrophys. J.* **670,** 313–331 (2007).
27. Mateo, M. L. Dwarf galaxies of the Local Group. *Annu. Rev. Astron. Astrophys.* **36,** 435–506 (1998).
28. Strigari, L. E. *et al.* A common mass scale for satellite galaxies of the Milky Way. *Nature* **454,** 1096–1097 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Our results are based on observations made with the W. M. Keck Observatory and the Hubble Space Telescope. S.V. is supported by a Pappalardo Fellowship at the Massachusetts Institute of Technology, L.V.E.K. is supported (in part) through an NWO-VIDI program subsidy, and D.J.L. and C.D.F. acknowledge support from the National Science Foundation. The authors are grateful to P. Marshall for comments and feedback.

Author Contributions S.V. and L.V.E.K. developed the gravitational imaging technique used for the detection of substructure. S.V. carried out the gravitational lens modelling of the data with help from L.V.E.K. and J.P.M. S.V., L.V.E.K. and J.P.M. wrote the manuscript with comments from all of the authors. C.D.F. was the principal investigator of the observing programme and was responsible, along with D.J.L., for acquiring the data. D.J.L. and M.W.A. reduced the data and performed the galaxy subtraction with help from C.D.F. M.W.A. calculated the systematic error in the substructure mass. C.D.F. calculated the galaxy luminosity.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.V. (svegetti@space.mit.edu)

LETTER

doi:10.1038/nature10707

Coherent singlet–triplet oscillations in a silicon-based double quantum dot

B. M. Maune[1], M. G. Borselli[1], B. Huang[1], T. D. Ladd[1], P. W. Deelman[1], K. S. Holabird[1], A. A. Kiselev[1], I. Alvarado-Rodriguez[1], R. S. Ross[1], A. E. Schmitz[1], M. Sokolich[1], C. A. Watson[1], M. F. Gyure[1] & A. T. Hunter[1]

Silicon is more than the dominant material in the conventional microelectronics industry: it also has potential as a host material for emerging quantum information technologies. Standard fabrication techniques already allow the isolation of single electron spins in silicon transistor-like devices. Although this is also possible in other materials, silicon-based systems have the advantage of interacting more weakly with nuclear spins. Reducing such interactions is important for the control of spin quantum bits because nuclear fluctuations limit quantum phase coherence, as seen in recent experiments in GaAs-based quantum dots[1,2]. Advances in reducing nuclear decoherence effects by means of complex control[3–5] still result in coherence times much shorter than those seen in experiments on large ensembles of impurity-bound electrons in bulk silicon crystals[6,7]. Here we report coherent control of electron spins in two coupled quantum dots in an undoped Si/SiGe heterostructure and show that this system has a nuclei-induced dephasing time of 360 nanoseconds, which is an increase by nearly two orders of magnitude over similar measurements in GaAs-based quantum dots. The degree of phase coherence observed, combined with fast, gated electrical initialization, read-out and control, should motivate future development of silicon-based quantum information processors.

Coherent control of single, isolated quantum bits (qubits) has now been demonstrated in a large variety of physical systems[8], but not in silicon. Silicon qubits have garnered interest for over a decade in part owing to the potential to exploit techniques and infrastructure from the established microelectronics industry[9,10]. Interest in silicon is further supported by the expectation of the material's superior spin coherence times relative to other semiconductors, a result of both the reduced hyperfine coupling to nuclear spins in the semiconducting host material and the reduced number of nuclear spins. Despite the significant interest in silicon-based electrically gated qubits, progress has been hampered by several fabrication and design challenges. For example, the larger in-plane effective electron mass in silicon (which is nearly three times larger than in GaAs) shrinks the electron wavefunctions where the quantum dots are formed. Smaller wavefunctions necessitate the fabrication of correspondingly smaller devices to facilitate the isolation of a single electron in each dot. Another concern is the valley degeneracy present in bulk silicon, which is removed by an appropriate combination of strain, spatial confinement and interface properties. Considerable progress has been made in overcoming these and other obstacles, and has led to recent measurements demonstrating sufficiently large valley splitting and sufficiently long spin relaxation times for spin-qubit initialization and read-out operations[11–19].

In the present work, we form quantum dots by depleting charge from a two-dimensional electron gas within a Si/SiGe undoped heterostructure with lithographically patterned electrostatic gates (Fig. 1). We adjust the carrier concentration within the silicon quantum well using an isolated global gate above the device, and we control the charge states of the two quantum dots using the other gate electrodes. After initially depleting the double dot of all but two electrons (with one in each dot), we optimize the coupling of the quantum dots and electron-bath tunnel barriers to allow coherent dot manipulations. This and similar devices and the design advances that have facilitated coherent operation are further described in ref. 11; the work presented here used device C of that report.

As previously described[1] for a GaAs double quantum dot, the coherent electron manipulations we discuss involve operating the silicon double dot in the (1,1) charge configuration, where two electrons are separated with one in each dot. In this configuration and at a finite magnetic field, which splits off the $m = \pm 1$ spin triplet states, $(1,1)T_{\pm}$, the spin singlet state, (1,1)S, and the $m = 0$ spin triplet state, $(1,1)T_0$, may be treated as the two states of a qubit (Supplementary Information). To measure these states, we rely on spin-to-charge conversion based on Pauli spin blockade. Figure 2a shows a spin blockade signature at the (0,2)–(1,1) anticrossing obtained by cyclically pulsing the gates to transfer the double dot through the $(0,1) \rightarrow (1,1) \rightarrow (0,2)$ sequence. During the $(0,1) \rightarrow (1,1)$ transition, an electron is loaded from the bath into the left-hand dot. Both singlet and triplet states, including states formed from different valleys, are populated. Whereas the lowest-energy singlet state is able to freely transfer into (0,2), other states, including the lowest-energy triplet, are blocked by the Pauli exclusion principle[15,20–22]. This blockade is indicated by a change in the conductance of a nearby quantum point contact (QPC), which discriminates between the (1,1) and (0,2) dot charge states. The measured (0,2) signal level is used to define 100% singlet probability in subsequent experiments. The conductance change between (1,1) and (0,2), together with the known fractional time spent in the measure phase and an experimentally estimated amount of blockade relaxation during measurement, allows us directly to convert measured signal to singlet probability for all data presented here (Supplementary Information).

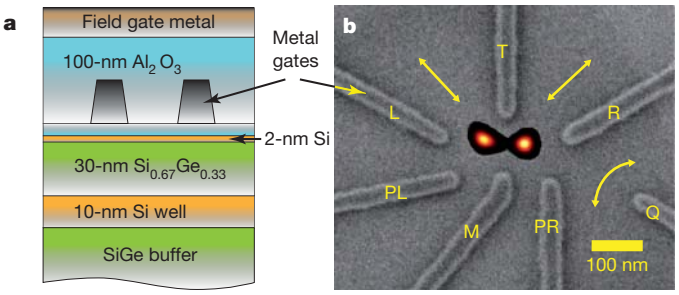


Figure 1 | Device design. **a**, Device cross-section showing undoped heterostructure, dielectric and gate stack. **b**, Scanning electron micrograph of actual device before dielectric isolation and field gate deposition. Electrostatic gates are labelled L (left), T, R (right), PL, M, PR and Q; gates L and R are used for fast pulsing. The straight arrows show current paths for transport experiments (Supplementary Information), whereas the curved arrow shows the path of current through the QPC. A numerical simulation of the electron density for the two-electron (1,1) state is superimposed on the micrograph.

[1]HRL Laboratories LLC, 3011 Malibu Canyon Road, Malibu, California 90265, USA.

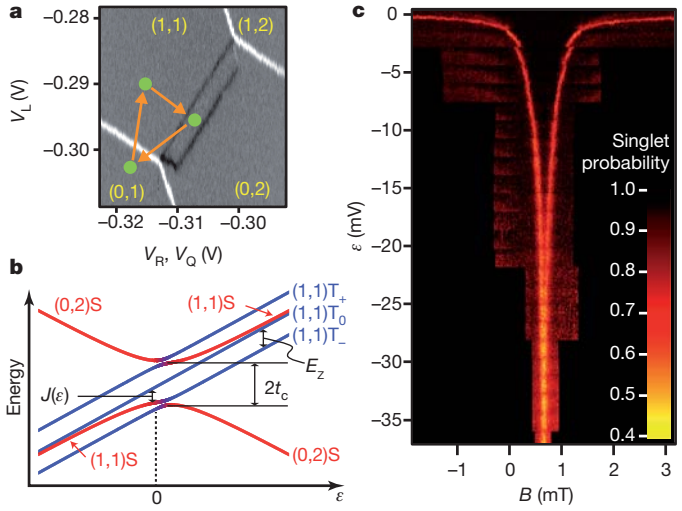


Figure 2 | Spin blockade in the double quantum dot. **a**, Subregion of the charge stability diagram, showing the differential transconductance (black and white lines) of the charge-sensing QPC as a function of gate voltages. The trapezoidal spin blockade signature is seen with cyclic pulsing through the sequence (0,1) → (1,1) → (0,2) (orange arrows). Being a differential measurement, the technique is sensitive only to changes in the average double-dot charge configuration, which is why only the edges of the trapezoidal blockade region are visible. The different colours indicate different responses of the QPC to electron movements on and between the dots. In particular, the current through the QPC decreases when electrons are added to either dot (white lines) but it increases when an electron transfers from the right dot to the left dot (black lines) (Supplementary Information). The voltage on gate Q (V_Q) was swept with that on gate R (V_R) to compensate for the capacitive coupling between gate R and the QPC and hence maintain constant sensitivity throughout the scan. **b**, Energy diagram of the (0,2)–(1,1) anticrossing, showing energies of the qubit states (1,1)S and (1,1)T_0 and the Zeeman split (1,1)$T_\pm$ states as functions of detuning, ε. The exchange-energy splitting between qubit states, $J(\varepsilon)$, the Zeeman splitting between triplet states, E_Z, and the tunnel coupling, t_c, are shown. **c**, Spin funnel obtained by measuring the degeneracy of (1,1)S and (1,1)T_- as a function of ε and B. The width of the spin funnel implies a tunnel coupling, t_c, of approximately 3 μeV. The functional form of $J(\varepsilon)$ extracted from the funnel data is shown in Fig. 5.

The spin blockade is lifted by the presence of an excited state, observed when biasing deeper into the (0,2) charge region. The excited state opens an additional unblocked pathway and results in the trapezoidal blockade region seen in Fig. 2a. The trapezoid height indicates an excited state at an energy of approximately 140 μeV. Numerical modelling suggests that this state is a (0,2) triplet involving an excited valley state of the right-hand dot, but it could alternatively be an orbital triplet if the right-hand dot is highly asymmetric[11,14]. In either case, the blockade region is sufficiently wide to distinguish the two qubit states (1,1)S and (1,1)T_0, and all other excited states are high enough in energy to have no further role in coherent operation.

The coherent evolution of the singlet and triplet qubit states is governed by two interactions. First, the singlet–triplet energy splitting, which is primarily a result of hybridization of (1,1) and (0,2) singlets, acts as an effective exchange interaction with energy J between the two spins. Second, the dot-to-dot difference in the instantaneous hyperfine coupling to ^{29}Si nuclear spins (which comprise 4.7% of the lattice nuclei in isotopically natural silicon) acts as a singlet–triplet coupling energy, Δ_{HF}. This coupling fluctuates from one experimental run to the next with a Gaussian distribution of standard deviation σ_{HF}. In a Bloch sphere picture with the singlet and triplet states at the poles, effective exchange causes a rotation about the z axis and the hyperfine energy introduces a random x component to this rotation axis. Experimentally, we have direct control over only J and are able to vary it by electrically pulsing the left- and right-hand gates to adjust the detuning, ε, between the dots (Fig. 2b). Because J and Δ_{HF} provide independent rotation axes and we only have direct control of J, exploration of the Bloch sphere requires us to achieve J values both much larger than and much smaller than σ_{HF}. For deep pulses into the (1,1) charge configuration, J is expected to be proportional to $t_c^2/|\varepsilon|$, where t_c is the interdot tunnel coupling[23]. The tunnel coupling was kept small to access J values below the small σ_{HF} in silicon.

An estimate of the tunnel coupling is provided by measuring $J(\varepsilon)$ near $\varepsilon = 0$ in a 'spin funnel' experiment[24]. An applied magnetic field, B, splits the (1,1)$T_\pm$ states from (1,1)T_0 by the Zeeman energy, $E_Z = \pm g\mu_B B$, where μ_B is the Bohr magneton and g is the effective electron g factor in silicon. The energetic crossing of (1,1)T_- with (1,1)S occurs where $J(\varepsilon) = |E_Z|$ (Fig. 2b). At this crossing, hyperfine-induced mixing of (1,1)S and (1,1)T_- occurs, which is detected by the charge sensor as a spin blockade of the (1,1)T_- state. The detuning dependence of the energetic crossing point, corresponding to the form of $J(\varepsilon)$ as a function of B, resembles a funnel whose width can be used to estimate t_c (Fig. 2c). Using the funnel, we tuned our device until we obtained an acceptable t_c value of about 3 μeV.

Controlled rotations between superpositions of the singlet and triplet states may be achieved by modifying J. Such an experiment, following ref. 1, produces data analogous to Rabi oscillations. We start with the (0,2)S ground state and then slowly separate the electrons by biasing the dots deep into the (1,1) charge configuration (Fig. 3a, b). As a result of adiabatic passage (Supplementary Information and ref. 1), a low-J ground state is prepared that is a superposition of the (1,1)S and (1,1)T_0 states owing to the hyperfine field but, because J is non-zero, is not a hyperfine eigenstate. This superposition is rotated on the Bloch sphere by the non-adiabatically applied, larger-J exchange pulse. The rotation axis depends on both J and Δ_{HF} (Fig. 3d). After the finite-duration exchange pulse, the gates are then pulsed back deep into (1,1) and adiabatically ramped into (0,2) for read-out. During this process, the low-J ground state is mapped back to the Pauli-unblocked hybridized singlet and the orthogonal low-J state is mapped to the Pauli-blocked triplet.

Figure 3c shows the resulting oscillation in the singlet probability, P_s, as a function of the exchange-pulse duration and detuning. At less negative detunings, corresponding to high values of J, the initial average visibility of the Rabi oscillations is 0.7 ± 0.1. At more negative values of ε, corresponding to lower values of J, the initial visibility is reduced and hyperfine-induced damping is observed. At intermediate detunings, as many as 20 oscillations can be resolved. The average visibility, the damping at low J values and the phase modulation of the oscillations closely follow the behaviour expected when averaging over a Gaussian distribution of hyperfine differences Δ_{HF}, as further described in Supplementary Information. Curve-fitting the data to this model allows us to extract $J(\varepsilon)$ over a wide bias range.

The practicality of implementing quantum computation using silicon-based double quantum dots depends on how long a separated (1,1)S state takes to dephase into an incoherent mixture of (1,1)S and (1,1)T_0 over multiple experimental runs. This timescale, T_2^*, directly impacts quantum control requirements, because it constrains either the speed or the complexity of high-fidelity gate operations. In III–V semiconductor quantum dots, T_2^* is reported to be a few to tens of nanoseconds[1,2,5,25]. The measurement of T_2^* is accomplished by preparing a state orthogonal to the hyperfine rotation axis and then allowing it to evolve freely in many trials. In each trial, the hyperfine rotation frequency is different, such that on averaging over those trials a dephasing-induced decay is observed. In our double-dot system, we prepare the initial singlet state by biasing the device into the (0,2)S state and quickly (within ~20 ns) pulsing the gates to a large negative detuning deep in the (1,1) charge configuration (Fig. 4a, b). The pulsing is non-adiabatic with respect to the nuclear mixing time, so the separated electrons are initially in the (1,1)S state. The subsequent state precession is a function of both the instantaneous hyperfine gradient and the remaining exchange energy. As a result, an ensemble average both oscillates and dephases into a mixture of (1,1)S and

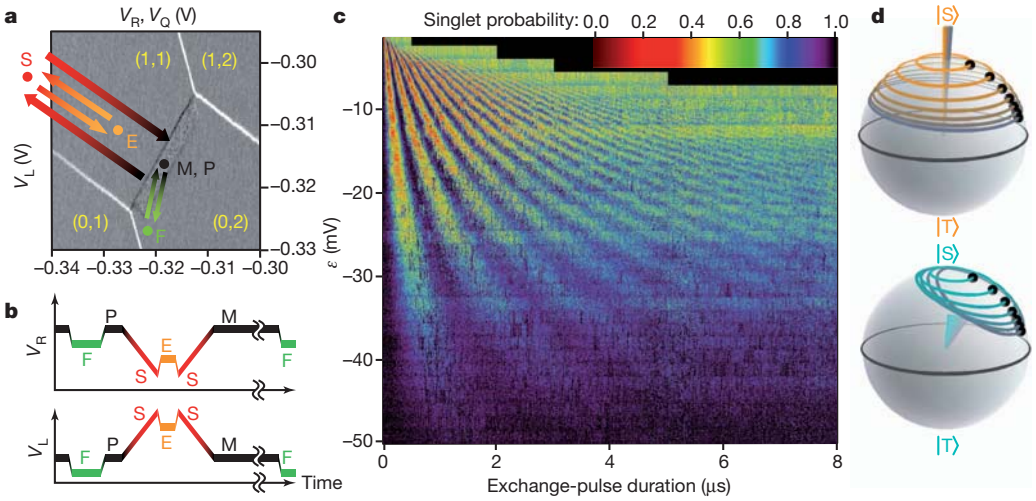


Figure 3 | Rabi oscillation pulse sequence and data. **a**, **b**, Essential elements of the pulse sequence used to obtain Rabi oscillations: V_L and V_R superimposed on the charge stability diagram (**a**); and represented as functions of time (**b**). The (0,2)S ground state is first prepared at point F with an exchange of an electron with the bath of the right-hand dot, if necessary. We then pulse to an intermediate point, P, before the subsequent adiabatic transition into the low-J ground state at point S, deep into (1,1). An exchange pulse and state rotation is then applied at E and followed by a return to S. The adiabatic transfer is then reversed and the dots are biased to point M for measurement, which generally consists of >90% of the pulse cycle. **c**, Rabi oscillations of the singlet probability, P_s, as a function of ε and exchange-pulse duration. **d**, Bloch sphere representations of state evolution for $J > \sigma_{HF}$ (top) and $J < \sigma_{HF}$ (bottom). Several instances of initial hyperfine ground states are shown as black dots (all with $\Delta_{HF} > 0$, for clarity), with corresponding rotation axes and trajectories. The data in **c** correspond to the average probability of returning to the initial prepared low-J ground state.

(1,1)T_0 (refs 23, 26). The gates are then pulsed quickly back into the (0,2) charge configuration for measurement.

Figure 4c shows the measured ensemble-averaged singlet probability as a function of separation time for different detunings, corresponding to different values of residual exchange energy, J, present during the separation pulse. For moderate negative detunings ($J > \sigma_{HF}$), the singlet state rotates primarily around the z axis and the singlet probability remains close to 100% (Fig. 4d). For large negative detunings ($J < \sigma_{HF}$), the singlet state rotates around the x axis and dephases, bringing the singlet probability to approximately 50%. Intermediate detunings result in oscillation frequencies and asymptotic probabilities between these two extremes. Fitting the data in Fig. 4c with a consistent model[23,26] for all traces provides estimates for the residual values of J and σ_{HF} (Supplementary Information). The extracted functional form of $J(\varepsilon)$ is shown in Fig. 5 and is in excellent agreement with the values extracted from the Rabi oscillation periods where the two data sets overlap. The model yields an estimate for σ_{HF} of 2.6 ± 0.2 neV. A theoretical calculation of the expected σ_{HF} value in isotopically natural silicon, using real-space, hybrid Poisson–Schrödinger/full-configuration-interaction simulation methods, and using our best estimates of the microscopic device geometry and tuning configuration, yields a value for σ_{HF} of 1.9 neV (Supplementary Information and ref. 27), in reasonable agreement with the experimentally measured value. This value corresponds to a nuclei-induced dephasing time of $T_2^* = \sqrt{2}\hbar/\sigma_{HF} = 360 \pm 30$ ns.

The reduced hyperfine coupling in silicon relative to other material systems will allow high-fidelity quantum control of future silicon-based qubits. Although the present device has only limited coherent

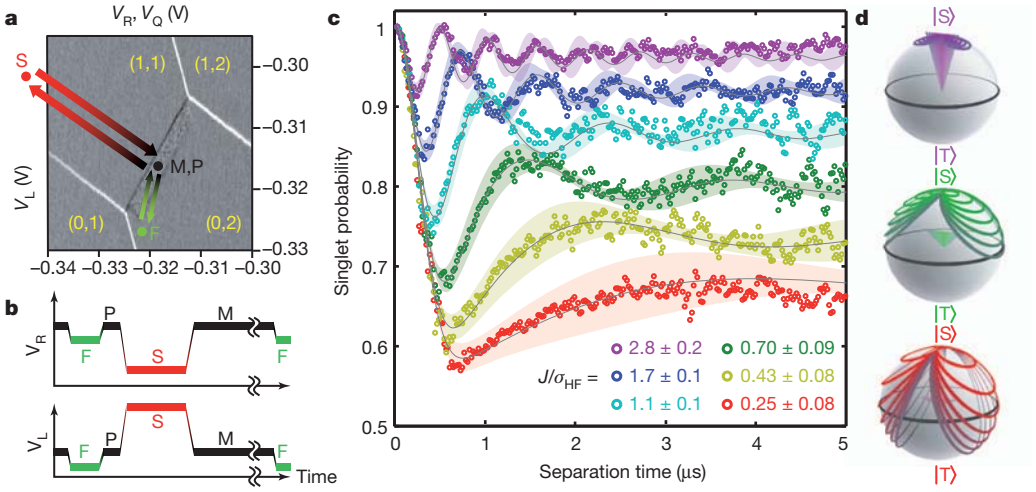


Figure 4 | T_2^* pulse sequence and data. **a**, **b**, Essential elements of T_2^* pulse sequence. Analogous with the Rabi oscillation pulse sequence (Fig. 3a, b), we initialize in the (0,2)S ground state at F and then pulse to intermediate point P. We then non-adiabatically pulse deeply into (1,1) to point S (this is the separation pulse) for a variable amount of time for singlet dephasing, and then return non-adiabatically to point M for measurement. **c**, Evolution of (1,1)S as indicated by P_s, as a function of separation time for several different separation-pulse detunings (residual J). Coloured circles show measured data; the least-squares fit (black lines) and 1-s.d. confidence intervals (coloured bands) for a hyperfine-averaged model[23,26] are superimposed. The values of J/σ_{HF} resulting from this fit are shown in the key. **d**, Bloch sphere representations of an ensemble of trajectories, with J/σ_{HF} values as indicated by colour in **c**. The ensemble of rotation axes is also shown in each case. The data in **c** correspond to the average projection of trajectories onto the z axis.

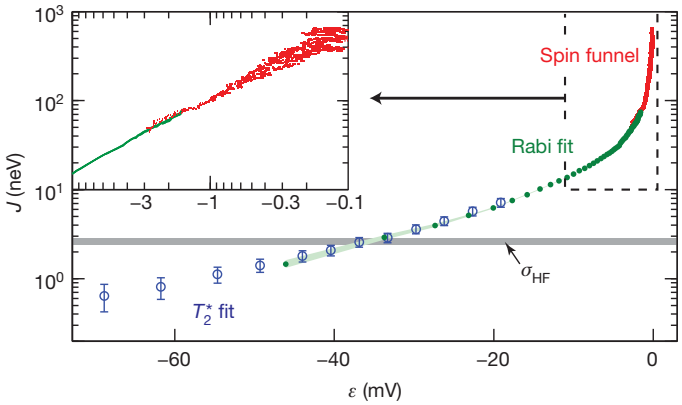


Figure 5 | Effective exchange energy versus detuning. $J(\varepsilon)$ extracted from the spin funnel data (Fig. 2c), Rabi oscillation data (Fig. 3c) and T_2* data (Fig. 4c). J ranges from 0.6 to 700 neV with excellent agreement where these data sets overlap (Supplementary Information). Error bars on the T_2* data (blue) and the confidence interval for the Rabi data (green) are 1-s.d. fit-parameter uncertainties; spin funnel symbol sizes correspond to Gaussian-fit transition widths. The grey band denotes the estimated σ_{HF} value from the T_2* data.

control capabilities owing to its reliance on the small and fluctuating hyperfine field for one axis of control, future devices will almost certainly not rely on hyperfine interactions. Other control methods, such as exchange-only control of three-dot systems[28], will allow complete coherent manipulation on a timescale much shorter than T_2*. There are no fundamental problems with delivering exchange pulses on subnanosecond timescales[1], which will further allow pulse compensation and dynamical decoupling techniques for high-fidelity gate operations[29,30]. These techniques can be used in future efforts to study and eliminate other decoherence sources not yet isolated in the experiments described here. Our results therefore indicate strong potential for the execution of the quantum error correction protocols that allow sustained quantum computation in silicon-based systems.

METHODS SUMMARY

The device is characterized in a dilution refrigerator with an estimated effective electron temperature of 150 mK. The QPC conductance changes when electrons are loaded, unloaded or transferred between the dots and is used to determine the charge state of the quantum dots. Fast voltage pulses are applied to the left- and right-hand gates using a Tektronix 5014 arbitrary wave generator with low-pass filtering at its output. Measurements are conducted with an external in-plane magnetic field of 30 mT, except for the spin funnel experiment, in which the field is swept.

The differential transconductance measurement technique used is implemented by superimposing a low-frequency voltage square wave (generally 600–900 Hz) on the left-hand gate and using standard lock-in amplification techniques with a 30-ms time constant and 60-ms delay between measurements. The low-frequency square wave is achieved by concatenating an appropriate number of the higher-frequency coherent manipulation pulse sequences to bring the lock-in measurement frequency down to near the noise floor of our measurement system. The added square wave is present only during the measurement phase of the pulse sequences. Apart from a background signal resulting from coupling between the low-frequency square wave and the QPC, the differential transconductance measurement technique is sensitive only to changes in the QPC conductance resulting from changes in the double-dot charge state configuration. As a result, we see only lines marking the outlines of the charge stability regions and of the trapezoidal triplet blockade region in (0,2), instead of the more commonly seen QPC conductance plateaux.

More details on experimental methods are available in Supplementary Information.

Received 24 June; accepted 1 November 2011.

1. Petta, J. R. *et al.* Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309,** 2180–2184 (2005).
2. Koppens, F. H. L., Nowack, K. C. & Vandersypen, L. M. K. Spin echo of a single electron spin in a quantum dot. *Phys. Rev. Lett.* **100,** 236802 (2008).
3. Reilly, D. J. *et al.* Suppressing spin qubit dephasing by nuclear state preparation. *Science* **321,** 817–821 (2008).
4. Barthel, C., Medford, J., Marcus, C. M., Hanson, M. P. & Gossard, A. C. Interlaced dynamical decoupling and coherent operation of a singlet-triplet qubit. *Phys. Rev. Lett.* **105,** 266808 (2010).
5. Bluhm, H. *et al.* Dephasing time of GaAs electron-spin qubits coupled to a nuclear bath exceeding 200 μs. *Nature Phys.* **7,** 109–113 (2011).
6. Simmons, S. *et al.* Entanglement in a solid-state spin ensemble. *Nature* **470,** 69–72 (2011).
7. Tyryshkin, A. M. *et al.* Electron spin coherence exceeding seconds in high purity silicon. *Nature Mater.* (in the press); preprint at ⟨http://arxiv.org/abs/1105.3772⟩ (2011).
8. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (2010).
9. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393,** 133–137 (1998).
10. Eriksson, M. A. *et al.* Spin-based quantum dot quantum computing in silicon. *Quantum Inf. Process.* **3,** 133–146 (2004).
11. Borselli, M. G. *et al.* Pauli spin blockade in undoped Si/SiGe two-electron double quantum dots. *Appl. Phys. Lett.* **99,** 063109 (2011).
12. Goswami, S. *et al.* Controllable valley splitting in silicon quantum devices. *Nature Phys.* **3,** 41–45 (2007).
13. Xiao, M., House, M. G. & Jiang, H. W. Parallel spin filling and energy spectroscopy in few-electron Si metal-on-semiconductor-based quantum dots. *Appl. Phys. Lett.* **97,** 032103 (2010).
14. Borselli, M. G. *et al.* Measurement of valley splitting in high-symmetry Si/SiGe quantum dots. *Appl. Phys. Lett.* **98,** 123118 (2011).
15. Lai, N. S. *et al.* Pauli spin blockade in a highly tunable silicon double quantum dot. *Sci. Rep.* **1,** 110 (2011).
16. Hayes, R. R. *et al.* Lifetime measurements (T1) of electron spins in Si/SiGe quantum dots. Preprint at ⟨http://arxiv.org/abs/0908.0173⟩ (2009).
17. Xiao, M., House, M. G. & Jiang, H. W. Measurement of the spin relaxation time of single electrons in a silicon metal-oxide-semiconductor-based quantum dot. *Phys. Rev. Lett.* **104,** 096801 (2010).
18. Morello, A. *et al.* Single-shot readout of an electron spin in silicon. *Nature* **467,** 687–691 (2010).
19. Simmons, C. B. *et al.* Tunable spin loading and T_1 of a silicon spin qubit measured by single-shot readout. *Phys. Rev. Lett.* **106,** 156804 (2011).
20. Johnson, A. C. *et al.* Triplet–singlet spin relaxation via nuclei in a double quantum dot. *Nature* **435,** 925–928 (2005).
21. Liu, H. W. *et al.* Pauli-spin-blockade transport through a silicon double quantum dot. *Phys. Rev. B* **77,** 073310 (2008).
22. Shaji, N. *et al.* Spin blockade and lifetime-enhanced transport in a few-electron Si/SiGe double quantum dot. *Nature Phys.* **4,** 540–544 (2008).
23. Laird, E. A. *et al.* Effect of exchange interaction on spin dephasing in a double quantum dot. *Phys. Rev. Lett.* **97,** 056801 (2006).
24. Petta, J. R. *et al.* Dynamic nuclear polarization with single electron spins. *Phys. Rev. Lett.* **100,** 067601 (2008).
25. Press, D. *et al.* Ultrafast optical spin echo in a single quantum dot. *Nature Photon.* **4,** 367–370 (2010).
26. Coish, W. A. & Loss, D. Singlet-triplet decoherence due to nuclear spins in a double quantum dot. *Phys. Rev. B* **72,** 125337 (2005).
27. Assali, L. V. C. *et al.* Hyperfine interactions in silicon quantum dots. *Phys. Rev. B* **83,** 165301 (2011).
28. DiVincenzo, D. P., Bacon, D., Kempe, J., Burkard, G. & Whaley, K. B. Universal quantum computation with the exchange interaction. *Nature* **408,** 339–342 (2000).
29. Tomita, Y., Merrill, J. T. & Brown, K. R. Multi-qubit compensation sequences. *N. J. Phys.* **12,** 015002 (2010).
30. West, J. R., Lidar, D. A., Fong, B. H. & Gyure, M. F. High fidelity quantum gates via dynamical decoupling. *Phys. Rev. Lett.* **105,** 230503 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. M. Marcus for discussions, J. R. Petta for assistance with measurement techniques and B. H. Fong for assistance with hyperfine calculations. Sponsored by United States Department of Defense. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the United States Department of Defense or the US Government. Approved for public release, distribution unlimited.

Author Contributions B.M.M., M.G.B., C.A.W., K.S.H. and A.T.H. contributed to device measurement and testing. B.H., P.W.D., I.A.-R., A.E.S. and M.S. contributed to material growth and device fabrication. B.M.M., T.D.L., A.A.K., R.S.R. and M.F.G. contributed to modelling and data analysis. B.M.M., T.D.L., A.A.K. and M.F.G prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.M.M. (bmmaune@hrl.com).

LETTER

doi:10.1038/nature10769

Reconfigurable self-assembly through chiral control of interfacial tension

Thomas Gibaud[1], Edward Barry[1], Mark J. Zakhary[1], Mir Henglin[1], Andrew Ward[1], Yasheng Yang[1], Cristina Berciu[2], Rudolf Oldenbourg[3], Michael F. Hagan[1], Daniela Nicastro[2], Robert B. Meyer[1] & Zvonimir Dogic[1]

From determining the optical properties of simple molecular crystals to establishing the preferred handedness in highly complex vertebrates, molecular chirality profoundly influences the structural, mechanical and optical properties of both synthetic and biological matter on macroscopic length scales[1,2]. In soft materials such as amphiphilic lipids and liquid crystals, the competition between local chiral interactions and global constraints imposed by the geometry of the self-assembled structures leads to frustration and the assembly of unique materials[3–6]. An example of particular interest is smectic liquid crystals, where the two-dimensional layered geometry cannot support twist and chirality is consequently expelled to the edges in a manner analogous to the expulsion of a magnetic field from superconductors[7–10]. Here we demonstrate a consequence of this geometric frustration that leads to a new design principle for the assembly of chiral molecules. Using a model system of colloidal membranes[11], we show that molecular chirality can control the interfacial tension, an important property of multi-component mixtures. This suggests an analogy between chiral twist, which is expelled to the edges of two-dimensional membranes, and amphiphilic surfactants, which are expelled to oil–water interfaces[12]. As with surfactants, chiral control of interfacial tension drives the formation of many polymorphic assemblages such as twisted ribbons with linear and circular topologies, starfish membranes, and double and triple helices. Tuning molecular chirality *in situ* allows dynamical control of line tension, which powers polymorphic transitions between various chiral structures. These findings outline a general strategy for the assembly of reconfigurable chiral materials that can easily be moved, stretched, attached to one another and transformed between multiple conformational states, thus allowing precise assembly and nanosculpting of highly dynamical and designable materials with complex topologies.

In experiments on chiral self-assembly, we used a two-component mixture consisting of 880-nm-long rod-like fd viruses and the non-adsorbing polymer Dextran. In aqueous suspension, fd viruses alone have purely repulsive interactions[13]. Adding non-adsorbing polymer to a dilute isotropic suspension of fd rods induces attractive interactions by the depletion mechanism and leads to their condensation into colloidal membranes, which are equilibrium structures consisting of one-rod-length-thick, liquid-like monolayers of aligned rods[11] (Fig. 1a). Despite having different structures on molecular length scales, the coarse-grained properties of colloidal membranes are identical to those of conventional lipid bilayers. However, unlike their amphiphilic counterparts, colloidal membranes do not form vesicles and are instead observed as freely suspended disks with exposed edges. Here we investigate the behaviour of these exposed edges in a manner analogous to previously studied liquid–liquid domains embedded in lipid bilayers[14–16]. For our experiments, it is essential that fd viruses are chiral, that is, a pair of aligned viruses minimizes their interaction energy when they are slightly twisted with respect to each other in a preferred direction. The strength of chiral interactions can be continuously tuned to zero through either genetic or physical methods[13,17] (Supplementary Fig. 1).

Before investigating chiral membranes, we determined the edge structure of a membrane composed of simpler, achiral rods using three complementary imaging techniques, namely two-dimensional (2D) and three-dimensional (3D) polarization microscopy and electron microscopy. The local tilting of the rods within a membrane was determined using a 2D LC-PolScope, which produces images in which the intensity of each pixel represents the local retardance of the membrane[18] (Fig. 1d). Such images can be quantitatively related to the tilting of the rods away from the layer normal[19] (the z axis). Rods in the bulk of a membrane are aligned along the z axis, and it follows that 2D LC-PolScope images appear black in that region (Fig. 1e). By contrast, the bright, birefringent ring along the membrane's periphery reveals local tilting of the rods (Fig. 1e and Supplementary Fig. 2). For achiral rods, this indicates that a membrane has a hemi-toroidal curved edge (Fig. 1b, c). By comparison with an untilted edge, a curved edge structure lowers the area of the rod–polymer interface, thus reducing interfacial tension, at the cost of increasing the elastic energy due to a twist distortion. This hypothesis is confirmed by visualizing the 3D membrane structure using electron tomography, which shows that the viruses' long-axis transitions from being parallel to the z axis in the membrane bulk to being perpendicular to the z axis and tangential to the edge along the membrane periphery (Fig. 1d and Supplementary Fig. 3). For achiral viruses, the spontaneous twist at the edges is equally likely to be clockwise or anticlockwise. However, the 2D LC-PolScope records only 2D projections of the birefringence map and thus cannot distinguish between these two possibilities. For this reason, we used a 3D LC-PolScope[20], which reveals that achiral rods are equally likely to tilt in either direction, thus confirming the spontaneously broken chiral symmetry at the edge of an achiral membrane (Fig. 1f–h and Supplementary Fig. 4). Simulations of hard, achiral spherocylinders and depletant molecules provide additional verification that interfacial effects alone cause spontaneous twisting of achiral rods at the membrane's edge (Supplementary Fig. 5).

When viewed with optical microscopy, a membrane's edge shows pronounced thermal fluctuations, the analysis of which yields the line tension, γ_{eff}, which is the energetic cost required to move rods from the membrane interior to the periphery[12,21,22] (Fig. 2a). A typical fluctuation spectrum for an achiral edge is shown in Fig. 2b. In the regime of small wavevector, q, the mean square Fourier amplitudes, $\langle a_q{}^2 \rangle$, are independent of q, allowing us to extract the effective line tension, $\gamma_{\mathrm{eff}} = k_{\mathrm{B}}T/\langle a_q{}^2 \rangle$ (ref. 12), where k_{B} is the Boltzmann constant and T is the temperature. In the large-q limit, fluctuations scale as $1/q^2$ and are therefore substantially suppressed relative to those of a simple interface. One possible explanation is that fluctuations are suppressed as a result of a quasi-one-dimensional (quasi-1D) nematic phase at the edge, which is a direct consequence of rod tilting (Fig. 1e). At large wavelengths, the elastic energy required to deform the quasi-1D nematic phase is negligible and the fluctuations are dominated by the line tension. With decreasing wavelength, there is an increasing energetic penalty associated with bending the quasi-1D nematic phase, resulting in the suppression of fluctuations.

[1]The Martin Fisher School of Physics, Brandeis University, 415 South Street, Waltham, Massachusetts 02454, USA. [2]Department of Biology, Brandeis University, 415 South Street, Waltham, Massachusetts 02454, USA. [3]Marine Biological Laboratory, 7 MBL Street, Woods Hole, Massachusetts 02543, USA.

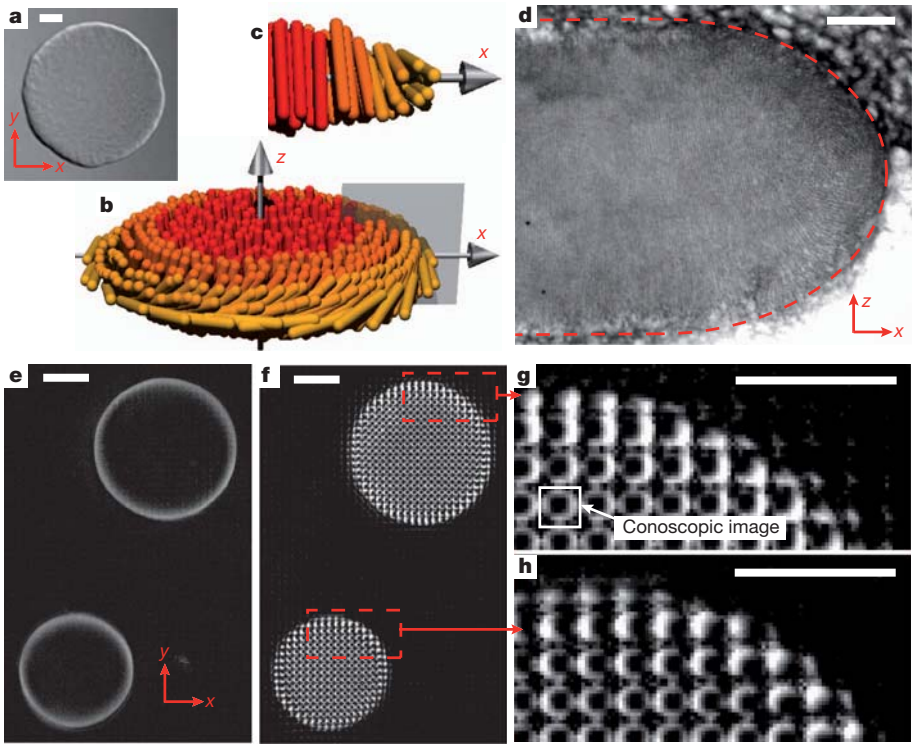


Figure 1 | Edge structure of membranes assembled from an achiral mixture of wild-type and Tyr21Met fd phages. **a**, Differential interference contrast (DIC) image of a colloidal membrane. **b**, Illustration of colloidal membrane indicating that its edge adopts a surface-tension-minimizing curved profile, forcing rods to twist locally. For clarity, the aspect ratio of rods has been reduced. **c**, Cross-section shown in **b**. **d**, Electron micrograph of a membrane directly visualizing the curved edge profile (Supplementary Fig. 3). **e**, 2D LC-PolScope birefringence map of two membranes. The bright band associated with the edges indicates local rod tilting (Supplementary Fig. 2). **f**, 3D LC-PolScope image composed of an array of conoscopic microimages, each sampling the local orientation of the rods. In each microimage, the displacement of the dark spot away from the centre determines the local tilting of the rods with respect to the layer normal (the *z* axis; Supplementary Fig. 4). **g**, **h**, Sections of 3D LC-PolScope images of two edges with clockwise (**g**) and anticlockwise (**h**) twists, respectively, illustrating the broken chiral symmetry found at the edge of achiral membranes. Scale bars, 5 μm (**a**, **e**–**g**, **h**); 0.2 μm (**d**).

Next we demonstrated that the chirality of the constituent rods controls the magnitude of γ_{eff}. The global constraints imposed by the 2D membrane topology are fundamentally incompatible with local twisting (chiral interactions) between constituent rods, leading to the expulsion of chirality from the bulk of the membrane[7] (untwisting of rods). It follows that introducing chirality raises the energy of rods in the membrane bulk while simultaneously lowering the energy of rods close to the edge, where they are free to twist and satisfy chiral interactions. Therefore, we propose that there are two contributions to the effective line tension, γ_{bare} and γ_{chiral}, such that $\gamma_{eff} = \gamma_{bare} - \gamma_{chiral}$, where γ_{bare} is the line tension of a membrane edge composed of achiral rods and γ_{chiral} is the chiral contribution to the line tension[19,23]. To measure γ_{bare} and γ_{chiral} independently, we used the temperature dependence of fd chirality (Supplementary Fig. 1). We measured γ_{bare} at high temperature (60 °C), where rods are achiral ($\gamma_{chiral} = 0$) and, therefore, $\gamma_{eff} = \gamma_{bare}$ (Fig. 2c). Decreasing the temperature increases the strength of chiral interactions, resulting in a smaller γ_{eff} value and demonstrating that chirality indeed reduces line tension (Fig. 2c). Temperature-dependent chiral contributions ($\gamma_{chiral}(T) = -\gamma_{eff}(T) + \gamma_{bare}$) were converted into a function that depends on the chiral wavevector, q_0, using measurements of the temperature dependence of the cholesteric twist (Supplementary Fig. 1). Notably, measurements of γ_{chiral} at different depletant concentrations collapse onto a universal curve (Fig. 2d), confirming that the two contributions to γ_{eff} are uncorrelated. Our data show that chirality can reduce the line tension by as much as $450 k_B T$ per micrometre of edge length (Fig. 2d). To test this hypothesis independently, we measured γ_{eff} for membranes composed of the Tyr21Met mutant fd virus. By contrast with wild-type fd virus, the chirality of this mutant fd virus is temperature independent (Supplementary Fig. 6). We find that γ_{eff} for the Tyr21Met mutant membranes is temperature independent over the entire measurement range, unambiguously demonstrating chiral control of line tension and ruling out any intrinsic temperature effects.

Our findings raise the possibility that at sufficiently low temperatures the chiral contribution to interfacial energy could dominate the bare line tension, lowering the energetic cost of creating edges and leading to

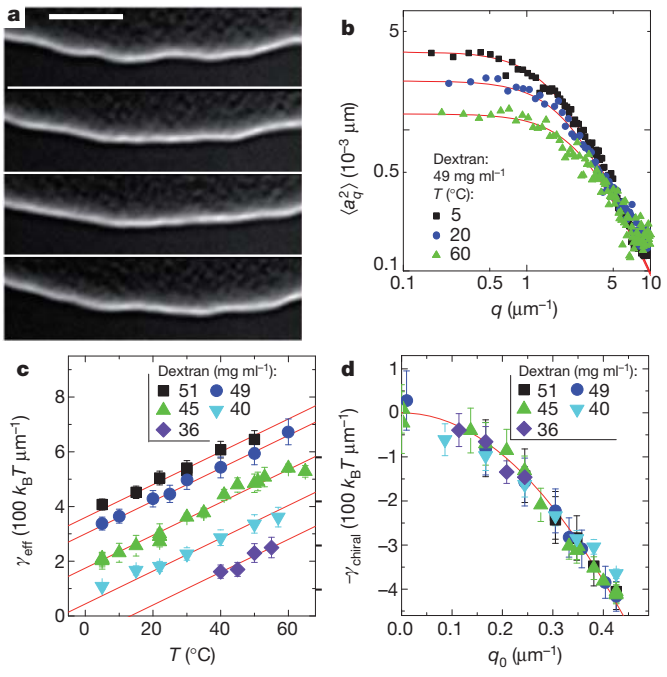


Figure 2 | Chiral control of effective line tension, γ_{eff}. **a**, DIC images, taken 1 s apart, illustrating the fluctuations of the membrane's edge. Scale bar, 5 μm. **b**, Fluctuation spectrum ($\langle a_q^2 \rangle$ versus q) plotted for three different temperatures. For small q, $\langle a_q^2 \rangle$ is independent of q and inversely proportional to the effective line tension, $\gamma_{eff} = k_B T / \langle a_q^2 \rangle$. For large q, the suppressed fluctuations are independent of depletant concentration (osmotic pressure) and sample temperature (fd chirality) and scale as $1/q^2$. Red lines are fits using equation (3) (Methods). **c**, γ_{eff} extracted from the low-q regime, plotted for a range of sample temperatures (fd chirality) and Dextran concentrations (osmotic pressures). In the achiral limit at 60 °C, $\gamma_{chiral} = 0$ and $\gamma_{bare} = \gamma_{eff}$. Increasing the Dextran concentration increases γ_{bare}. Decreasing the temperature reduces γ_{eff} because γ_{chiral} increases. The red lines of fixed slope are guides to the eyes illustrating the universal scaling of γ_{eff} with chirality. **d**, γ_{chiral} as a function of the twist wavevector, q_0 (measured in the cholesteric phase), for various depletant concentrations. The temperature dependence of q_0 is extracted from Supplementary Fig. 1. The red curve is quadratic fit to the data: $\gamma_{chiral} = C q_0^2$. Error bars, s.d.; $n = 10$.

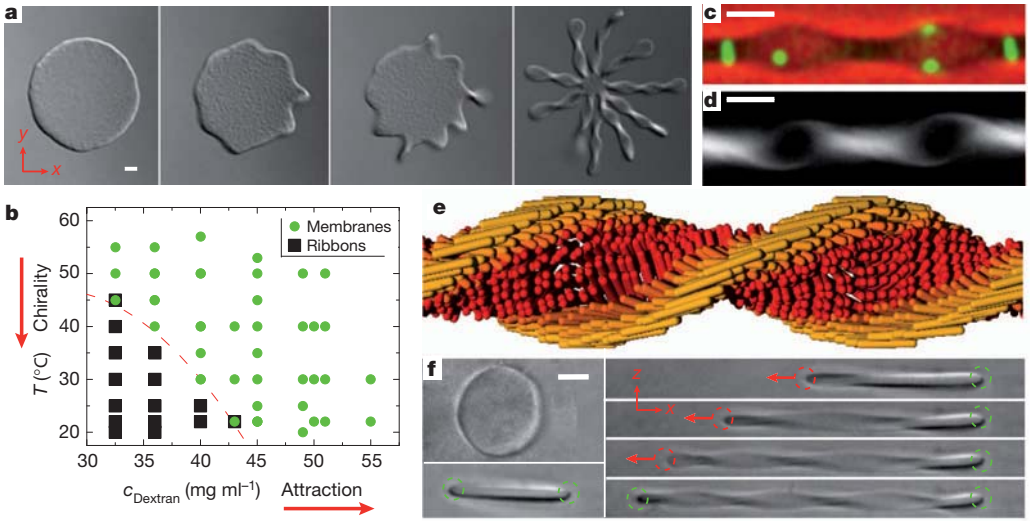


Figure 3 | Transition of a 2D disk into 1D twisted ribbons. **a**, A temperature quench reduces the line tension, inducing a transition of the 2D membrane into 1D twisted ribbons (Supplementary Movie 1). **b**, The stability diagram indicates regions of phase space where ribbons and membranes are observed as a function of temperature (chirality) and Dextran concentration ($c_{Dextran}$). The boundary indicates the temperature above which no stable or metastable ribbons are observed at a given Dextran concentration. **c**, An overlaid phase contrast (red)/fluorescence (green) image of a stable ribbon containing a low volume fraction of fluorescently labelled rods (Supplementary Movie 2). **d**, LC-PolScope image indicating rod tilting that penetrates from the edge of the ribbon towards its centre. The intensity at each pixel is proportional to $\sin^2(\theta)$, where θ is the local rod tilt angle away from the image plane. **e**, Schematic structure of twisted ribbons as deduced from optical microscopy. **f**, A 2D colloidal membrane is trapped with a dual-beam optical trap and stretched, inducing the transition to a 1D twisted ribbon (Supplementary Movie 3). Red circles indicate moving traps and green circles indicate fixed traps. Scale bars, 2 μm.

their spontaneous formation. In this respect, we expect chirality to have a role similar to that of amphiphilic surfactants in oil–water mixtures. Such surfactants lower surface tension and lead to the formation of excess interfaces through the assembly of microemulsions. To investigate this possibility, we quenched a membrane assembled at low depletant (Dextran) concentration from the high-temperature achiral phase. With decreasing temperature, fluctuations of the membrane edge become more pronounced, indicating that γ_{eff} is decreasing. Eventually the edge becomes unstable, resulting in a remarkable polymorphic transition in which twisted, ribbon-like structures grow along the entire periphery of the disk, generating a starfish-shaped membrane (Fig. 3a and Supplementary Movie 1). This polymorphic transition is reversible; with increasing temperature, the starfish membrane collapses back into a disk. The stability diagram indicates that 1D twisted ribbons are stable at low Dextran concentrations and low temperatures (small γ_{bare} values and large γ_{chiral} values), whereas flat 2D membranes are stable for large γ_{bare} values or small γ_{chiral} values (Fig. 3b). DIC microscopy reveals the overall structure of the ribbons, such as their pitch and width (Fig. 3a); fluorescence microscopy shows the liquid-like dynamics of single rods within the assemblage (Fig. 3c and Supplementary Movie 2); and 2D LC-PolScope imaging quantitatively determines the local tilt of the rods (Fig. 3d). A schematic representation of the twisted ribbons based on this information, indicating inhomogeneous tilting of the rods within a ribbon, is shown in Fig. 3e. Such data are in agreement with recent theoretical predictions[24].

Chiral self-assembly produces structures that combine seemingly divergent properties of high elasticity and fluidity. To demonstrate these properties, we trapped two opposite sides of a flat disk with a dual-beam optical trap and applied an extensional force, causing the transition from a stretched disk into a twisted ribbon (Fig. 3f and Supplementary Movie 3). This mechanically induced disk-to-ribbon transition is reversible; on removal of the optical trap, the highly elastic ribbon relaxes back into its original shape. Furthermore, using optical tweezers it is possible to change the architecture of the chiral assemblages systematically, allowing new pathways for assembling nanomaterials with highly complex topologies. For example, two ends of a linear twisted ribbon are easily joined together to assemble a closed, ring-like polymeric structure. The disk-to-ribbon transition is tightly coupled to the assemblage topology (Fig. 4d). On increasing the temperature, all linear ribbons transform into achiral disks (Supplementary Movie 4), whereas closed, ring-like ribbons remain twisted owing to the constraint imposed by the ring topology (Supplementary Movie 5).

The chiral assembly pathways described here are hierarchical, opening up the possibility that simple changes on microscopic (ångström) length scales can be used to control structures on macroscopic (millimetre) length scales[25,26]. At the smallest relevant length scale, each virus is assembled from DNA that is coated with thousands of

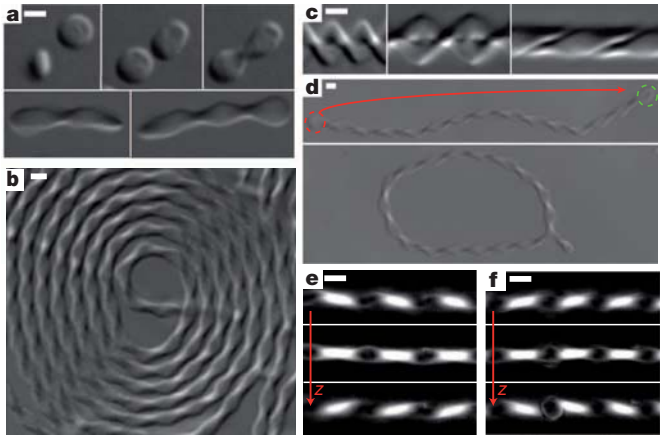


Figure 4 | Hierarchical self-assembly: from isolated viruses and metastable disks to singly and doubly twisted ribbons. **a**, In the ribbon region of the phase diagram in Fig. 3b, fd viruses condense into metastable, self-limited disks. On longer timescales, these disks coalesce and form 1D twisted ribbons. **b**, Image of a single ribbon that has collapsed into a toroid. **c**, Doubly twisted ribbons are assembled by wrapping around each other two singly twisted ribbons with a well-defined phase difference (Supplementary Fig. 8). **d**, Directed assembly of a ring-like supramolecular polymer from a linear twisted ribbon using optical tweezers. **e**, **f**, Single point mutation of the major coat protein permits microscopic control of ribbon chirality. Wild-type fd virus assembles into left-handed ribbons (**e**), whereas the Tyr21Met fd virus forms right-handed ribbons (**f**), as demonstrated by the *z*-stack series of 2D LC-PolScope images. Scale bars, 2 μm.

copies of the major coat protein. At the mesoscopic length scale, attractive depletion interactions condense dilute, rod-like viruses into self-limited, metastable, mesoscopic colloidal disks whose diameter is approximately one twist penetration length[19,23,27,28] (Fig. 4a). Once assembled, mesoscopic disks themselves experience mutual attractions, resulting in lateral coalescence and formation of twisted doublets. On longer timescales, this coalescence process continues, producing ribbons with complex topologies and multiple branching points (Supplementary Fig. 7). The length of these ribbons can reach macroscopic (millimetre) dimensions. This hierarchical assembly process describes a robust synthesis of ribbons that behave as supramolecular polymers and can assemble into even more-complex structures at higher levels of hierarchy, such as toroids (Fig. 4b). Under certain conditions, ribbons also wrap around each other to produce a great variety of double- and triple-helical structures (Fig. 4c and Supplementary Fig. 8). A challenge associated with hierarchical assembly pathways is to control the final macroscopic assemblage by specific modification of relevant microscopic parameters. Previous work showed that even a single amino-acid mutation of the major coat protein can alter the microscopic chirality of individual viruses[13]. We used this finding to demonstrate that such amino-acid exchange propagates across all hierarchical levels, switching the ribbon handedness and allowing for microscopic control of macroscopic ribbon chirality (Fig. 4e, f).

We have shown that the chirality of the constituent molecules can be used to tune the line tension associated with the exposed edge of a colloidal monolayer membrane. The potential of hierarchical chiral self-assembly is demonstrated by assembling reconfigurable materials with unique mechanical properties and complex topologies that can easily switch between multiple polymorphic forms[29,30].

METHODS SUMMARY

Wild-type and Tyr21Met fd viruses were purified according to previously published protocols[13]. To reduce the number of longer viruses, which destabilize monolayer structures such as colloidal membranes or twisted ribbons, virus suspensions were cycled through isotropic–nematic phase coexistence. Only the isotropic portion of the samples, enriched in shorter viruses, was used for experiments[11]. The polydispersity of viruses was quantified using gel electrophoresis (Supplementary Fig. 9). All samples were prepared in a buffer containing 20 mM Tris and 100 mM NaCl at pH 8.0. Addition of the non-adsorbing polymer Dextran (relative molecular mass, 500,000; Sigma-Aldrich) induces condensation of rod-like viruses into either flat colloidal membranes or various other chiral assemblages. The strength of chiral interactions between rods was tuned either by changing the sample temperature or by varying the ratio of wild-type fd virus to Tyr21Met fd virus in membranes containing a mixture of the two[13]. Optical microscopy data was taken with an inverted microscope (Nikon TE-2000) equipped with DIC optics, a homemade heating/cooling stage and a cooled charged-coupled-device camera (QImaging, Retiga Exi). The 2D and 3D LC-PolScope imaging set-ups have been described elsewhere[18,20]. To prevent binding of viruses to surfaces, microscope slides and coverslips were cleaned and subsequently coated with a silane agent (3-(trimethoxysilyl) propyl methacrylate, Sigma Aldrich) to which a polyacrylamide brush was polymerized from methacrylate groups.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 July; accepted 6 December 2011.
Published online 4 January 2012.

1. Pasteur, L. On the relations that can exist between crystalline form, and chemical composition and the sense of rotary polarization. *Ann. Chim. Phys.* **24,** 442–459 (1848).
2. Hirokawa, N., Tanaka, Y., Okada, Y. & Takeda, S. Nodal flow and the generation of left-right asymmetry. *Cell* **125,** 33–45 (2006).
3. Harris, A. B., Kamien, R. D. & Lubensky, T. C. Molecular chirality and chiral parameters. *Rev. Mod. Phys.* **71,** 1745–1757 (1999).
4. Oda, R., Huc, I., Schmutz, M., Candau, S. J. & MacKintosh, F. C. Tuning bilayer twist using chiral counterions. *Nature* **399,** 566–569 (1999).
5. Aggeli, A. *et al.* Hierarchical self-assembly of chiral rod-like molecules as a model for peptide beta-sheet tapes, ribbons, fibrils, and fibers. *Proc. Natl Acad. Sci. USA* **98,** 11857–11862 (2001).
6. Kamien, R. D. & Selinger, J. V. Order and frustration in chiral liquid crystals. *J. Phys. Condens. Matter* **13,** R1–R22 (2001).
7. de Gennes, P. G. An anology between superconductors and smectics A. *Solid State Commun.* **88,** 1039–1042 (1993).
8. Renn, S. R. & Lubensky, T. C. Abrikosov dislocation lattice in a model of the cholesteric to smectic-A transition. *Phys. Rev. A* **38,** 2132–2147 (1988).
9. Matsumoto, E. A., Alexander, G. P. & Kamien, R. D. Helical nanofilaments and the high chirality limit of smectics *A. Phys. Rev. Lett.* **103,** 257804 (2009).
10. Hough, L. E. *et al.* Helical nanofilament phases. *Science* **325,** 456–460 (2009).
11. Barry, E. & Dogic, Z. Entropy driven self-assembly of nonamphiphilic colloidal membranes. *Proc. Natl Acad. Sci. USA* **107,** 10348–10353 (2010).
12. Safran, S. *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* 79–85 (Addison Wesley, 1994).
13. Barry, E., Beller, D. & Dogic, Z. A model liquid crystalline system based on rodlike viruses with variable chirality and persistence length. *Soft Matter* **5,** 2563–2570 (2009).
14. Baumgart, T., Hess, S. T. & Webb, W. W. Imaging coexisting fluid domains in biomembrane models coupling curvature and line tension. *Nature* **425,** 821–824 (2003).
15. Honerkamp-Smith, A. R. *et al.* Line tensions, correlation lengths, and critical exponents in lipid membranes near critical points. *Biophys. J.* **95,** 236–246 (2008).
16. Lee, K. Y. C. & McConnell, H. M. Quantitized symmetry of liquid monolayer domains. *J. Phys. Chem.* **97,** 9532–9539 (1993).
17. Dogic, Z. & Fraden, S. Cholesteric phase in virus suspensions. *Langmuir* **16,** 7820–7824 (2000).
18. Oldenbourg, R. & Mei, G. New polarized light microscope with precision universal compensator. *J. Microsc.* **180,** 140–147 (1995).
19. Barry, E., Dogic, Z., Meyer, R. B., Pelcovits, R. A. & Oldenbourg, R. Direct measurement of the twist penetration length in a single smectic a layer of colloidal virus particles. *J. Phys. Chem. B* **113,** 3910–3913 (2009).
20. Oldenbourg, R. Polarized light field microscopy: an analytical method using a microlens array to simultaneously capture both conoscopic and orthoscopic views of birefringent objects. *J. Microsc.* **231,** 419–432 (2008).
21. Aarts, D. G. A. L., Schmidt, M. & Lekkerkerker, H. N. W. Direct visual observation of thermal capillary waves. *Science* **304,** 847–850 (2004).
22. Fradin, C. *et al.* Reduction in the surface energy of liquid interfaces at short length scales. *Nature* **403,** 871–874 (2000).
23. Pelcovits, R. A. & Meyer, R. B. Twist penetration in single-layer smectic A discs of colloidal virus particles. *Liq. Cryst.* **36,** 1157–1160 (2009).
24. Kaplan, C. N., Tu, H., Pelcovits, R. A. & Meyer, R. B. Theory of depletion-induced phase transition from chiral smectic-A twisted ribbons to semi-infinite flat membranes. *Phys. Rev. E* **82,** 021701 (2010).
25. Srivastava, S. *et al.* Light-controlled self-assembly of semiconductor nanoparticles into twisted ribbons. *Science* **327,** 1355–1359 (2010).
26. Chung, W. J. *et al.* Biomimetic self-templating supramolecular structures. *Nature* **478,** 364–368 (2011).
27. Grason, G. M. & Bruinsma, R. F. Chirality and equilibrium biopolymer bundles. *Phys. Rev. Lett.* **99,** 098101 (2007).
28. Claessens, M., Semmrich, C., Ramos, L. & Bausch, A. R. Helical twist controls the thickness of F-actin bundles. *Proc. Natl Acad. Sci. USA* **105,** 8819–8822 (2008).
29. Nguyen, T. D. & Glotzer, S. C. Switchable helical structures formed by the hierarchical self-assembly of laterally tethered nanorods. *Small* **5,** 2092–2098 (2009).
30. Nguyen, T. D. & Glotzer, S. C. Reconfigurable assemblies of shape-changing nanorods. *ACS Nano* **4,** 2585–2594 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the US National Science Foundation (NSF-MRSEC-0820492, NSF-DMR-0955776, NSF-MRI-0923057, NSF-CMMI-1068566) and the Petroleum Research Fund (ACS-PRF 50558-DNI7). We acknowledge use of the Brandeis MRSEC optical microscopy facility.

Author Contributions T.G., E.B., M.J.Z., R.B.M. and Z.D. designed the experiments and interpreted the results. T.G., E.B. and M.J.Z. performed the experiments. A.W. performed the optical trapping experiments. E.B., C.B. and D.N. performed the electron microscopy imaging. M.H. performed the experiments on mutant viruses. R.O. performed the LC-PolScope imaging. Y.Y. and M.F.H. designed and performed the computer simulations. T.G., E.B., M.J.Z. and Z.D. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Z.D. (zdogic@brandeis.edu).

METHODS

Virus stock solutions. Filamentous viruses were grown using standard biological techniques that have been discussed elsewhere[31]. The resulting virus suspensions contain a low fraction of multimeric virus particles (dimers and trimers) with longer contour lengths. Because high monodispersity in particle length is essential for assembly of membranes and ribbons, the polydispersity was reduced by fractionating the samples through the isotropic–nematic phase transition, as longer viruses preferentially dissolve in the nematic liquid-crystalline phase. Suspensions of fd viruses (2–8 ml) were prepared at isotropic–nematic phase coexistence. Only the isotropic (top) portion of the sample, enriched in monomeric viruses, was used for subsequent experiments. The fractionation was repeated once or twice to obtain sufficiently monodisperse suspensions that reproducibly assemble into colloidal membranes and ribbons. The polydispersity of virus suspensions was quantified using gel electrophoresis of intact viruses. Approximately 20 µl of a 0.5 mg ml^{-1} virus solution was loaded into a 1.8% agarose gel and run at 1.4 V cm^{-1} for approximately 8 h (ref. 31). The virus coat protein was dissolved by immersing gels in 0.2 M NaOH for 30 min, and the remaining DNA was visualized with ultraviolet light after staining with ethidium bromide[31] (Supplementary Fig. 9). Only virus suspensions with a monodispersity greater than 95% were used for experiments. For florescence microscopy, viruses were labelled with florescent dyes as described previously[32]. The wild-type fd virus was labelled with Alexa 488-NHS ester (Invitrogen) and the Tyr21Met fd virus was labelled with DyLight 546-NHS ester (Pierce). All samples were buffered in 20 mM Tris, pH 8.0, to which 100 mM NaCl was added to screen electrostatic repulsion between viruses. The concentration of the viruses was determined using absorption spectroscopy[13].

Tuning the twist of the cholesteric phase. The strength of chiral interactions between viruses was quantified by measuring the pitch of a cholesteric liquid-crystalline sample. Glass cylindrical capillaries, 0.7 mm in diameter (Charles Supper Company), were filled with fd virus suspensions. For the cholesteric texture to develop properly, it was necessary to clean the capillaries in a hot detergent solution (Hellmanex) after soaking them in 5 M NaOH. After equilibration for several hours, the samples show a well-known fingerprint texture, seen by polarization microscopy, that is characteristic of cholesteric liquid-crystalline phases[13,17] (Supplementary Fig. 1). The twist wavevector, $q_0 = 2\pi/P$, was determined by measuring the cholesteric pitch, P. Bulk samples were prepared at 100 mg ml^{-1}. Previous work, as well as experiments described in Supplementary Fig. 4, indicate that this is the effective virus concentration in membranes and ribbons[11]. We assume that q_0 is the same for bulk cholesteric phase and 2D membranes.

Two complimentary methods were used to tune the strength of chiral interactions between fd viruses: genetic mutations of the major coat protein and physical control of sample temperature. Most aspects of the wild-type fd virus phase behaviour, such as the location of the isotropic–nematic phase transition, are independent of temperature, as is expected for entropic suspensions of hard, rod-like particles. An exception is that the wild-type fd virus persistence length is weakly dependent on temperature[33]. Data shown in Supplementary Fig. 1 indicates that decreasing temperature increases the strength of chiral interactions between the viruses, as evidenced by the measurements of the cholesteric pitch[17]. Notably, the twist wavevector shows evidence of a second-order phase transition from a high-temperature achiral phase to a low-temperature cholesteric phase. By contrast with the temperature-dependent chirality of wild-type fd virus, the chirality of Tyr21Met fd virus does not show any temperature dependence (Supplementary Fig. 6).

An alternative control of chirality involves modifying the composition of the major coat protein. For example, Tyr21Met fd virus, a mutant that differs from wild-type fd virus by a single point mutation of the major coat protein, forms a cholesteric helix with the opposite handedness[13]. By controlling the ratio of wild-type to Tyr21Met fd virus, it is possible to tune the chirality of the resulting suspension[13]. Such mixtures still form colloidal membranes, and fluorescence microscopy indicates that the two components are uniformly dispersed on optically resolvable length scales (~300 nm). We have used both temperature and genetic methods to tune the chirality of the constituent rods.

Light microscopy sample preparation. Light microscopy chambers were assembled using glass slides and coverslips (Goldseal, Fisher Scientific) with a layer of unstretched Parafilm as a spacer. Glass slides and coverslips were coated with an acrylamide brush to prevent non-specific binding of viruses to the glass surfaces, by a procedure described elsewhere[34]. We mixed fd virus with the non-adsorbing polymer Dextran (relative molecular mass, 500,000; Sigma-Aldrich) to from membrane and ribbons samples. The virus concentration remained fixed at 10 mg ml^{-1}, and the Dextran concentration varied between 30 and 50 mg ml^{-1} as indicated in the main text. The samples, sealed using ultraviolet-cured glue (Norland Optical) and stored at 4 °C, remained good for weeks to months.

Optical microscopy. Samples were characterized using a number of complimentary optical microscopy techniques. Except those made using the 3D LC-PolScope, all observations were made using a ×100 oil-immersion objective (Plan Fluor; numerical aperture, 1.3) on an inverted microscope (Nikon TE-2000). Images were recorded with cooled charge-coupled-device (CCD) cameras (CoolSnap HQ, Photometrics, or Retiga Exi, QImaging). The microscope was equipped with traditional polarization optics, a DIC module, a fluorescence imaging module and 2D LC-Polscope microscopy[18]. For 3D PolScope measurements, we used a Zeiss Axiovert 200M microscope with a Plan Apochomat oil-immersion objective (×63; numerical aperture, 1.4) and a monochrome CCD camera[20] (Retiga 4000R, QImaging).

Sample temperature was tuned between 4 and 60 °C with a home-made Peltier module equipped with a proportional–integral–derivative temperature controller (ILX Lightwave LPT 5910). The temperature-controlling side of the Peltier device was attached to a copper ring fitted around the microscope objective, which heats or cools the sample through the immersion oil. A thermistor, placed in the copper ring adjacent to the sample, enabled the proportional–integral–derivative feedback necessary to adjust the temperature. Excess heat was removed using a constant flow of room-temperature water.

3D LC-PolScope microscopy. Membrane edges with left-handed and right-handed twists appear indistinguishable when viewed with 2D LC-PolScope because this technique provides only a 2D projection of the sample birefringence. To determine the 3D orientation of rods, we used 3D LC-PolScope microscopy[20]. The 3D LC-Polscope extends the capabilities of the 2D LC-PolScope by placing a microlens array in the image plane of the objective lens. The microlens array generates a hybrid image consisting of a series of conoscopic microimages, each probing a different spatial area in the image plane. Analysis of each conoscopic microimage reveals the local birefringence as a function of the propagation direction of transmitted light rays. Supplementary Fig. 4a illustrates the path of the optical rays from the specimen plane to the back focal plane of the 3D LC-PolScope microlens array. Rays shown in green and red pass perpendicularly through the membrane, parallel to its optical axis, and experience no differential retardation. Rays shown in black pass at an angle to the optical axis and are differentially retarded. Rays with the same tilt angle in the specimen plane are focused at the same point in the back focal plane of the objective. The array of microlenses, placed in the image plane of the objective, refocuses these rays into a plane behind the microlens array. Behind each microlens, an image of the objective's aperture appears containing multiple conoscopic images, each specific to rays that have passed through a small region of the specimen. For example, rays emanating from the region X of the sample fall onto the microlens X'. Therefore, only rays that passed through X contribute to the conoscopic image behind X'. The same argument applies to rays passing through Y and Y', and so on.

The conoscopic image of a single microlens focused on the bulk portion of the membrane shows a radially symmetric distribution of the retardance that increases away from the centre of the image (Supplementary Fig. 4b). In the centre, the rays of light are aligned with the orientation of the viruses and the retardance is zero. Away from the centre, the retardance increases because the rays of light become tilted with respect to the orientation of the viruses. When the viruses are tilted with respect to the optical axis, for example near the edge of the membrane, the radial profile obtained from a conoscopic image becomes asymmetric and indicates local tilting of the molecules. For colloidal membranes composed of an achiral mixture of wild-type and Tyr21Met fd viruses, the viruses at the edge of a membrane are found to twist in clockwise and anticlockwise directions with equal probability (Supplementary Fig 4c–e).

Analysis of the 3D LC-PolScope retardance profile is used to determine the concentration of viruses within a membrane, because the membrane birefringence is proportional to the local rod concentration. It follows that the local retardance of the membrane is given by $R = \Delta n_{sat} dcS\sin^2(\theta)/\sin(\alpha)$, where θ is the local tilting of the rods, c is the concentration of rods within a membrane, d is the thickness of the colloidal membrane, S is the local order parameter (assumed to be ~1), Δn_{sat} is the specific birefringence of a fully aligned bulk sample of fd virus at unit concentration as determined by X-ray scattering experiments[35] and α is the orientation of the rays of light as a function of the distance, r, to the centre of the conoscopic image. This orientation is given by $\sin(\alpha) = r\mathrm{NA}/r_{\mathrm{NA}}n$, where $r_{\mathrm{NA}} = 3.9$ µm is the size of the conoscopic image, NA = 1.4 is the numerical aperture and $n = 1.33$ is the index of refraction of the solvent. The fitting of this functional form to the retardance profile yields a mean concentration of viruses in the bulk of the membrane of $c = 103 \pm 10$ mg ml^{-1} (Supplementary Fig. 4f). Similar values where obtained by an independent method that involves counting the number of fluorescently labelled particles[11]. Finally, we also note that the concentration of the viruses does not appreciably change with varying Dextran concentration (Supplementary Fig. 4g).

Measurement of the line tension. Line tension was measured by directly imaging and analysing thermal fluctuations of the membrane's edge with DIC optical microscopy using well-established methods[12]. Lower line tension indicates that

less energy is required to place the particles at an interface, resulting in larger fluctuations of the membrane's edge. The conformation of the membrane's edge is described by $h(x)$, the local height of the interface, where x is the lateral position along the interface. The total length of the interface is

$$L = \int \mathrm{d}x \sqrt{1 + \left(\frac{\mathrm{d}h}{\mathrm{d}x}\right)^2} \quad (1)$$

For small fluctuations ($\mathrm{d}h/\mathrm{d}x = \tan(\theta) \approx \theta$), the expression for the excess length of the edge relative to a flat interface can be approximated as

$$\Delta L = \int \mathrm{d}x \sqrt{1+\theta^2} - L \approx \frac{1}{2}\int \mathrm{d}x\, \theta^2(x)$$

If line tension, γ, is the energetic cost per unit length of the edge, the free-energy difference between fluctuating and straight edges is given by $\Delta F = \gamma \Delta L$. Next we decompose $\theta(x)$ into a Fourier series:

$$\theta(x) = \sum\nolimits_q \sqrt{\frac{2}{L}} a_q \cos(qx)$$

Here L is the length of the interface and the wavevector, q, is given by $q = n\pi/L$. The energetic cost of the membrane conformation specified by Fourier amplitudes a_q is given by

$$\Delta F = \gamma \Delta L = \frac{\gamma}{2}\int_0^L \mathrm{d}x\, \theta^2(x) = \frac{\gamma}{2}\sum\nolimits_q a_q^2$$

By invoking the equipartition theorem, it is possible to relate the line tension to fluctuations of the interface:

$$\langle a_q^2 \rangle = k_\mathrm{B}T/\gamma \quad (2)$$

To measure $\langle {a_q}^2 \rangle$ experimentally, we acquired a series of uncorrelated images of the interfacial contour. To ensure optimal contrast, the DIC shear axis was always perpendicular to the membrane's edge. Under these conditions, intensity profile cuts taken perpendicular to the edge could be fitted to a Gaussian, yielding the conformation of the edge with subpixel accuracy. Each conformation is described in terms of the Fourier amplitudes, a_q. Averaging over a sufficient number of uncorrelated images yields a fluctuation spectrum as shown in Fig. 2b, where the mean square amplitudes, $\langle {a_q}^2 \rangle$, are plotted as functions of q. In the limit of small q, $\langle {a_q}^2 \rangle$ is independent of wavenumber as predicted by equation (2). As q increases, the fluctuations are suppressed and scale as q^{-2}. The entire fluctuation spectrum can be quantitatively described by the following equation:

$$\langle a_q^2 \rangle = \frac{k_\mathrm{B}T}{\gamma + \kappa q^2} \quad (3)$$

The $1/q^2$ regime can be obtained by including terms that are proportional to the square of the gradient of $\theta(x)$ in equation (1). This indicates that there is a bending energy, κ, associated with fluctuations of the interface, which might be related to the existence of a thin, quasi-1D nematic phase located at the edge of the membrane, where the rods are tilted and the layer's smectic ordering is destroyed.

Laser tweezers. Colloidal membranes and ribbons are easily manipulated using optical tweezers. The laser tweezer set-up was built around an inverted Nikon TE-2000 microscope. A 1,064-nm laser beam (Compass 1064, Coherent) is time-shared between two points by means of a pair of orthogonally oriented paratellurite (TeO_2) acousto-optic deflectors (Intra-Action). The laser is projected onto the back focal plane of an oil-immersion objective (Plan Fluor ×100, NA = 1.3) and subsequently focused onto the imaging plane. Using custom LABVIEW software, multiple trap locations were specified and used to stretch and manipulate membranes and ribbons.

Electron microscopy. For transmission electron microscopy, membranes were assembled as described above but with the following differences. The monolayer membranes are relatively small and only stable in solution; therefore, to keep them intact throughout fixation and staining, we immobilized the membrane assemblies by embedding them in 1% low-melting-temperature agarose (SeaKem; gelation temperature, 25 ± 5 °C), in 100 mM NaCl, 20 mM Tris, pH 8.15, and 150 mM sucrose as a cryoprotectant. Virus samples were mixed in 1.5-ml Eppendorf tubes with fluid agarose at 30 °C and kept warm for 3–6 h to allow formation of relatively large membranes as confirmed by light microscopy. Subsequently, a drop of the fluid sample containing membranes was transferred into the 0.1-mm-high cavity of an aluminium planchette for high-pressure freezing (Wohlwend), which was then closed. Then the sample temperature was decreased to 4 °C for 30–45 min. The gelled samples were rapidly frozen using a Leica HPM-100 high-pressure freezer (Leica Microsystems). Using a Leica AFS-2 device, the frozen samples were freeze-substituted at low temperatures (starting at −90 °C) over the course of three days in a solution containing 1% osmium tetroxide (EMS), 0.5% anhydrous glutaraldehyde (EMS) and 2% water in anhydrous acetone (AC32680-0010, Fisher Scientific). After the temperature was raised to 4 °C, the sample was infiltrated and embedded in EMbed 812-Resin (EMS). Ultrathin, 70-nm, sections were collected on slot grids covered with Formvar support film and post-stained with uranyl acetate (supersaturated solution) and 0.2% lead citrate, before inspection on a FEI Morgan 268 transmission electron microscope with a 1k CCD camera (GATAN) and on a 300-keV Tecnai F30 intermediate-voltage transmission electron microscope (FEI) with a 4k CCD camera (GATAN). For electron tomography, we used 150-nm-thick sections that were coated with 10-nm colloidal gold particles. Tilt series were recorded over an angular range of ±60° with 1° increments on the Tecnai F30 using the microscope control software SERIALEM[36]. Tomographic reconstructions were calculated with the IMOD software using fiducial alignment and weighted back-projection[37].

Computer simulations of the rod/polymer mixture. In the computational model, the rods were represented as hard spherocylinders and the non-adsorbing polymers (depletant) were modelled as ghost spheres, which act as hard spheres when interacting with rods but can freely interpenetrate one another[38]. We performed Metropolis Monte Carlo simulation with periodic boundary conditions. The system contained 6,000 rods in a box with dimensions 140 × 140 × 50 (in units of the rod diameter). A constant osmotic pressure was maintained by coupling ghost spheres to a constant-chemical-potential bath through insertion and deletion moves. The simulations thus sampled the $N_\mathrm{rod}\mu_\mathrm{sphere}VT$ ensemble, where N_rod is the number of rods, μ_sphere is the sphere chemical potential and V is the volume. The rod aspect ratio was 20, the ghost-sphere diameter was 1.5 and on average there were about 520,000 spheres. Simulations were initialized with rods in a monolayer, and Monte Carlo simulation was performed until rod density within the membrane and rod orientations equilibrated. Rods at the membrane edge twisted spontaneously regardless of the initial rod orientations in the membrane.

31. Maniatis, T., Sambrook, J. & Fritsch, E. *Molecular Cloning* Ch. 3 (1989).
32. Lettinga, M. P., Barry, E. & Dogic, Z. Self-diffusion of rod-like viruses in the nematic phase. *Europhys. Lett.* **71,** 692–698 (2005).
33. Tang, J. X. & Fraden, S. Nonmonotonic temperature dependence of the flexibility of bacteriophage fd. *Biopolymers* **39,** 13–22 (1996).
34. Lau, A. W. C., Prasad, A. & Dogic, Z. Condensation of isolated semi-flexible filaments driven by depletion interactions. *Europhys. Lett.* **87,** 48006 (2009).
35. Purdy, K. R. *et al.* Measuring the nematic order of suspensions of colloidal fd virus by X-ray diffraction and optical birefringence. *Phys. Rev. E* **67,** 031708 (2003).
36. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152,** 36–51 (2005).
37. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116,** 71–76 (1996).
38. Yang, Y., Barry, E., Dogic, Z. & Hagan, M. F. Self-assembly of 2D membranes from mixtures of hard rods and depleting polymers. *Soft Matter* **8,** 707–714 (2012).

LETTER

doi:10.1038/nature10740

Kimberlite ascent by assimilation-fuelled buoyancy

James K. Russell[1], Lucy A. Porritt[1], Yan Lavallée[2] & Donald B. Dingwell[2]

Kimberlite magmas have the deepest origin of all terrestrial magmas and are exclusively associated with cratons[1–3]. During ascent, they travel through about 150 kilometres of cratonic mantle lithosphere and entrain seemingly prohibitive loads (more than 25 per cent by volume) of mantle-derived xenoliths and xenocrysts (including diamond)[4,5]. Kimberlite magmas also reputedly have higher ascent rates[6–9] than other xenolith-bearing magmas[10,11]. Exsolution of dissolved volatiles (carbon dioxide and water) is thought to be essential to provide sufficient buoyancy for the rapid ascent of these dense, crystal-rich magmas. The cause and nature of such exsolution, however, remains elusive and is rarely specified[6,9]. Here we use a series of high-temperature experiments to demonstrate a mechanism for the spontaneous, efficient and continuous production of this volatile phase. This mechanism requires parental melts of kimberlite to originate as carbonatite-like melts. In transit through the mantle lithosphere, these silica-undersaturated melts assimilate mantle minerals, especially orthopyroxene, driving the melt to more silicic compositions, and causing a marked drop in carbon dioxide solubility. The solubility drop manifests itself immediately in a continuous and vigorous exsolution of a fluid phase, thereby reducing magma density, increasing buoyancy, and driving the rapid and accelerating ascent of the increasingly kimberlitic magma. Our model provides an explanation for continuous ascent of magmas laden with high volumes of dense mantle cargo, an explanation for the chemical diversity of kimberlite, and a connection between kimberlites and cratons.

Our current understanding of kimberlite ascent is greatly hampered by uncertainty about the compositions of primary kimberlite melts[1,3,5,12]. These compositions remains elusive for three reasons: kimberlites contain abundant (>30%) xenocrystic material; kimberlite melt quenched to glass has not been observed; and kimberlites are highly susceptible to alteration owing to their ultrabasic composition[1] and age[2]. Many strategies have been used to estimate kimberlite melt compositions[12,13] but the range of postulated compositions remains large and contentious (SiO_2 < 25–35%, MgO 15–25%)—especially the content of dissolved volatiles (CO_2 and H_2O)[1,3,12,13]. Physical properties estimated for the putative range of melt compositions include densities similar to basaltic or komatiitic melts (2,800–2,900 kg m^{-3}) and low viscosities (10^{-2}–10^{2} Pa s)[12].

Mantle xenoliths transported by kimberlite constrain the formation depths of these magmas to the base of the cratonic mantle lithosphere (CML) or deeper[4]. Mechanical disaggregation of xenoliths produces the mantle-olivine-dominated (>25 vol.%) suite of xenocrysts characteristic of most kimberlites (Fig. 1a). Notably, orthopyroxene (opx), which comprises ~15–27% of the CML[14–16] and is the most silicic mineral present (Fig. 1b), is rare-to-absent in kimberlite. Where observed, opx texturally records severe disequilibrium (Fig. 1c)[5,17,18]. The high crystal contents suggest bulk densities of the kimberlite magma (melt+crystal) in the range of 3.0–3.1 g cm^{-3}. Buoyant ascent of kimberlite magma from depths >120 km can be greatly enhanced by the presence of a low-density exsolved fluid phase (density of CO_2 fluid $\rho_{CO_2} \approx 1.2$ g cm^{-3} at pressure $P \approx 2$ GPa).

Here we present a mechanism for the rapid ascent of these deep-seated, high-density magmas. Our model is inspired by new

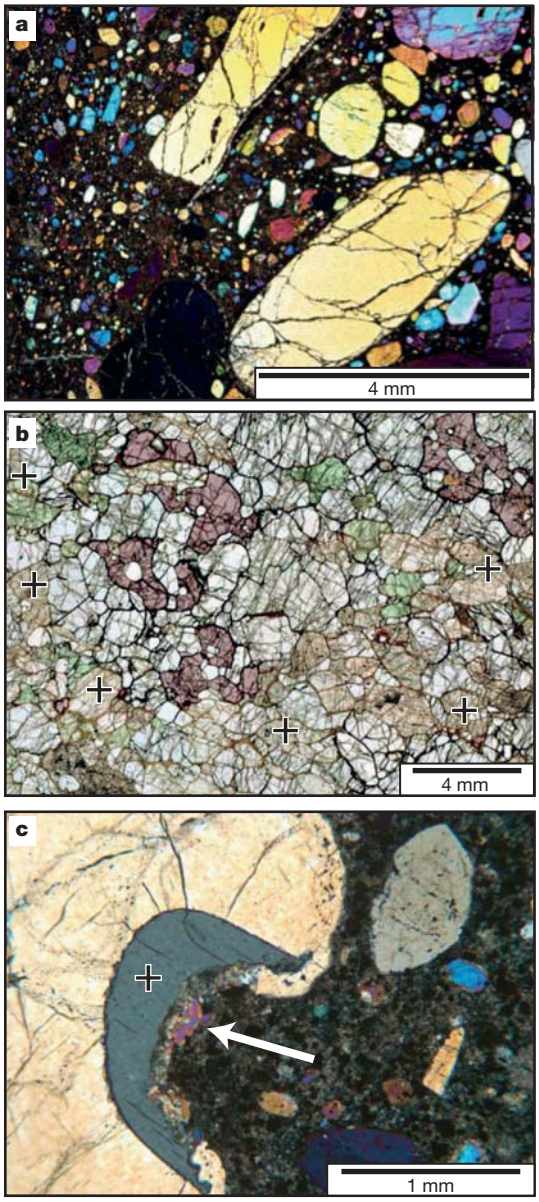


Figure 1 | Transmitted-light optical microscope images of thin sections of kimberlite, peridotite and orthopyroxene. **a**, Kimberlite hosts abundant rounded centimetre- to millimetre-scale mantle-derived xenocrysts of olivine (ol) and subordinate amounts of xenocrystic clinopyroxene (cpx), garnet (gar) and ilmenite (ilm). Orthopyroxene (opx) is conspicuously absent. **b**, Peridotitic xenolith from kimberlite, showing original mantle-equilibrated mineralogy (ol > opx > cpx > gar), textures, and grain size and shape distributions of mantle minerals. Grains of opx are marked with '+' (image courtesy of M. Kopylova). **c**, Kimberlite containing a fragment of disaggregated mantle peridotite (arrow), comprising a grain of partially dissolved opx (+; grey grain) within larger ol grain (yellow grain) (image courtesy of C. Brett).

[1]Volcanology and Petrology Laboratory, Earth and Ocean Sciences, University of British Columbia, Vancouver V6T 1Z4, Canada. [2]Department of Earth and Environmental Sciences, Ludwig Maximilian University of Munich, München 80333, Germany.

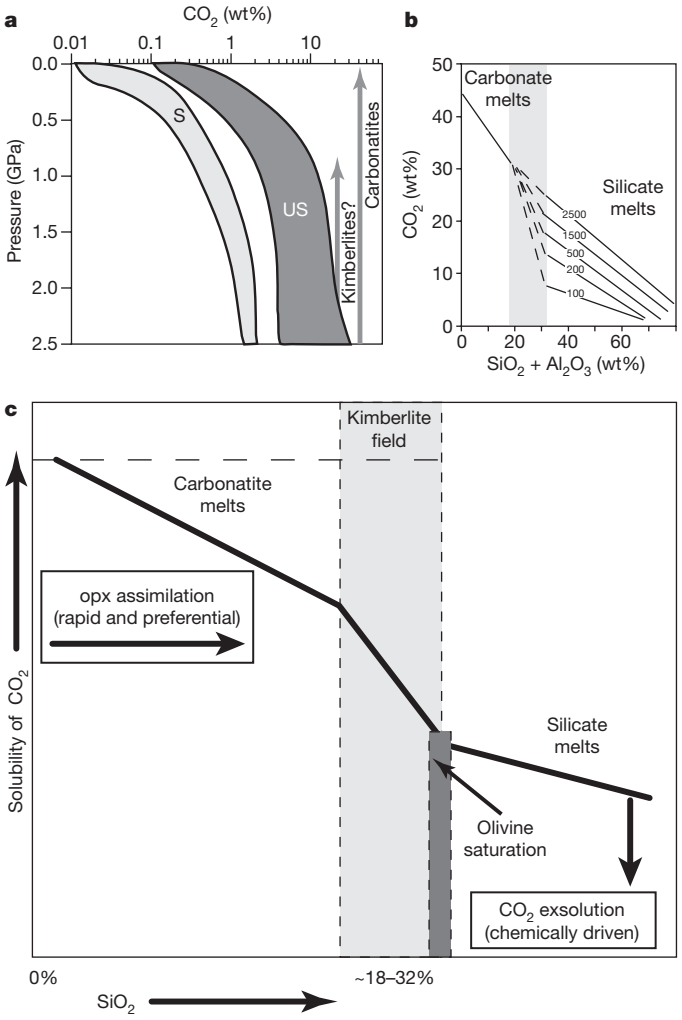


Figure 2 | CO_2 solubilities in silicic to carbonate melts (after ref. 19). **a**, CO_2 solubility limits for silica-saturated (S) and undersaturated (US) melts and hypothetical solubilities of carbonatite and kimberlite melt. **b**, Pressure and composition dependence of CO_2 solubility across the carbonate–silicate transition. Note limited effects of pressure (numbers on lines, MPa) on CO_2 solubility versus the strong effects of composition (that is, $SiO_2 + Al_2O_3$ wt%). **c**, Schematic model for assimilation-induced fluid-exsolution (that is, vesiculation) of carbonatitic or proto-kimberlitic melts. Opx assimilation drives non-silicate melts (left side) to more silicic compositions (towards right side), causing a decrease in CO_2 solubility and exsolution of a fluid phase. Continued assimilation causes continued exsolution, and results in a kimberlitic melt and olivine saturation at about 18–32 wt% SiO_2.

experimental data on the solubility of volatiles (that is, CO_2) in melts across the silicate–carbonate transition[19] (Fig. 2). This work elucidates the pressure and compositional controls on CO_2 solubility in silicic (for example, basalt to melilite) to carbonatitic melts (Fig. 2a,b) and provides three main insights. First, the solubility of CO_2 (and H_2O) in silicic melts increases with pressure but, importantly, in magmas with $SiO_2 + Al_2O_3 > 35$ wt%, solubility is limited to ~10–15%. This precludes extraordinary amounts of volatile being sequestered in parental kimberlitic melts[9,20]. Second, carbonatitic melts (that is, $SiO_2 + Al_2O_3 < 20$ wt%) have substantially higher CO_2 solubilities that are essentially independent of pressure (Fig. 2b). Solubility of CO_2 in carbonate melts is limited only by melt stoichiometry, and can be >40% (refs 21, 22). Last, the transition from carbonatitic to silicic (that is, kimberlitic) melt compositions is accompanied by a pronounced decrease in CO_2 solubility (Fig. 2b).

We suggest that melts parental to kimberlite originate as carbonatitic or near-carbonatitic melts. Such melts begin with near-stoichiometric CO_2 contents (~40 wt%) and are buoyant relative to the cratonic mantle lithosphere (density $\rho \approx 2.6$–2.8 g cm^{-3})[22]. Carbonatitic melts can also accommodate substantial amounts of water (10 wt% at 0.1 GPa) without reducing their CO_2 contents[23]. The melts entering the CML initiate the mechanical incorporation and disaggregation of xenoliths[6,24], thereby ensuring that newly liberated xenocrysts are continuously fed into the melt. At these pressures ($P < 2.5$ GPa) and temperatures ($T > 1{,}250$ °C), all mantle minerals are out of equilibrium and react with the low-a_{SiO_2} carbonatitic melt[25] (a_{SiO_2} is the activity of SiO_2). Orthopyroxene, as the most silica-saturated phase, has the highest affinity for dissolution and is assimilated rapidly and preferentially over other phases[5,17,25–27].

Assimilation of opx drives the melt towards more silicic compositions (Fig. 2c). The increased SiO_2 content of the contaminated carbonatitic melt causes a decrease in CO_2 solubility that is expressed by immediate fluid exsolution deep within the CML. Continued opx assimilation is attended by continuous fluid exsolution, leading to ever-increasing buoyancy. The process is gradual until opx assimilation produces a 'kimberlitic' melt composition ($SiO_2 > 18$ wt%), which initiates a catastrophic drop in CO_2 solubility. This results in massive exsolution of a volatile phase that reduces magma density, increases buoyancy, and supports rapid and accelerating ascent.

Results from a series of high-temperature melting experiments test these inductive ideas, and illustrate the potential speed and efficiency of the process. Our experiments involve melting powdered mixtures of Na_2CO_3 and natural mantle-derived opx ($MgSiO_3$; Fig. 3a), and measuring the weight loss as a proxy for CO_2 loss. The procedures and resulting data are fully described in Methods and Supplementary Table 1.

In non-steady-state ('transient') experiments, the melt rapidly assimilates the opx, creating a more silicic melt that is oversaturated in dissolved CO_2. The melt exsolves dissolved CO_2 (that is, undergoes decarbonation), causing a weight loss proportional to the opx content (Fig. 3b). These experiments demonstrate the spontaneous nature and rapid (~20 min) rate of the assimilation and decarbonation reactions. Reaction (that is, weight loss) continues until all of the opx has dissolved, the melt has reached its maximum SiO_2 content, and the CO_2 solubility limit of the new more silicic melt is reached (Fig. 3b). Static experiments track the change in melt composition and the production of CO_2 via melt decarbonation (Fig. 3c,d). The SiO_2 and MgO contents of the hybrid melts are greater than the bulk compositions (Fig. 3a) owing to enrichment by CO_2 loss. Na_2O content shows an initial increase and then decreases owing to the combined effects of dilution (addition of opx) and enrichment (decarbonation). CO_2 decreases dramatically. The addition of ~42 wt% opx produces SiO_2-rich (~32.7 wt%) melts in which all CO_2 is lost due to decarbonation (Fig. 3d). The maximum amount of CO_2 that would be released from the magma during assimilation to fuel the ascent of kimberlite is ~25–30 wt% (Fig. 3c, d).

Primary mantle-derived carbonatitic to near-carbonatitic (that is, low-a_{SiO_2}) melts provide an appealing model for parental kimberlite magma. Carbonatite melts are accepted as a common product of partial melting of carbonated mantle sources at pressures >2.5 GPa (refs 21, 28). Recent experimental work shows that the solidus melts of carbonated peridotite at >2.5 GPa will always be CO_2 rich and SiO_2 poor as long as carbonate is stable in the mantle assemblage[21]. These experimentally produced melts contain in excess of 40% dissolved CO_2, and can even accommodate substantial dissolved H_2O (ref. 23); we suggest that the asthenospheric production of such melts marks the onset of kimberlite ascent (Fig. 4a, b). Our mechanistic model for kimberlite ascent (Fig. 4b, c) supports petrogenetic models that interpret the diversity of kimberlite bulk rock compositions as mechanical mixing of mantle olivine (70–80%) and a carbonate-rich melt[29] combined with assimilation of mantle orthopyroxene (20–30%)[30].

We modelled the chemical evolution of the ascending melt by assuming that the amount of opx assimilated is linearly related to the

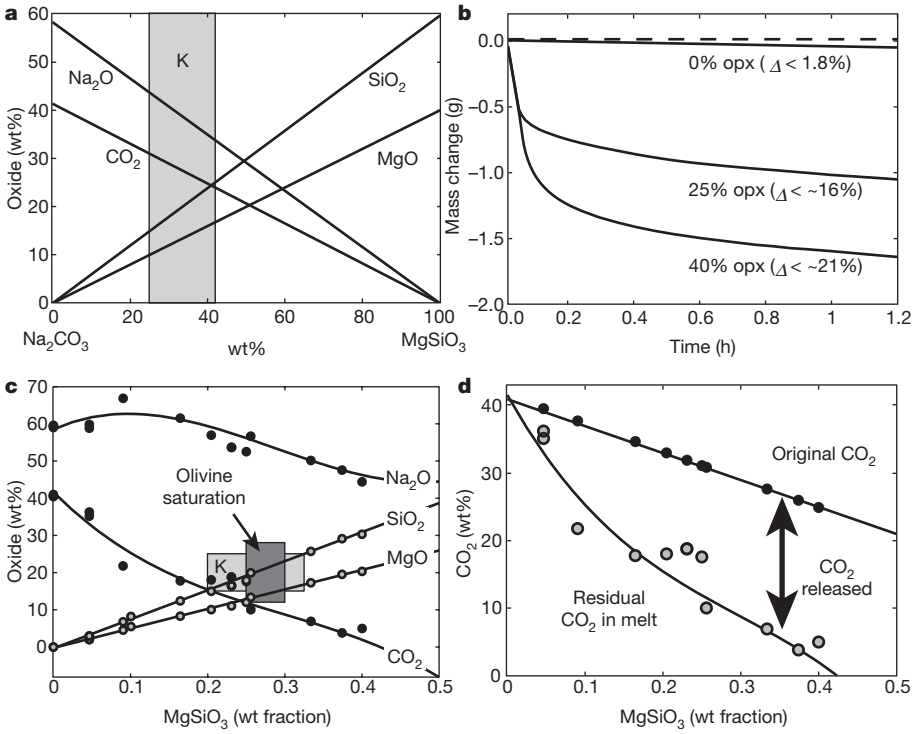


Figure 3 | High-temperature analogue experiments. **a**, Compositions produced by mechanical mixing of sodium carbonate (Na_2CO_3) and orthopyroxene (opx, $MgSiO_3$). (Data given in Supplementary Table 1). Shaded field (K) denotes range of SiO_2 contents for kimberlite melts based on Igwisi Hills kimberlite (Supplementary Table 2). **b**, Results of transient high-temperature experiments, showing mass change with time due to decarbonation driven by opx assimilation. Total weight loss (Δ) is reported in parentheses. **c**, Melt compositions produced by chemical reaction during melting of Na_2CO_3-opx mixtures versus opx content. Shaded fields mark kimberlite (K) melts and condition for olivine saturation and crystallization. **d**, Post-experiment distribution of CO_2 between hybrid melt and the exsolved fluid phase. Distance between two lines is total exsolved (released) CO_2.

distance travelled in the mantle lithosphere, and that assimilation and decarbonation are essentially instantaneous (left dashed line, Fig. 4d; Supplementary Table 2). This process can be sustained energetically by the adiabatic decompression of the magma during its ascent, which can result in considerable (150–200 °C) superheat[31]. After assimilation of ~15–20% opx, a carbonatite melt will evolve to a composition close to melt compositions estimated for the Igwisi Hills kimberlite lavas[32] (SiO_2 16–22 wt%; Supplementary Table 2; Fig. 4d). The amount of CO_2 released during assimilation and available to fuel the ascent of the contaminated carbonatite (kimberlite in the widest sense) is ~25%. Ultimately, continued opx assimilation drives the melt towards olivine saturation (Fig. 2c), allowing for late-stage magmatic crystallization of olivine as overgrowths observed on rounded olivine xenocrysts or as microphenocrysts[5,18,25,26] (Fig. 2c). Olivine crystallization would logically coincide with the later stages of kimberlite ascent and emplacement, and clearly follows on from an earlier and more extensive period of olivine disequilibrium.

The assimilation of opx is the key to this process, because it dissolves on timescales that compete with kimberlite ascent rates[25–27]. The rapid and efficient dissolution of opx, relative to other mantle xenocrysts, explains the absence of xenocrystic opx in kimberlite[5,17]. However, the main mechanistic consequence of opx assimilation is to decrease the CO_2 solubility of the melt, thereby driving deep-seated exsolution and effervescence of a CO_2-rich volatile phase (Fig. 4c). The exsolved fluid enhances the buoyancy of the ascending magma within the deep mantle, and provides an explanation for its rapid ascent and its ability to continuously entrain and transport mantle cargo to the Earth's surface.

Our model for kimberlite ascent is unique in that it begins with a low-a_{SiO_2} melt containing stoichiometric amounts of dissolved CO_2. The CO_2 is stable within the melt across a wide range of pressure–temperature conditions and, thus, exsolution of the fluid phase is not driven by depressurization, as it is in most magma ascent models. Rather, CO_2 fluid exsolution is driven by an unavoidable chemical reaction (that is, assimilation) that creates a new more silicic melt whose CO_2 solubility is exceeded; the exsolution is spontaneous and immediate. Although the initial fluid production is not driven by a change in pressure, the buoyancy of the fluid-saturated magma will be greatly enhanced by the volumetric expansion of the fluids attending depressurization. This expansion allows for continuous acceleration of the magma, entrainment of greater quantities of mantle material, and decoupling (that is, separation) of a CO_2-rich fluid phase from the ascending magma to create precursory fenitization (alkali metasomatism) events[33].

Our model provides an explanation for the linkage between kimberlites and cratons; it is clear that, if the parental magmas are carbonatitic, that a ready source of highly reactive opx is necessary to produce a more silica-rich melt (that is, kimberlite) that will exsolve fluid in the deep mantle. The cratonic mantle lithosphere has two attributes that support this transformation: first, it is enriched in opx (15–30%) relative to other mantle lithosphere[14–16]; and second, it is thick (90–120 km), and thus provides ample opportunity for sampling of opx-rich mantle. As long as the evolving carbonatitic melt is exposed to new opx, the assimilation-induced exsolution of CO_2 will continue and buoyancy is maintained or increased. Therefore, we might expect the diversity of kimberlite at the Earth's surface to actually reflect the mineralogical composition, thermal state and the thickness of the underlying mantle lithosphere rather than the source region of the melt.

The corollary is that if opx were absent, this process would not be viable, because the dissolution rates for other mantle silicates (which could also promote an assimilation-driven exsolution of fluid) are too low to keep pace with ascent rate. In addition, the enthalpy of opx dissolution is less than the heat of crystallization for olivine, implying that opx assimilation can be balanced energetically by smaller quantities of olivine crystallization and a little cooling[18]. Dissolution of olivine,

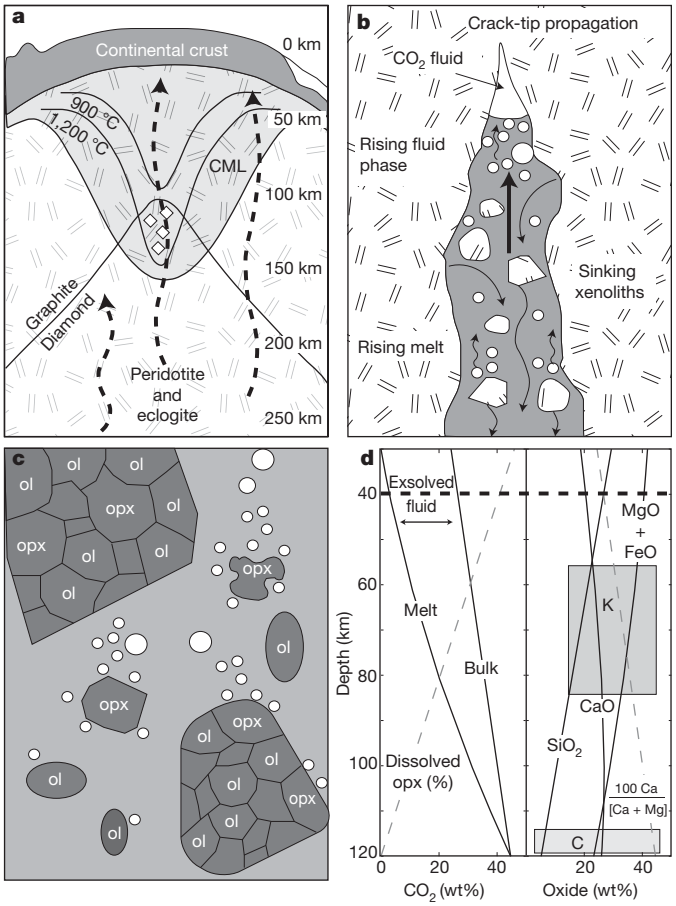


Figure 4 | Model for assimilation-fuelled ascent of kimberlite. **a**, Diverse ascent paths through cratonic mantle lithosphere (CML) shown as dashed arrows. Also shown is the line below which diamond is stable relative to graphite. **b**, Melts produced by melting of carbonated peridotite transit mantle lithosphere as dykes by crack-tip propagation, liberating dense (sinking) xenoliths to the CO_2-rich silica-undersaturated melt, and causing effervescence of buoyant (rising) CO_2-fluid. **c**, Xenoliths disaggregate and release individual mineral grains (for example, ol) to carbonatitic melt; opx grains are assimilated, preferentially promoting volatile exsolution. Deep-seated volatile production supports continued, crack-propagation-limited magma ascent. **d**, Chemical evolution of melt during ascent, including: left panel, wt% of opx assimilated (dashed line, dissolved), and CO_2 content in bulk system and in melt (solid lines); right panel, oxide (SiO_2, MgO + FeO, CaO) contents of the evolving melt (solid lines), and the 100 Ca/(Ca + Mg) ratio of melt (dashed line). Initial carbonatitic and derivative kimberlite melts (shaded boxes) defined by SiO_2 contents (Supplementary Table 2).

for example, would require substantial cooling of the magma (as well as longer times). In parts of the Earth where carbonatitic magmas do not transit the cratonic mantle lithosphere, there are only two possible outcomes. The carbonatitic melt successfully travels through a relatively thin lithosphere to be emplaced or erupted as carbonatite, or the carbonatite melt incorporates opx-poor mantle and crystallizes at depth because it fails to assimilate mantle silicates efficiently enough to promote a deep-seated fluid phase and the requisite buoyancy.

We suggest the possibility that all kimberlites start life as carbonatitic melts produced through partial melting of carbonated peridotite in the deep subcratonic mantle. As the melts transit the mantle lithosphere they continuously incorporate and disaggregate peridotite xenoliths. The low a_{SiO_2} of the carbonatitic melt causes the rapid and preferential dissolution of opx relative to the other silicate mantle minerals. This increases the silica content of the melt and triggers immediate exsolution of CO_2, reducing the buoyancy of the magma and facilitating rapid ascent. The faster the magma rises, the more mantle material will be entrained and the more opx will be dissolved. Providing opx is available, this mechanism enables the continuous and accelerating ascent of the magma and the evolution of the melt from carbonatitic to kimberlitic compositions.

METHODS SUMMARY

Our experiments use synthetic melts of Na_2CO_3 (melting temperature $T_m = 851\,^{\circ}C$) as an analogue for a natural carbonatitic-like melt. The experiments involve the melting of mixtures of Na_2CO_3 and crushed natural opx at temperatures of 1,000–1,100 °C for ~1 h; the mixture is open to the air. Pure Na_2CO_3 melt contains ~44 wt% dissolved CO_2 and is stable at these conditions. Melting of Na_2CO_3-opx mixtures produces vigorous decarbonation expressed by weight loss. Decarbonation results from reduced CO_2 solubility in the hybrid melt produced by opx assimilation (Fig. 2). Both transient and static experiments were performed. In the former, an initial mass of sample (Na_2CO_3 + opx) of known composition was lowered from a balance into a preheated furnace; cumulative weight loss was recorded as a function of time. The transient experiment on pure Na_2CO_3 (>1 h at 1,100 °C) shows only slight (<1.8%) change in mass and demonstrates the stability of stoichiometric amounts of CO_2 in the melt. Transient melting of Na_2CO_3-opx mixtures causes immediate reaction as opx is assimilated; the resulting hybrid melts exsolve CO_2 causing a continuous but decreasing loss in weight, until the equilibrium solubility limit of the new melt composition is reached, and weight loss ceases (<20 min). The magnitude of weight loss is taken as a quantitative measure of CO_2 loss. The tangents to the weight loss curves give the relative rates of assimilation and degassing. Static experiments saw known masses and compositions of Na_2CO_3 + opx placed into a pre-heated oven (1,000–1,100 °C) for a fixed time (Supplementary Table 1; Fig. 3a). The weight loss at the end of each experiment was measured to provide a quantitative estimate of decarbonation.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 July; accepted 28 November 2011.

1. Mitchell, R. H. *Kimberlites: Mineralogy, Geochemistry and Petrology* (Plenum, 1986).
2. Nixon, P. H. The morphology and nature of primary diamondiferous occurrences. *J. Geochem. Explor.* **53,** 41–71 (1995).
3. Kjarsgaard, B. A. in *Mineral Deposits of Canada: A Synthesis of Major Deposit-Types, District Metallogeny, The Evolution Of Geological Provinces, and Exploration Methods* (ed. Goodfellow, W. D.) 245–271 (Geological Association of Canada, Mineral Deposits Division, 2007).
4. Boyd, F. R. & Nixon, P. H. in *Lesotho Kimberlites* (ed. Nixon, P. H.) 254–268 (Lesotho National Development Corporation, Maseru, 1973).
5. Mitchell, R. H. Petrology of hypabyssal kimberlites: relevance to primary magma compositions. *J. Volcanol. Geotherm. Res.* **174,** 1–8 (2008).
6. Anderson, O. L. in *Kimberlites, Diatremes and Diamonds: Their Geology, Petrology and Geochemistry* (eds Meyer, H. O. A. & Boyd, F. R.) 344–353 (AGU, 1979).
7. Canil, D. & Fedortchouk, Y. Garnet dissolution and the emplacement of kimberlites. *Earth Planet. Sci. Lett.* **167,** 227–237 (1999).
8. Sparks, R. S. J. *et al.* Dynamical constraints on kimberlite volcanism. *J. Volcanol. Geotherm. Res.* **155,** 18 (2006).
9. Wilson, L. & Head, J. W. An integrated model of kimberlite ascent and eruption. *Nature* **447,** 53–57 (2007).
10. Sparks, R. S. J., Pinkerton, H. & Macdonald, R. The transport of xenoliths in magmas. *Earth Planet. Sci. Lett.* **35,** 234–238 (1977).
11. Spera, F. K. Carbon dioxide in petrogenesis III: role of volatiles in the ascent of alkaline magma with special reference to xenolith-bearing mafic lavas. *Contrib. Mineral. Petrol.* **88,** 217–232 (1984).
12. Sparks, R. S. J. *et al.* The nature of erupting kimberlite melts. *Lithos* **112,** 429–438 (2009).
13. Price, S. E., Russell, J. K. & Kopylova, M. G. Primitive magma from the Jericho Pipe, N.W.T., Canada: constraints on primary kimberlite melt chemistry. *J. Petrol.* **41,** 789–808 (2000).
14. Boyd, F. R. Compositional distinction between oceanic and cratonic lithosphere. *Earth Planet. Sci. Lett.* **96,** 15–26 (1989).
15. Kopylova, M. G. & Russell, J. K. Chemical stratification of cratonic lithosphere: constraints from the Northern Slave craton, Canada. *Earth Planet. Sci. Lett.* **181,** 71–87 (2000).
16. McDonough, W. F. Constraints on the composition of the continental lithospheric mantle. *Earth Planet. Sci. Lett.* **101,** 1–18 (1990).
17. Mitchell, R. H. Composition of olivine, silica activity and oxygen fugacity in kimberlite. *Lithos* **6,** 65–81 (1973).
18. Brett, R. C., Russell, J. K. & Moss, S. Origin of olivine in kimberlite: phenocryst or imposter? *Lithos* **112,** 201–212 (2009).
19. Brooker, R. A., Sparks, R. S. J., Kavanagh, J. & Field, M. The volatile content of hypabyssal kimberlite magmas: some constraints from experiments on natural rock compositions. *Bull. Volcanol.* **73,** 959–981 (2011).
20. Sparks, R. S. J., Brown, R. J., Field, M. & Gilbertson, M. A. Kimberlite ascent and eruption. *Nature* **450,** E21 (2007).
21. Dalton, J. A. & Presnall, D. C. The continuum of primary carbonatitic-kimberlitic melt compositions in equilibrium with lherzolite: data from the system CaO-MgO-Al_2O_3-SiO_2-CO_2 at 6 GPa. *J. Petrol.* **39,** 1953–1964 (1998).

22. Dobson, D. P. *et al.* In-situ measurement of viscosity and density of carbonate melts at high pressure. *Earth Planet. Sci. Lett.* **143,** 207–215 (1996).
23. Keppler, H. Water solubility in carbonatite melt. *Am. Mineral.* **88,** 1822–1824 (2003).
24. Lensky, N. G., Niebo, R. W., Holloway, J. R., Lyakhovsky, V. & Navon, O. Bubble nucleation as a trigger for xenolith entrapment in mantle melts. *Earth Planet. Sci. Lett.* **245,** 278–288 (2006).
25. Luth, R. W. The activity of silica in kimberlites: revisited. *Contrib. Mineral. Petrol.* **158,** 283–294 (2009).
26. Shaw, C. S. J. Dissolution of orthopyroxene in basanitic magma between 0.4 and 2 GPa: further implications for the origin of Si-rich alkaline glass inclusions in mantle xenoliths. *Contrib. Mineral. Petrol.* **135,** 114–132 (1999).
27. Edwards, B. R. & Russell, J. K. Time scales of magmatic processes: new insights from dynamic models for magmatic assimilation. *Geology* **26,** 1103–1106 (1998).
28. Wyllie, P. J. & Huang, W. L. Peridotite, kimberlite, and carbonatite explained in the system CaO-MgO-SiO_2-CO_2. *Geology* **3,** 621–624 (1975).
29. Canil, D. & Bellis, A. J. Phase equilibria in a volatile-free kimberlite at 0.1 MPa and the search for primary kimberlite magma. *Lithos* **105,** 111–117 (2008).
30. Patterson, M., Francis, D. & McCandless, T. Kimberlites: magmas of mixtures? *Lithos* **112,** 191–200 (2009).
31. Kavanagh, J. & Sparks, R. S. J. Temperature changes in ascending kimberlite magma. *Earth Planet. Sci. Lett.* **286,** 404–413 (2009).
32. Dawson, J. B. Quaternary kimberlitic volcanism on the Tanzania Craton. *Contrib. Mineral. Petrol.* **116,** 473–485 (1994).
33. Ferguson, J., Danchin, R. V. & Nixon, P. H. in *Lesotho Kimberlites* (ed. Nixon, P. H.) 207–213 (Lesotho National Development Corporation, Maseru, Lesotho, 1973).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the laboratory support of W. Ertel-Ingritsch and S. Laumann at Ludwig Maximilian University. Funding for this research is from the Natural Sciences and Engineering Research Council (J.K.R.), a Marie Curie outbound fellowship (L.A.P.) and an ERC Advanced Researcher Grant (D.B.D.). The original manuscript benefitted from review by S. Sparks.

Author Contributions J.K.R. conceptualized the original idea, performed the static experiments at the University of Munich, and wrote the original draft paper. L.A.P. provided kimberlite expertise and context and drafted the figures. Y.L. performed the transient experiments and processed the experimental data. D.B.D. hosted, provided and repaired the experimental facility, and advised on the optimal experimental methods. All co-authors contributed to producing a final draft for review and in revising the manuscript after review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.K.R. (krussell@eos.ubc.ca).

METHODS

Experiments. We have used a series of high-temperature melting experiments to illustrate aspects of the proposed model for the ascent of kimberlite. Our experiments use synthetic melts of sodium carbonate as an analogue for a natural carbonatitic melt that could be parental (that is, pre-assimilation) to kimberlites. The experimental results provide a semiquantitative test of the ideas advanced in our manuscript. The experiments (for example, conditions, melt compositions, and scales) are not intended to mimic the natural process exactly, but to illustrate what is possible and how the process may operate during magma transport.

Our experimental design follows ref. 34, in which was explored the melting reactions between superliquidus sodium carbonate (Na_2CO_3) and silica sand. Pure Na_2CO_3 melt contains ~44 wt% (50 mol%) dissolved CO_2. Fusion of the mechanical mixtures of Na_2CO_3 and silica sand at temperatures above melting ($T_m > 851\,^{\circ}C$) caused the Na_2CO_3 melt to react vigorously with the silica grains and exsolve CO_2 to the atmosphere[34].The magnitude of CO_2 loss was quantitatively expressed as a loss in weight and trivial weight loss due to Na_2O volatilization was demonstrated[34].The decarbonation of the melts continues until the SiO_2 contents of the hybrid melts exceed the stoichiometry of sodium orthosilicate ($Na_4SiO_4 \approx 32.7$ wt%) causing crystallization of the silicate phase.

Our experiments used Na_2CO_3 melts at superliquidus temperatures of 1,000–1,100 °C into which natural orthopyroxene (opx, $MgSiO_3$) had been added (Supplementary Table 1). The high superliquidus temperatures were chosen to ensure complete reaction and to suppress crystallization of new oxide or silicate minerals and are not necessarily applicable to the natural process (that is, kimberlite ascent). The opx derives from a fresh sample of cratonic mantle peridotite from the Jericho kimberlite, NWT, Canada[35]. The sample was coarsely crushed and opx grains were hand-picked under a binocular microscope; an aliquot of >150 g of orthopyroxene grains was then washed, dried and finely crushed to a coarse powder (<500 μm). We then prepared batches (15–20 g) of pre-mixed starting materials comprising precisely measured masses of oven dried Na_2CO_3 powder and coarsely powdered opx; the mass fractions of the batches covered a range of opx contents from 0 to 40 wt% (Supplementary Table 1). The preparation of large batches of starting mixtures facilitated performing replicate experiments on the same bulk compositions.

Two types of experiments were performed. First, we performed transient experiments wherein a known mass of sample (Na_2CO_3 + opx) of known composition (0, 25 and 40 wt% opx) was placed in a platinum crucible and lowered from an electronic balance into a preheated furnace. The experiments were held at the controlled furnace temperature (1,000–1,100 °C) for 1–1.5 h. Each experiment provides a record of total change in mass of the crucible plus sample as a function of time. The transient experimental set-up was also used on a pure Na_2CO_3 melt in which no opx was introduced (0% opx, Supplementary Table 1). This experiment (Fig. 3b) demonstrates the long-term stability of stoichiometric amounts of dissolved CO_2 (~40 wt%) in sodium carbonate melts. Over the course of the experiment (>1 h at 1,100 °C) there is only a slight (<1.8%) cumulative change in mass, indicating that, even though performed in air, there is little to no dissociation of the melt or degassing. This stability of the Na_2CO_3 melt is also demonstrated by the static experiments (Supplementary Table 1, see below).

In the other transient experiments, the Na_2CO_3 powder melts and begins to react immediately with the mechanically mixed-in grains of opx. The Na_2CO_3 melts assimilate the opx powder creating a more silicic melt oversaturated in dissolved CO_2. The melt is forced to exsolve (effervesce) the dissolved CO_2 (that is, decarbonation), causing a continuous loss in weight (16% and 21%; Fig. 3b). The gas exsolution reaction continues until the solubility limit of the new melt composition is reached and the loss of weight with time ceases.

Interpretation of results. The transient experiments provide two insights. First, the final weight loss is a measure of the total CO_2 lost and defines the residual CO_2 in the melt; that value represents the new equilibrium CO_2 solubility of the hybrid melt. Second, the tangent to the mass loss versus time curve is a record of the relative rates of assimilation and degassing. The slope and implied rates are high initially and then decrease exponentially as the equilibrium solubility limit for the new hybrid melt is reached. The transient experiments show the spontaneous nature and rapid rate of the assimilation and decarbonation reactions; virtually all of the loss in mass occurs in the first <20 min, after which there is little change, indicating that the assimilation-induced decarbonation process is over. The experiments rapidly converge to an equilibrium state wherein all of the opx has been dissolved, the melt has reached its maximum SiO_2 content and the CO_2 solubility has been decreased as far as possible. At this point, the melt retains the amount of CO_2 dictated by its composition; the addition of more opx would foster more reaction, more weight loss, and a new more silicic melt further depleted in CO_2.

We also performed static experiments wherein a known mass of sample (Na_2CO_3 + opx) of known composition (0, 25 and 40 wt% opx) was measured into a platinum crucible, weighed, and then placed within a pre-heated oven (1,000–1,100 °C) for a fixed time (>35–75 min) (Supplementary Table 1; Fig. 3a). The sample was then removed, cooled in a desiccator, weighed and the total weight loss determined. Most experiments were run in duplicate or triplicate and several duplicates are reported in Supplementary Table 1.

Data analysis. All hybrid melt compositions are computed assuming that all available opx dissolves into carbonate melt without crystallization to create a more siliceous melt (Supplementary Table 1). Each experiment was examined visually for signs of crystallization immediately on being removed from the furnace at the end of the experimental dwell time. In every case, the crucibles contained only a clear translucent liquid. The residual CO_2 content of the melt is also calculated assuming all mass loss is due to decarbonation (that is, Na_2O volatilization is negligible[34]) and establishes the solubility limits of the hybrid melts.

These experimental results are used to predict the chemical evolution of an idealized carbonatitic melt ascending through the mantle lithosphere and preferentially assimilating opx (Fig. 4d; Supplementary Table 2). These model calculations use the following end-member compositions and relationships. We use as the parental carbonatitic melt composition KM29 from ref. 36, a melt composition produced by melting (1,430 °C, 7 GPa) of carbonated peridotite. The opx is assumed to have a single composition equivalent to that reported in sample 25-9 of ref. 35.We have modelled the CO_2 content (in wt%) of the evolving melt (W_f) using the experimentally derived (Fig. 4d) expression:

$$W_f = W_i(1 - 2.17w_{opx}) \quad (1)$$

where W_i is the CO_2 content (in wt%) of the idealized parental carbonatitic melt, and w_{opx} is the weight fraction of opx assimilated.

The range of melt compositions generated in this fashion are reported in Supplementary Table 2, where they are compared against the melt compositions estimated for the Igwisi Hills kimberlite[32].The Igwisi Hills compositions include two measurements on matrix-rich portions of kimberlite lava and two model compositions produced by correcting one of the lava compositions for the compositional effects of xenocrystic and phenocrystic olivine. For the purposes of comparison, Supplementary Table 2 reports the Igwisi Hills compositions on a reduced basis wherein MgO represents the sum of MgO and FeO. The best agreement between our model melt compositions resulting from progressive assimilation of opx and the Igwisi Hills kimberlite melt estimates occurs at between 15 and 20 wt% assimilation.

34. Hrma, P. Reaction between sodium carbonate and silica sand at 874°C < T 1022°C. *J. Am. Ceram. Soc.* **68,** 337–341 (1985).
35. Kopylova, M. G., Russell, J. K. & Cookenboo, H. Petrology of peridotite and pyroxenite xenoliths from the Jericho kimberlite: implications for the thermal state of the mantle beneath the Slave Craton, Northern Canada. *J. Petrol.* **40,** 79–104 (1999).
36. Dalton, J. A. & Presnall, D. C. Carbonatitic melts along solidus of model lherzolite in the system CaO–MgO–Al_2O_3–SiO_2–CO_2 from 3 to 7 GPa. *Contrib. Mineral. Petrol.* **131,** 123–135 (1998).

LETTER

doi:10.1038/nature10723

Recovery rates reflect distance to a tipping point in a living system

Annelies J. Veraart[1], Elisabeth J. Faassen[1], Vasilis Dakos[1], Egbert H. van Nes[1], Miquel Lürling[1,2] & Marten Scheffer[1]

Tipping points, at which complex systems can shift abruptly from one state to another, are notoriously difficult to predict[1]. Theory proposes that early warning signals may be based on the phenomenon that recovery rates from small perturbations should tend to zero when approaching a tipping point [2,3]; however, evidence that this happens in living systems is lacking. Here we test such 'critical slowing down' using a microcosm in which photo-inhibition drives a cyanobacterial population to a classical tipping point when a critical light level is exceeded. We show that over a large range of conditions, recovery from small perturbations becomes slower as the system comes closer to the critical point. In addition, autocorrelation in the subtle fluctuations of the system's state rose towards the tipping point, supporting the idea that this metric can be used as an indirect indicator of slowing down[4,5]. Although stochasticity prohibits prediction of the timing of critical transitions, our results suggest that indicators of slowing down may be used to rank complex systems on a broad scale from resilient to fragile.

Systems ranging from the brain and society to ecosystems and the climate can have tipping points at which minor perturbations can invoke a critical transition to a contrasting state[6]. The complexity of such systems prohibits accurate predictive modelling. However, it has been suggested that even without mechanistic insight, the proximity of a tipping point may be inferred from generic features of fluctuations and spatial patterns that can be interpreted as early warning indicators[1,7–9]. This idea is based on the phenomenon that at bifurcation points at which stability of an equilibrium changes, the dominant real eigenvalue becomes zero[10]. As a result, the rate of recovery from perturbation should go to zero as such bifurcations are approached (Supplementary Notes 1). This phenomenon, which is known as critical slowing down, is well established in physics but it has only recently been suggested that the recovery rate from perturbations could be an indicator of the distance to a tipping point in complex living systems such as ecosystems[3].

Although the prospect that the fragility of living systems could be probed this way is attractive, experimental evidence has so far been lacking. Instead, much work has been focused on ways to infer slowing down from indirect indicators such as autocorrelation and variance. However, although these indicators are linked to slowing down in simple stochastically forced models[1,5,9], recent theoretical studies indicate that the indirect indicators will not always respond in simple ways[11] (Brock, W. A. & Carpenter, S. R., submitted). This is confirmed by empirical studies on the climate[5], the food web of a lake[12] and laboratory populations of water fleas[13]. In these systems, trends in indirect indicators occurred but were not all consistent. Here we use a controlled system in which there is positive feedback between organisms and their physical environment to test critical slowing down directly from recovery rates.

We exposed cyanobacteria in chemostat microcosms to increasing light stress. This is a well understood system for which models have shown alternative stable states and tipping points[14,15]. Cyanobacteria provide the shade needed for their own growth, creating a positive feedback, and this constitutes the mechanism behind the bistability. Although light is needed for photosynthesis, light levels that are too high are detrimental to primary producers such as the cyanobacteria that we used. Mutual shading can ameliorate this stress, and is thus one of the ways in which facilitation can outweigh competitive interactions under harsh conditions[16]. Such feedback between organisms and their environment is the mechanism behind alternative stable states in a range of ecosystems[17]. Indeed, as a result of the facilitative shading, our system can maintain a high biomass under incident light levels that prohibit growth in low biomass cultures. A fold bifurcation that represents a classical tipping point occurs at the light level at which this mechanism becomes too weak to allow persistence of the population[14,15] (see Supplementary Notes 1 for a model analysis). Here, shading becomes insufficient to prevent growth inhibition, and the resulting loss of biomass further weakens the stabilizing shading effect. This implies that a positive feedback is driving the system towards a crash.

We cultured cyanobacteria in two independently controlled chemostat microcosms (M1 and M2) and increased incident light daily in small steps to the point at which the population collapsed (see Methods). We perturbed the populations every 4–5 days by removing 10% of their biomass through dilution. Consistent with the model results (Supplementary Fig. 1.1) the populations maintained a relatively high biomass throughout the experiment until they collapsed rather markedly when a tipping point was reached (Fig. 1 and Supplementary Notes 3). Recovery rates of both systems decreased gradually

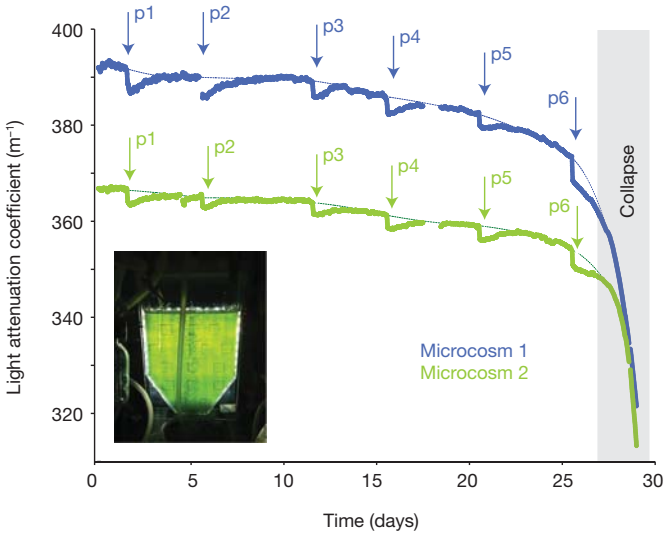


Figure 1 | The response of two populations of cyanobacteria (*Aphanizomenon flos-aquae*) to dilution events under a regime of gradually increasing light levels. Dilution events are indicated as perturbations p1–p6. The light attenuation coefficient is a measure of population density. Thin curve segments represent the baselines that were used for computing recovery rates. The inset shows the experimental system.

[1]Department of Aquatic Ecology and Water Quality Management, Wageningen University, PO Box 47, NL-6700 AA, Wageningen, The Netherlands. [2]Department of Aquatic Ecology, Netherlands Institute of Ecology, Royal Netherlands Academy of Arts and Sciences, PO Box 50, 6700AB, Wageningen, The Netherlands.

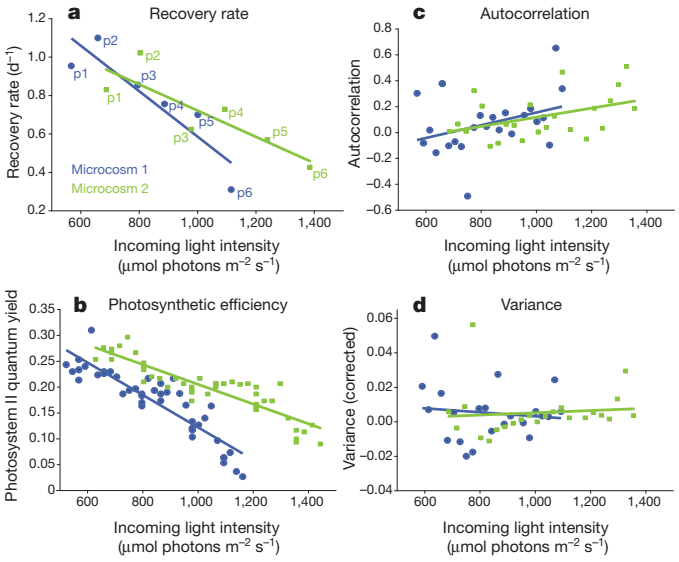


Figure 2 | Indicators of slowing down as a function of light intensity. **a**, Recovery rates after perturbation (p1–p6). **b**, Photosynthetic efficiency (photosystem II quantum yield). **c**, Autocorrelation in the population density estimator for each day based on 30 min average I_{out} data. **d**, Corrected variance in daily time series (see Methods and Supplementary Notes 4). Fisher's combined (one-tailed) significance test of the slopes of the regression line's ($n = 6$) in both microcosms: recovery rate, $\chi^2_4 = 18.48$, $P = 0.001$ (if the p6 perturbations are excluded ($n = 5$), $\chi^2_4 = 12.39$, $P = 0.015$); photosynthetic efficiency $n = 45$, $\chi^2_4 = 147.66$, $P < 0.001$; autocorrelation, $n = 23$, $\chi^2_4 = 14.00$, $P = 0.007$; variance, not significant, $n = 23$, $\chi^2_4 = 7.72$, $P = 0.102$.

towards the tipping point, starting far from the bifurcation (Fig. 2a), and tended to decline more rapidly towards the tipping point. This was also predicted by our model (Supplementary Fig. 1.1).

In most complex systems, the mechanisms that are involved in causing the slowing down will be difficult to unravel. However, in our particular system the photosynthetic capacity of the cyanobacteria that is at the heart of their growth potential can be sensed through measurements of the efficiency of their light harvesting system (see Methods). Whereas biomass remained relatively high in the trajectory towards the critical threshold, this specific indicator of their vigour declined linearly with increasing light stress, approaching zero at the point of collapse (Fig. 2b). This is an independent confirmation that the light-induced stress to the cyanobacteria in the two independently operated microcosms does indeed undermine the resilience of the system to the point at which collapse is inevitable.

In systems that are subject to stochastic perturbation regimes, slowing down is predicted to be reflected in characteristic changes in fluctuations of the state. In particular, it has been proposed that increases in autocorrelation and variance can be interpreted as indirect indicators of slowing down[1,4,9]. Although our experiment was not primarily designed to study the effect of stochastic perturbation regimes, there are continuous subtle fluctuations in our measurements of the density of the cyanobacterial population. These fluctuations will reflect a mixture of factors including measurement noise and instabilities in the lighting system as well as true population fluctuations induced by the slightly fluctuating conditions in the bubbled microcosms. We studied how autocorrelation and variance in these subtle fluctuations changed as our system approached the critical point. For this we analysed all the stretches of continuous measurements of 1 day between the experimental interruptions caused by the daily stepwise increase of light intensity and the dilution perturbation events (see Methods). Although autocorrelation was quite variable in the time series that we studied, there was a significant increase towards the tipping point in both experimental systems (Fig. 2c). No trend in variance was apparent in the systems (Fig. 2d, see also Supplementary Notes 4). These results are consistent with the prediction that measuring recovery rates from perturbations is a robust way to detect critical slowing down[18], whereas indirect indicators of slowing down may or may not increase towards a critical threshold[11]. Also, our findings are consistent with the prediction that autocorrelation is usually directly related to slowing down and may therefore provide a more robust signal than variance in some situations[11].

Perturbation experiments such as the ones in our experiment will often be impossible in large complex systems, leaving indirect indicators as the only tool by which to infer slowing down. However, experimental probing of recovery rates may be feasible in some smaller systems, as long as the timescales are appropriate and stochastic fluctuations are small relative to experimental perturbations. Even in larger systems, local perturbations may be an option to probe resilience, allowing adaptive management to steer the system away from the brink of collapse.

Perhaps most importantly, our experimental demonstration of slowing down implies a proof of concept, providing a fundamental basis to the current search for generic early warning signals in systems ranging from the brain and ecosystems to society and the climate[1]. The fact that slowing down in our system started far from the critical point suggests that recovery rates as well as indirect indicators may be used to rank such complex systems on a broad scale from stable to critical. This does not mean that slowing down can be used to actually predict transitions. Stochastic shocks will trigger critical transitions always before the bifurcation point is reached, indicating that there is inherent unpredictability in systems. Nevertheless, the prospect of having generic indicators of resilience is a potentially large step forward. Mechanistic models to predict tipping points in nature and society accurately are simply beyond our reach, leaving empirical estimation of fragility as one of the key challenges in complex systems science today[1].

METHODS SUMMARY

Experiments were performed in two identical flat chemostat microcosms[19] (M1 and M2) in which we cultured cyanobacteria (*Aphanizomenon flos-aquae* (L.) Ralfs) on a nutrient-rich growth medium, modified from BG11 medium[20]. Light irradiance was increased in steps of 23 μmol photons m^{-2} s^{-1} for M1 and 29 μmol photons m^{-2} s^{-1} each day for M2. Photosynthetic efficiency was measured from diurnal samples. The intensity of the light passing through the chemostats was averaged at 5-min intervals, and light attenuation coefficients (ε, m^{-1}) were calculated as an indicator of biomass:

$$\varepsilon = \frac{-\ln(I_{out}/I_{in})}{d}$$

where I_{in} is the intensity of the incoming light (μE), I_{out} is the intensity of the outgoing light (μE) and d is the depth of the chemostats (m). External perturbations were performed every 4–5 days by diluting the culture with 170 ml of sterile medium. A baseline for calculating recovery rates was constructed for each perturbation event (Supplementary Notes 2 and Supplementary Fig. 2.1) by fitting a quadratic curve from the period just before perturbation to the period just before the next perturbation (thin curves in Fig. 1). Recovery rates after each perturbation (λ, *per day*) were then calculated from a linear regression of $-\ln(\varepsilon_0 - \varepsilon_c)$ against time, where ε_0 is the light attenuation of the baseline (m^{-1}) and ε_c is the light attenuation of the chemostat (m^{-1})[3].

The lag 1 autocorrelation and variance were analysed for each uninterrupted period between the daily manipulations after removing the trends from each period by fitting polynomials of 2 degrees. To check for effects of nonlinear propagation of measurement noise, we constructed a null model that assumed that all residuals were due to uncorrelated normally distributed noise. As variance showed a trend towards the bifurcation in this null model (Supplementary Table 4.1), we corrected the observed variance by subtracting the median of the null model (Fig. 2d).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 September; accepted 18 November 2011.
Published online 25 December 2011.

1. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).

2. Wissel, C. A universal law of the characteristic return time near thresholds. *Oecologia* **65,** 101–107 (1984).
3. van Nes, E. H. & Scheffer, M. Slow recovery from perturbations as a generic indicator of a nearby catastrophic shift. *Am. Nat.* **169,** 738–747 (2007).
4. Ives, A. R. Measuring resilience in stochastic systems. *Ecol. Monogr.* **65,** 217–233 (1995).
5. Dakos, V. *et al.* Slowing down as an early warning signal for abrupt climate change. *Proc. Natl Acad. Sci. USA* **105,** 14308–14312 (2008).
6. Scheffer, M. *Critical Transitions in Nature and Society* (eds Levin, S.A. & Strogatz, S.H.) (Princeton Univ. Press, 2009).
7. Kleinen, T., Held, H. & Petschel-Held, G. The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. *Ocean Dyn.* **53,** 53–63 (2003).
8. Held, H. & Kleinen, T. Detection of climate system bifurcations by degenerate fingerprinting. *Geophys. Res. Lett.* **31,** L23207 (2004).
9. Carpenter, S. R. & Brock, W. A. Rising variance: a leading indicator of ecological transition. *Ecol. Lett.* **9,** 311–318 (2006).
10. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering* 1st edn (Addison-Wesley, 1994).
11. Dakos, V., van Nes, E. H., D' Odorico, P. & Scheffer, M. How robust are variance and autocorrelation as early-warning signals for critical transitions? *Ecology* (submitted); preprint at http://dx.doi.org/10.1890/11-0889.1.
12. Carpenter, S. R. *et al.* Early warnings of regime shifts: A whole-ecosystem experiment. *Science* **332,** 1079–1082 (2011).
13. Drake, J. M. & Griffen, B. D. Early warning signals of extinction in deteriorating environments. *Nature* (2010).
14. Huisman, J. *The Struggle for Light.* PhD thesis, Univ. Groningen (1997).
15. Gerla, D. J., Mooij, W. M. & Huisman, J. Photoinhibition and the assembly of light-limited phytoplankton communities. *Oikos* **120,** 359–368 (2011).
16. Holmgren, M., Scheffer, M. & Huston, M. A. The interplay of facilitation and competition in plant communities. *Ecology* **78,** 1966–1975 (1997).
17. Scheffer, M., Carpenter, S. R., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413,** 591–596 (2001).
18. Dakos, V., Kéfi, S., Rietkerk, M., van Nes, E. H. & Scheffer, M. Slowing down in spatially patterned ecosystems at the brink of collapse. *Am. Nat.* **177,** E153–E166 (2011).
19. Huisman, J. *et al.* Principles of the light-limited chemostat: theory and ecological applications. Antonie Leeuwenhoek **81,** 117–133 (2002).
20. Andersen, R. A. Berges, J.A., Harrison, P.J. & Watanabe M.M. in *Algal culturing techniques* 1st edn 435–436 (Elsevier, 2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. B. Gonçalves Souza for discussions on the experimental set up. We thank D. Waasdorp and W. Beekman-Lukassen for assistance with the experiments and C. ter Braak for statistical advice. A.J.V., E.J.F., V.D., E.H.v.N. and M.S. are supported by a European Research Council Advanced grant and M.S. is the recipient of a Spinoza award.

Author Contributions A.J.V., E.J.F. and M.L. performed the experiments. A.J.V., E.J.F, E.H.v.N. and V.D. analysed the data. M.S., A.J.V., E.H.v.N., E.J.F. and V.D. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.H.v.N. (egbert.vannes@wur.nl).

METHODS

Experimental conditions. Experiments were performed in two identical flat chemostats (V 1.7 l, 0.05 m optical path length)[19]. In these chemostats we cultured the cyanobacterium *Aphanizomenon flos-aquae* (L.) Ralfs on a nutrient-rich sterile growth medium that was modified from BG11 medium[20]. The chemostats were kept at a stable temperature of 21 °C. A continuous flow of moistened air of 60–100 ml min^{-1} was supplied through a sintered glass sieve at the bottom of the vessel to ensure homogenous mixing of the culture. The air was mixed with CO_2 to satisfy the inorganic carbon need of the culture. CO_2 flow was adjusted, when needed, to maintain a pH of between 7.1 and 8.1. The chemostats were run at a dilution rate of 0.18 per day for chemostat 1 (named M1) and 0.21 per day for chemostat 2 and (M2). They were illuminated using white LED lamps (SL3500w, Photon Systems Instruments). Light irradiance was increased by 23 μmol photons m^{-2} s^{-1} per day for M1 and 29 μmol photons m^{-2} s^{-1} per day for M2 by a Light Studio 1.3 12C interface (Photon Systems Instruments, Brno).

Daily maintenance and measurements. Each day the walls of the chemostats were scraped with a magnetic stirrer to prevent cyanobacterial attachment. After scraping, we took samples to determine chlorophyll *a* concentrations and photosystem II quantum yield (in triplicate) using a PhytoPAM (phyto-ED), and to determine biovolume in a 400-μl sample volume (in triplicate) using a Casy TT Cell Counter, with a 150-μm capillary (Innovatis AG Casy Technology). The intensity of the light penetrating through the chemostat was recorded continuously using RA100 light sensors (Bottemanne Weather Instruments) that were attached to the outer wall of the chemostat and stored as 5-min averages on Squirell SQ1000 dataloggers (Grant Instruments). The light sensors were removed during scraping of the chemostats.

Perturbations. At an incoming light intensity of 477 μmol photons m^{-2} s^{-1} for M1 and 571 for M2, the light attenuation coefficients of the chemostats became stable. From this moment on, perturbations were performed every 4–5 days by diluting the culture with 170 ml of sterile medium. The dilution was always performed 2 h after the daily stepwise increase in light.

Recovery rates. We used the calculated light attenuation (Fig. 1 and Supplementary Note 2) as a measure of the cyanobacterial biomass[21]. Before calculation of recovery rates, the light data were corrected for sensor attachment differences (Supplementary Note 2). Vertical light attenuation (ε, m^{-1}) was calculated from the corrected light data:

$$\varepsilon = \frac{-\ln(I_{out}/I_{in})}{d}$$

where I_{in} is the intensity of the incoming light, I_{out} is the intensity of the outgoing light (both measured in μmol photons m^{-2} s^{-1}) and d is the optical path length of the chemostats (m). Light attenuation data were smoothed for calculation of recovery rates by taking a moving average of 1 h.

To calculate recovery rates, a baseline was constructed for each perturbation event. This baseline was obtained by fitting a quadratic curve from the period just before perturbation to the period just before the next perturbation (Fig. 1). Parameters for the baseline were estimated by forcing it through the sets of ε and t (time, day) at the start and end of the curve, and by forcing the slope at the start of the curve. The start slope was determined by the slope of the light attenuation data in the 20 h before disturbance.

Recovery rate after perturbation (λ, per day) was defined by an exponential model[3]:

$$\frac{d\varepsilon}{dt} = -\lambda(\varepsilon_0 - \varepsilon_c)$$

where ε_0 is the light attenuation coefficient of the baseline and ε_c is the light attenuation of the chemostat (both m^{-1}). We calculated λ by a linear regression of $-\ln(\varepsilon_0 - \varepsilon_c)$ against time. To avoid the effect of the light data correction for sensor position (Supplementary Note 2), recovery rates were calculated only on the first 18–20 h after perturbation. In this period there was no change in light meter position. Finally, the recovery rates were linearly regressed against incoming light intensity.

Autocorrelation and variance. Autocorrelation and variance of the continuous small fluctuations in our time series were analysed for each uninterrupted period between the daily manipulations and light increments. We performed all analyses on untransformed data (5-min averages of light attenuation data) as well as on data that were averaged over non-overlapping periods of 30 min. We removed the trends from each period with a constant light level by fitting polynomials of 2 degrees to the light attenuation, and we used the residuals to calculate the autocorrelation by fitting an autoregressive model of lag 1 and variance by estimating sample variance per day.

We analysed the effect of measurement noise using a null model (see Supplementary Notes 4).

21. Huisman, J. & Weissing, F. J. Light-limited growth and competition for light in well-mixed aquatic environments: an elementary model. *Ecology* **75,** 507–520 (1994).

LETTER

doi:10.1038/nature10724

Evolution of increased complexity in a molecular machine

Gregory C. Finnigan[1]*, Victor Hanson-Smith[2,3]*, Tom H. Stevens[1] & Joseph W. Thornton[2,4,5]

Many cellular processes are carried out by molecular 'machines'—assemblies of multiple differentiated proteins that physically interact to execute biological functions[1–8]. Despite much speculation, strong evidence of the mechanisms by which these assemblies evolved is lacking. Here we use ancestral gene resurrection[9–11] and manipulative genetic experiments to determine how the complexity of an essential molecular machine—the hexameric transmembrane ring of the eukaryotic V-ATPase proton pump—increased hundreds of millions of years ago. We show that the ring of Fungi, which is composed of three paralogous proteins, evolved from a more ancient two-paralogue complex because of a gene duplication that was followed by loss in each daughter copy of specific interfaces by which it interacts with other ring proteins. These losses were complementary, so both copies became obligate components with restricted spatial roles in the complex. Reintroducing a single historical mutation from each paralogue lineage into the resurrected ancestral proteins is sufficient to recapitulate their asymmetric degeneration and trigger the requirement for the more elaborate three-component ring. Our experiments show that increased complexity in an essential molecular machine evolved because of simple, high-probability evolutionary processes, without the apparent evolution of novel functions. They point to a plausible mechanism for the evolution of complexity in other multi-paralogue protein complexes.

Comparative genomic approaches suggest that the components of many molecular machines have appeared sequentially during evolution and that complexity increased gradually by incorporating new parts into simpler assemblies[2–8]. Such horizontal analyses of extant systems, however, cannot decisively test these hypotheses or reveal the mechanisms by which additional parts became obligate components of larger complexes. In contrast, vertical approaches that combine computational phylogenetic analysis with gene synthesis and molecular assays allow changes in the sequence, structure and function of reconstructed ancestral proteins to be experimentally traced through time.[9–11] Here we apply this approach to characterize the evolution of a small molecular machine and dissect the mechanisms that caused it to increase in complexity.

The vacuolar H^+-ATPase (V-ATPase) is a multisubunit protein complex that pumps protons across membranes to acidify subcellular compartments; this function is required for intracellular protein trafficking, coupled transport of small molecules and receptor-mediated endocytosis[1]. V-ATPase dysfunction has been implicated in human osteoporosis, in acquired drug resistance in human tumours, and in pathogen virulence[12–14]. A key subcomplex of the V-ATPase is the V_0 protein ring, a hexameric assembly that uses a rotary mechanism to move protons across organelle membranes (Fig. 1a)[15,16]. Although the V-ATPase is found in all eukaryotes, the V_0 ring varies in subunit composition among lineages. In animals and most other eukaryotes,

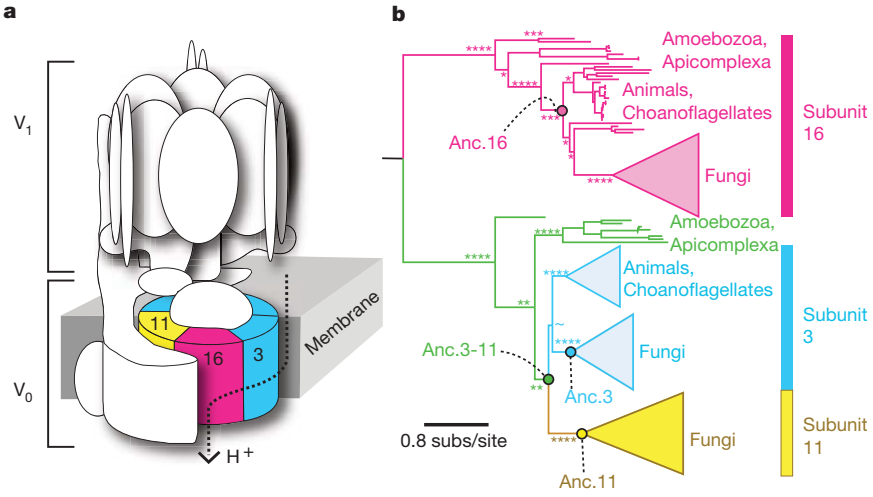


Figure 1 | Structure and evolution of the V-ATPase complex. a, In *S. cerevisiae*, the V-ATPase contains two subcomplexes: the octameric V_1 domain is on the cytosolic side of the organelle membrane, and the hexameric V_0 ring is membrane bound. Protein subunits Vma3, Vma11 and Vma16 are labelled and coloured. **b**, Maximum likelihood phylogeny of V-ATPase subunits Vma3, Vma11 and Vma16. All eukaryotes contain subunits 3 and 16, but Fungi also contain subunit 11. Circles show ancestral proteins reconstructed in this study. Colours correspond to those of subunits in panel **a**; unduplicated orthologues of Vma3 and Vma11 are green. Asterisks show approximate likelihood ratios for major nodes: ****, $>10^3$; ***, $>10^2$; **, >10; *, <10; ~, <2. The complete phylogeny is presented in Supplementary Information, section 2.

[1]Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA. [2]Institute for Ecology and Evolution, University of Oregon, Eugene, Oregon 97403, USA. [3]Department of Computer and Information Science, University of Oregon, Eugene, Oregon 97403, USA. [4]Howard Hughes Medical Institute, Eugene, Oregon 97403, USA. [5]Departments of Human Genetics and Ecology & Evolution, University of Chicago, Chicago, Illinois 60637, USA.
*These authors contributed equally to this work.

the ring consists of one subunit of Vma16 protein and five copies of its paralogue, Vma3 (Fig. 1b)[1]. In Fungi, the ring consists of one Vma16 subunit, four copies of Vma3 and one Vma11 subunit, arranged in a specific orientation[17]. All three proteins are required for V-ATPase function in Fungi[18,19], but the mechanisms are unknown by which both Vma3 and Vma11 became obligate components with specific positional roles in the complex.

To understand how the three-component ring evolved, we reconstructed ancestral V_0 proteins from just before and after the increase in complexity, synthesized and functionally characterized them in a yeast genetic system, and used manipulative methods to identify the genetic and molecular mechanisms by which their functions changed. We first inferred the phylogeny and best-fit evolutionary model of the protein family of which Vma3, Vma11 and Vma16 are members, using the sequences of all 139 extant family members available in GenBank (Supplementary Table 1). The maximum likelihood phylogeny (Fig. 1b and Supplementary Information, section 2) indicates that Vma3 and Vma11 are sister proteins that were produced by duplication of an ancestral gene (Anc.3-11) before the last common ancestor of all Fungi (~800 million years ago[20]). Whether this duplication occurred before or after the divergence of Fungi from other eukaryotes (~1 billion years ago[20]) is not clearly resolved, although the latter scenario is more parsimonious. The Vma3/Vma11 and Vma16 lineages, in turn, descend from an older gene duplication deep in the eukaryotic lineage (Fig. 1b). We used a maximum likelihood algorithm[21] to infer the ancestral amino acid sequences with the highest probability of producing all the extant sequence data, given the best-fit phylogeny and model. We reconstructed the ancestral proteins (Anc.3-11 and Anc.16) that made up the ancient two-paralogue eukaryotic ring, as well as the duplicated subunits Anc.3 and Anc.11 from the three-component ring in the common ancestor of all Fungi (Supplementary Information, sections 3 and 4).

To characterize the functions of these reconstructed proteins, we synthesized coding sequences and transformed them into *Saccharomyces cerevisiae* deficient for various ring components and therefore incapable of growth in the presence of elevated $CaCl_2$ (ref. 22). We found that the ancestral two-subunit ring can functionally replace the three-subunit ring of extant yeast. When the resurrected Anc.3-11 was transformed into yeast deficient for Vma3 (*vma3Δ*) or Vma11(*vma11Δ*), growth in the presence of elevated $CaCl_2$ was rescued, indicating that the functions of the present-day Vma3 and Vma11 proteins were already present before the duplication that generated them (Fig. 2a). Furthermore, Anc.3-11—unlike either of its present-day descendants—can partially rescue growth in yeast that are doubly deficient for both Vma3 and Vma11 (*vma3Δ vma11Δ*). The reconstructed Anc.16 also rescued growth in Vma16-deficient *S. cerevisiae* (*vma16Δ*) (Fig. 2b), and co-expression of Anc.3-11 and Anc.16 together rescued cell growth in *vma3Δ vma11Δ vma16Δ* yeast, which lack all three ring subunits (Fig. 2c). The ancestral genes specifically restore proper V-ATPase function in acidification of the vacuolar lumen (Fig. 2g). In addition, mutation of the ancestral subunits to remove glutamic acid residues known to be essential for V-ATPase enzyme function[17,23] abolished their ability to rescue growth on $CaCl_2$ (Supplementary Information, section 7). These inferences about the functions of Anc.3-11 and Anc.16 are robust to uncertainty about ancestral amino acid states. We reconstructed alternative versions of Anc.3-11 and Anc.16 by introducing amino acid states with posterior probability >0.2, but none of these abolished the ability of the ancestral genes to substitute functionally for the extant subunits (Supplementary Information, section 8). These results establish that during the increase in complexity, neither the V_0 complex nor its component proteins evolved new functions required for growth under the conditions in which the ring is known to be important.

Similar experiments with the components of the ancestral three-component ring show that after the duplication of Anc.3-11, its descendants Anc.3 and Anc.11 both became necessary for a functional complex because of complementary losses of ancestral functions. Unlike Anc.3-11, expression of Anc.3 can rescue growth and vacuole acidification in *vma3Δ* but not *vma11Δ* yeast, and Anc.11 can rescue growth in *vma11Δ* but not *vma3Δ* yeast (Fig. 2d, e, g). Furthermore, both Anc.3 and Anc.11 are required to rescue growth fully in *vma3Δ vma11Δ* yeast (Fig. 2f). These data indicate that after its origin by gene duplication, Anc.11 lost the ancestral protein's ability to carry out at least some functions of Vma3, and Anc.3 lost the ancestral capacity to carry out those of Vma11.

We conjectured that Vma3 and Vma11 evolved their specialized roles because they lost specific interfaces present in their ancestor that are required for ring assembly. Previous experiments with fusions of

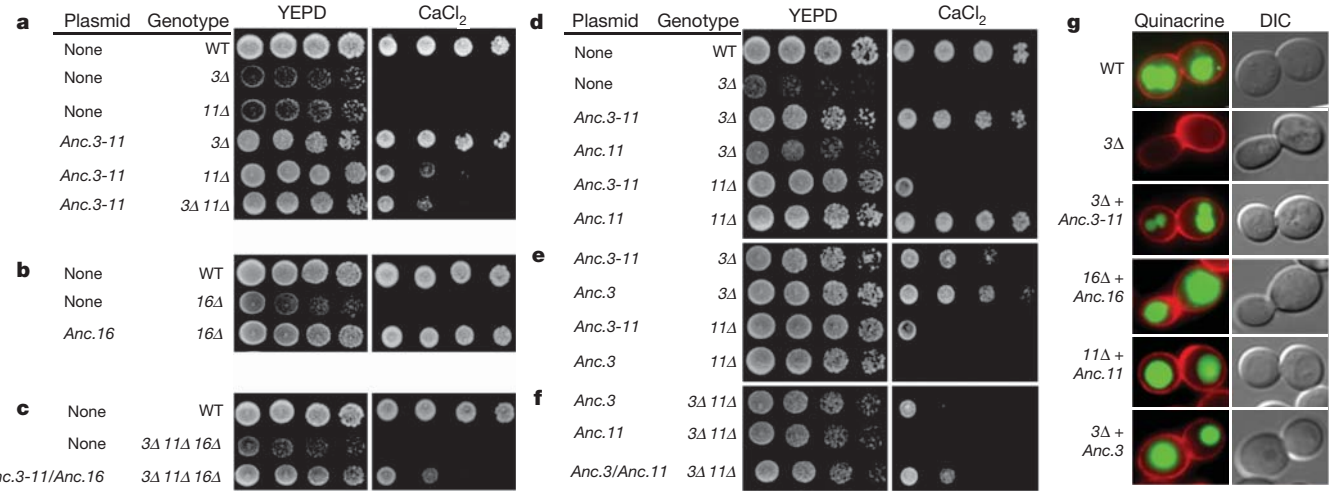


Figure 2 | Two reconstructed ancestral V_0 subunits functionally replace the three-paralogue ring in extant yeast. *S. cerevisiae* were plated in decreasing concentrations on permissive medium (YEPD) buffered with elevated $CaCl_2$. **a**, Expression of Anc.3-11 rescues growth in yeast that are deficient for endogenous subunit Vma3 (*3Δ*), subunit Vma11 (*11Δ*) or both (*3Δ 11Δ*). Growth of wild-type (WT) yeast is shown for comparison. **b**, Anc.16 rescues growth in yeast that are deficient for subunit Vma16 (*16Δ*). **c**, Expression of Anc.3-11 and Anc.16 together rescues growth in yeast that are deficient for Vma3, Vma11 and Vma16. **d**, Anc.11 rescues growth in *vma11Δ* but not in *vma3Δ* yeast. **e**, Anc.3 rescues growth in *vma3Δ* but not *vma11Δ* yeast. **f**, Anc.3 and Anc.11 together rescue growth in *vma3Δ vma11Δ* mutants. **g**,Yeast expressing reconstructed ancestral subunits properly acidified the vacuolar lumen. Red signal shows yeast cell walls; green signal (quinacrine) shows acidified compartments. Yeast were visualized by differential interference contrast microscopy.

extant yeast proteins have shown that the arrangement of subunits in the ring is constrained by the capacity of each subunit to form specific interfaces (which we labelled P, Q and R) with the other subunits[24]. Specifically, Vma11 is restricted to a single position between Vma16 and Vma3, because its clockwise interface can participate only in interface R with Vma16, and its anticlockswise interface can participate only in interface P with the clockwise side of Vma3 (Fig. 3). By contrast, copies of Vma3 occupy several positions in the ring, because they form interface P with other copies of Vma3 or Vma11, as well as interface Q with Vma16. However, Vma3 cannot form interface R with Vma16. As a result, both Vma3 and Vma11 are required in extant yeast to form a complete ring with Vma16.

To determine whether interaction interfaces were lost during evolution, we engineered fusions of ancestral ring proteins to assess the capacity of each to form the specific interfaces with other subunits that are required for a functional complex. Because Anc.3-11 can complement the loss of both Vma3 and Vma11, we proposed that the Anc.3-11 subunit could participate in all three specific interaction interfaces, and that these capacities were then partitioned between Anc3 and Anc11 after the duplication of Anc.3-11 (Fig. 3a, b). To test this hypothesis, we created six reciprocal gene fusions between yeast subunit Vma16 and ancestral subunits Anc.3-11, Anc.3 and Anc.11 (Fig. 3c and Supplementary Information, section 9). Each fusion constrains the structural position of subunits relative to subunit Vma16, making it possible to determine which arrangements yield a functional ring. As predicted, Anc.3-11 functioned on either side of Vma16 (Fig. 3d), indicating that it could form all three interfaces P, Q and R. By contrast, Anc.3 functioned when constrained to participate in interface Q with Vma16 and interface P with Vma3; however, ring function was lost when Anc.3 was constrained to form interface R with Vma16 (Fig. 3e). Anc.11 functioned when constrained to participate in interface R with Vma16 and interface P with Vma3, but ring function was lost when Anc.11 was constrained to participate in interface Q with Vma16 and interface P with Vma3. This result indicates that Anc.11 lost the capacity to form one or both of these interfaces during its post-duplication divergence from Anc.3-11 (Fig. 3f).

Taken together, these data indicate that the specificity of the ring arrangement and the obligate roles of Vma3 and Vma11 evolved by complementary loss of asymmetric interactions with other members of the ring (Fig. 3g, h). Before Anc.3-11 duplicated, the protein ring contained copies of only undifferentiated subunit Vma3/Vma11 and subunit 16. Immediately after Anc.3-11 duplicated, the two descendant subunits must have been functionally identical, so the protein ring could have assembled with many possible combinations of the two descendants, including copies of only one of the descendant proteins. This flexibility disappeared when Anc.3 lost the ancestral interface that allowed it to interact with the anticlockwise side of Vma16, and Anc.11 lost the ability to interact with the clockwise side of Vma16 and/or the anticlockwise side of Vma3. These complementary losses are sufficient to explain the specific arrangement of contemporary subunits in reconstructed and present-day fungal transmembrane rings.

To establish the genetic basis for the partitioning of the functions of Anc.3-11 between Vma3 and Vma11, we introduced historical mutations into Anc3.11 by directed mutagenesis and determined whether they recapitulated the shifts in function that occurred during the evolution of Anc.3 and Anc.11. The two phylogenetic branches leading from Anc.3-11 to Anc.3 and to Anc.11 contain 25 and 31 amino acid substitutions, respectively, but only a subset of these are strongly conserved in subunits Vma3 or Vma11 from extant Fungi (Fig. 4a). We introduced each of these 'diagnostic' substitutions into Anc.3-11 and

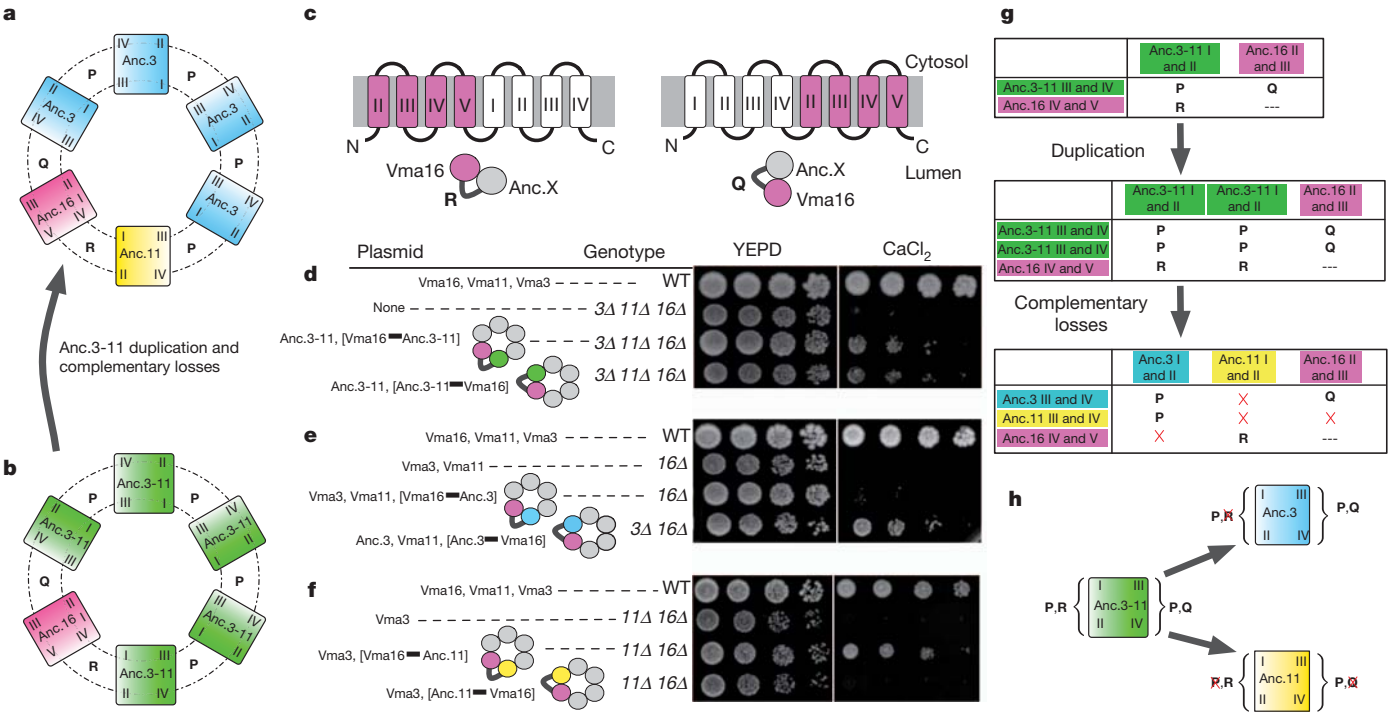


Figure 3 | Increasing complexity by complementary loss of interactions in the fungal V_0 ring. **a**, Model of the ancestral three-paralogue ring, arranged as in extant yeast[24]. Unique intersubunit interfaces are labelled P, Q and R. Subunits are colour-coded as in Fig. 1. **b**, Model of the ancestral two-paralogue ring, before duplication of Anc.3-11. **c**, To constrain the location of specific subunits, gene fusions were constructed by tethering an ancestral subunit to either the amino- or carboxy-terminal side of yeast Vma16. Roman numerals indicate the locations of transmembrane helices (I, II, III, IV and V)[24]. **d**–**f**, Growth assays of yeast with fused V_0 subunits identify the interfaces that ancestral subunits can form. For each experiment, expressed V_0 subunits are listed. Tethered subunits are in brackets and connected by a thick line. Cartoons show the constrained location of the tethered subunit relative to Vma16. Anc.3-11 can function on either side of Vma16 (**d**). Anc.3 can function only on the clockwise side of Vma16 (**e**). Anc.11 can function only on the anticlockwise side of Sc.16 (**f**). **g**, Interfaces that are formed by V_0 subunits before and after duplication and complementary loss of interfaces, based on the data in panels **d**–**f**. Red crosses indicate lost interfaces. **h**, Schematic of interfaces formed by Anc.3-11 that were lost in Anc.3 and Anc.11, based on data in panels **d**–**f**.

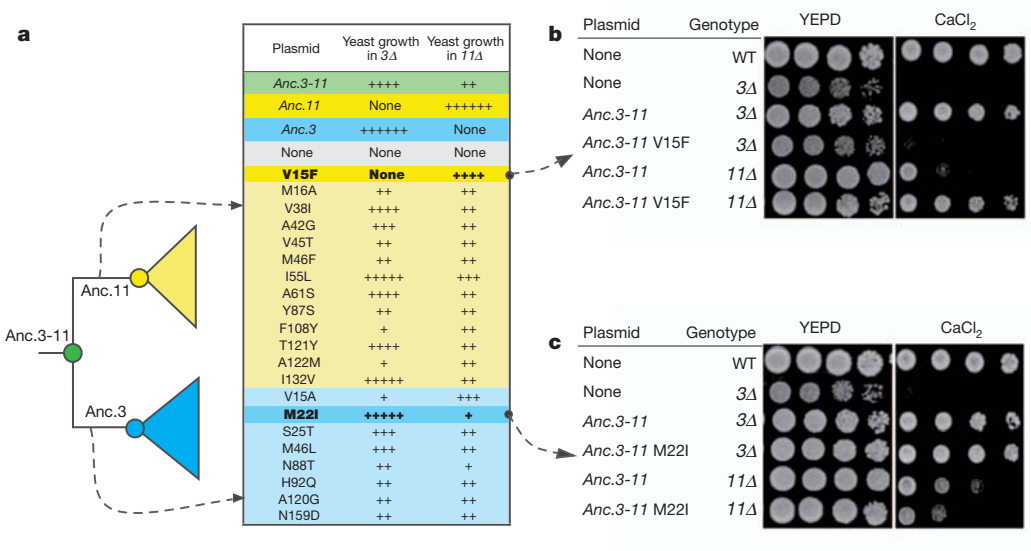


Figure 4 | Genetic basis for functional differentiation of Anc.3 and Anc.11. **a**, Experimental analysis of historical amino acid replacements. The table lists replacements that occurred on the branches leading from Anc.3-11 to Anc.11 (yellow) or to Anc.3 (blue) and that were subsequently conserved. Each derived residue was introduced singly into Anc.3-11; the variant genes were transformed into *S. cerevisiae*, and growth was assayed on elevated $CaCl_2$. The table shows growth semiquantitatively from zero (none) to wild type (++++++). Bold mutations entirely or partly recapitulate the functional evolution of Anc.11 and Anc.3. **b**, Replacement V15F abolishes the capacity of Anc.3-11 to function as subunit 3 and enhances the capacity of Anc.3-11 to function as subunit 11. **c**, Replacement M22I impairs the capacity of Anc.3-11 to function as subunit 11 without affecting its capacity to function as subunit 3.

experimentally evaluated whether they recapitulated the loss by Anc.3 or Anc.11 of the capacity to complement Vma gene deletions. We found that a single amino-acid replacement that occurred on the branch leading to Anc.11 (V15F) abolished the capacity of Anc.3-11 to function as subunit 3; it also enhanced the ability of Anc.3-11 to function as subunit 11 (Fig. 4b). V15F is located in transmembrane helix I, which participates in the P interface that our experiments indicate may have been lost on the same branch (Fig. 3 and Supplementary Information, section 4). Conversely, a single historical replacement (M22I) on the branch leading to Anc.3 radically reduced the capacity of Anc.3-11 to function as subunit 11 (Fig. 4c). M22I is also in transmembrane helix I, which participates in formation of the R interface that was lost on this branch (Fig. 3 and Supplementary Information, section 4). The Anc.3-11 M22I mutant retains some of the capacity of the ancestral protein to rescue growth in the Vma11-deficient background, suggesting that other mutations also contributed to the functional evolution of Vma3. One other historical mutation (N88T) on this branch also impaired the capacity of Anc.3-11 to function as subunit 11, but it reduced the capacity of the protein to function as Vma3 as well, suggesting that epistatic interactions with other residues allow this mutation to be tolerated in Anc.3 and its descendants. Several of the replacements on the branch leading to Anc.11 show a similar pattern, reducing the capacity of the protein to replace Vma3, indicating that these historical replacements function better together than in isolation.

How complexity and novel functions evolve has been a longstanding question in evolutionary biology[25–27], because mutations that compromise existing functions are far more frequent than those that generate new ones[28]. Our results indicate that the architectural complexity of molecular assemblies can evolve because of a few simple, relatively high-probability mutations that degrade ancestral interfaces but leave other functions intact. The specific roles of subunits Vma3 and Vma11 seem to have been acquired when duplicated genes lost some, but not all, of the capacity of the ancestral protein to participate in interactions with copies of itself and another protein required for proper ring assembly. Because complementary losses occurred in both lineages, the two descendant subunits became obligate components, and the complexity of the ring increased. It is possible that specialization of the duplicated subunits allowed increases in fitness, but genome-wide interaction screens and the phenotype of *vma11*Δ yeast provide no evidence that Vma11 evolved novel functions in addition to those that it inherited from Anc.3-11 in the V_0 ring[29].

We are aware of no other mechanistic analyses of a molecular machine's evolutionary trajectory, so the generality of our observations is unknown. By definition, however, all molecular machines involve differentiated parts in specific spatial orientations, and many such complexes are entirely or partially composed of paralogous proteins[2–8]. In the evolution of any such assembly, additional paralogues could become obligate components because of gene duplication[30] and subsequent mutations that cause specific interaction interfaces among them to degenerate.

This view of the evolution of molecular machines is related to recent models that explain other biological phenomena—such as the retention of large numbers of duplicate genes and mobile genetic elements within genomes—as the product of degenerative processes acting on modular biological systems[27]. Although mutations that enhanced the functions of individual ring components may have occurred during evolution, our data indicate that simple degenerative mutations are sufficient to explain the historical increase in complexity of a crucial molecular machine. There is no need to invoke the acquisition of 'novel' functions caused by low-probability mutational combinations.

METHODS SUMMARY

Ancestral protein sequences were inferred using maximum-likelihood phylogenetics from an alignment of 139 protein sequences of extant subunits 3, 11 and 16 from Amoebozoa, Apicomplexa, Metazoa, Choanoflagellida and Fungi. Ancestral genes were synthesized, cloned into yeast expression vectors and tested for complementation in various *S. cerevisiae* mutants. V-ATPase function was assayed by growth tests on medium buffered with $CaCl_2$, as described previously[31]. Steady-state levels of Vph1 were determined by western blot. Quinacrine staining and Vph1–GFP (green fluorescent protein) fusion constructs were visualized by fluorescence microscopy.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 September; accepted 21 November 2011.
Published online 9 January 2012.

1. Forgac, M. Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. *Nature Rev. Mol. Cell Biol.* **8,** 917–929 (2007).

2. Pallen, M. J. & Matzke, N. J. From the origin of species to the origin of bacterial flagella. *Nature Rev. Microbiol.* **4,** 784–790 (2006).
3. Liu, R. & Ochman, H. Stepwise formation of the bacterial flagellar system. *Proc. Natl Acad. Sci. USA* **104,** 7116–7121 (2007).
4. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nature Rev. Microbiol.* **5,** 892–899 (2007).
5. Dolezal, P., Likic, V., Tachezy, J. & Lithgow, T. Evolution of the molecular machines for protein import into mitochondria. *Science* **313,** 314–318 (2006).
6. Clements, A. *et al.* The reducible complexity of a mitochondrial molecular machine. *Proc. Natl Acad. Sci. USA* **106,** 15791–15795 (2009).
7. Archibald, J. M., Logsdon, J. M. Jr & Doolittle, W. F. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol. Biol. Evol.* **17,** 1456–1466 (2000).
8. Gabaldón, T., Rainey, D. & Huynen, M. A. Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (complex I). *J. Mol. Biol.* **348,** 857–870 (2005).
9. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Rev. Genet.* **5,** 366–375 (2004).
10. Liberles, D. (ed.) *Ancestral Sequence Reconstruction* (Oxford Univ. Press, 2007).
11. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **20,** 360–366 (2010).
12. Frattini, A. *et al.* Defects in TCIRG1 subunit of the vacuolar proton pump are responsible for a subset of human autosomal recessive osteopetrosis. *Nature Genet.* **25,** 343–346 (2000).
13. Pérez-Sayáns, M., Somoza-Martìn, J. M., Barros-Angueira, F., Rey, J. M. & Garcìa-Garcìa, A. V-ATPase inhibitors and implication in cancer treatment. *Cancer Treat. Rev.* **35,** 707–713 (2009).
14. Xu, L. *et al.* Inhibition of host vacuolar H^+-ATPase activity by a *Legionella pneumophila* effector. *PLoS Pathog.* **6,** e1000822 (2010).
15. Hirata, T. *et al.* Subunit rotation of vacuolar-type proton pumping ATPase: relative rotation of the g and c subunits. *J. Biol. Chem.* **278,** 23714–23719 (2003).
16. Imamura, H. *et al.* Rotation scheme of V_1-motor is different from that of F_1-motor. *Proc. Natl Acad. Sci. USA* **102,** 17929–17933 (2005).
17. Powell, B., Graham, L. A. & Stevens, T. H. Molecular characterization of the yeast vacuolar H^+-ATPase proton pore. *J. Biol. Chem.* **275,** 23654–23660 (2000).
18. Umemoto, N., Yoshihisa, T., Hirata, R. & Anraku, Y. Roles of the *VMA3* gene product, subunit c of the vacuolar membrane H^+-ATPase on vacuolar acidification and protein transport. A study with *VMA3*-disrupted mutants of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **265,** 18447–18453 (1990).
19. Umemoto, N., Ohya, Y. & Anraku, Y. *VMA11*, a novel gene that encodes a putative proteolipid, is indispensable for expression of yeast vacuolar membrane H^+-ATPase activity. *J. Biol. Chem.* **266,** 24526–24532 (1991).
20. Taylor, J. W. & Berbee, M. L. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia* **98,** 838–849 (2006).
21. Yang, Z., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141,** 1641–1650 (1995).
22. Kane, P. M. The where, when, and how of organelle acidification by the yeast vacuolar H^+-ATPase. *Microbiol. Mol. Biol. Rev.* **70,** 177–191 (2006).
23. Hirata, R., Graham, L. A., Takatsuki, A., Stevens, T. H. & Anraku, Y. Vma11 and vma16 encode second and third proteolipid subunits of the *Saccharomyces cerevisiae* vacuolar membrane H^+-ATPase. *J. Biol. Chem.* **272,** 4795–4803 (1997).
24. Wang, Y., Cipriano, D. J. & Forgac, M. Arrangement of subunits in the proteolipid ring of the V-ATPase. *J. Biol. Chem.* **282,** 34058–34065 (2007).
25. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
26. Jacob, F. Evolution and tinkering. *Science* **196,** 1161–1166 (1977).
27. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA* **104,** 8597–8604 (2007).
28. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. Experimental illumination of a fitness landscape. *Proc. Natl Acad. Sci. USA* **108,** 7896–7901 (2011).
29. Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813 (2004).
30. Pereira-Leal, J. B., Levy, E. D., Kamp, C. & Teichmann, S. A. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* **8,** R51 (2007).
31. Ryan, M., Graham, L. A. & Stevens, T. H. Voa1p functions in V-ATPase assembly in the yeast endoplasmic reticulum. *Mol. Biol. Cell* **19,** 5131–5142 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was supported by National Institutes of Health (NIH) grants R01-GM081592 (to J.W.T.) and R01-GM38006 (to T.H.S.), National Science Foundation (NSF) grants IOB-0546906 (to J.W.T.) and DEB-0516530 (to J.W.T.), NIH Genetics Training grant T32-GM007257 (to G.C.F.), NSF IGERT grant DGE-9972830 (to V.H.-S.) and the Howard Hughes Medical Institute (J.W.T.). We thank L. Graham, G. Butler and B. Houser for generating yeast strains and other assistance. We thank members of the Stevens and Thornton laboratories for helpful comments.

Author Contributions V.H.-S. performed the phylogenetic analysis and statistical reconstructions. G.C.F. performed functional experiments. All authors conceived the experiments, interpreted the results and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.W.T. (joet@uoregon.edu).

METHODS

***In silico* reconstruction of ancestral protein sequences.** V_0 complex subunits Vma3, Vma11 and Vma16 are sometimes referred to as subunits c, c′ and c″ in the literature. We searched GenBank for all eukaryote V-ATPase V_0 ring sequences (Supplementary Information, section 1). Our query returned subunit 3, 11 and 16 protein sequences for 26 species in Fungi, and subunit 3 and 11 sequences for 35 species in Metazoa, Amoebozoa and Apicomplexa. We aligned the sequences using PRANK v0.081202 (refs 32, 33). We selected the best-fit model (WAG with gamma-distributed rate variation and a proportion of invariant sites) using the Akaike Information Criterion as implemented in PROTTEST[34,35]. With this model, we used PhyML v3.0 to infer the maximum likelihood topology, branch lengths and model parameters[36]. We optimized the topology using the best result from nearest-neighbour interchange and subtree pruning and regrafting; we optimized all other free parameters using the default hill-climbing algorithm in PhyML. Phylogenetic support was calculated as the approximate likelihood ratio (converted from the approximate likelihood ratio statistic (aLRS) for branches reported by PhyML, using the equation aLR = exp[aLRS/2]) and as the likelihood ratio-based SH-like branch supports[37]. Nematoda subunit 3 and 11 sequences were connected by a very long branch basal to the Chromalveolata lineages. This result is inconsistent with the expectation that Nematoda are animals[38], so we excluded Nematoda data from further downstream analysis.

We inferred ML ancestral states and posterior probability distributions at each site for all ancestral nodes in the ML phylogeny using our own set of Python scripts, called Lazarus, which wraps PAML version 4.1 (ref. 39). Lazarus parsimoniously places ancestral gap characters according to Fitch's algorithm[40]. We characterized the overall support for Anc.3-11, Anc.16, Anc.3 and Anc.11 by binning the posterior probability of the ML state at each site into 5%-sized bins and then counting the proportion of total sites within each bin (Supplementary Information, section 2).

Robustness to alignment uncertainty. To assess the robustness of ancestral reconstructions to alignment uncertainty, we performed alignment using four algorithms: CLUSTAL version 2.0.10 (ref. 41), MUSCLE v3.7 (ref. 42), AMAP v2.2 (ref. 43), and PRANK v0.081202 (refs 32, 33). We then inferred the ML phylogeny and branch lengths for each alignment, using the methods described above. The resultant alignments varied in length from 347 sites (using CLUSTAL) to 683 sites (using PRANK), but all four alignments yielded the same ML topology with nearly identical ML branch lengths.

To determine which alignment algorithm yields the most accurate ancestral inferences under V-ATPase phylogenetic conditions, we simulated sequences across the V-ATPase ML phylogeny using insertion and deletion rates ranging from 0.0 to 0.1 indels per site. For each indel rate, we generated ten random unique indel-free ancestral sequences of 400 amino acids in length and then used INdelible[44] to simulate the ancestral sequence evolving along the branches of our ML phylogeny under the conditions of the WAG model with a proportion of invariant sites (+I) and a discrete gamma distribution of evolutionary rates (+G) with indel events randomly injected according to the specified indel rate. The size of each indel event was drawn from a Zipfian distribution with coefficient equal to 1.1 and the maximum length limited to 10 amino acids. We aligned the descendant sequences of each replicate using AMAP, CLUSTAL, MUSCLE and PRANK. For each alignment, we inferred the ML topology, branch lengths and model parameters using the methods described above. We used Lazarus to reconstruct all of the ancestral states, and queried Lazarus for the most-recent shared ancestor for opisthokont subunit 3/11 and opisthokont subunit 16 sequences. We measured the error of ancestral reconstructions as the proportion of ancestral sites that incorrectly contained an indel character (see Supplementary Information, section 6).

Plasmids and yeast strains. Bacterial and yeast manipulations were performed using standard laboratory protocols for molecular biology[45]. Plasmids that were used are listed in Supplementary Information, section 5. Ancestral sequences (pGF140, pGF139, pGF506 and pGF508) were synthesized by GenScript with a yeast codon bias. Triple haemagglutinin epitope tags were included before each stop codon. The Anc.3-11, Anc.16, Anc.3 and Anc.11 genes were subcloned to single-copy, *CEN*-based yeast vectors. The *ADH* terminator sequence (247 base pairs (bp)) and *Nat*^{r} drug resistance marker[46] were amplified using polymerase-chain-reaction (PCR) containing 40-bp tails homologous to the 3′ end of each coding region and vector sequence. Vectors were gapped, co-transformed into SF838-1Dα yeast with PCR fragments and cells were selected for *Nat*^{r}. A second round of *in vivo* ligation was used to place the ancestral genes under 500 bp of the *VMA3* or *VMA16* promoters to create pGF140 and pGF139, respectively. For vectors pGF240, pGF241, 1pGF252, pGF253, pGF503–pGF508, pGF510, pGF512–pGF515, pGF517–pGF519, pGF521, pGF523, pGF528, pGF529, pGF531, pGF534–pGF537 and pGF542, the relevant locus (Anc.3-11, Anc.16 or Anc.3) was PCR amplified with 5′ and 3′ untranslated flanking sequence and cloned into pCR4Blunt-TOPO (Invitrogen). When necessary, a modified Quikchange protocol[47] was used to introduce point mutations before the gene was subcloned into a yeast vector (pRS316 or pRS415). To generate pGF502, sequence from codon 31 to the stop codon of Anc.16 was amplified with the *ADH::Nat*^{r} cassette from pGF139, cloned into TOPO, and *in vivo* ligated downstream of the *VMA16* promoter (including a start codon) in pRS415.

A triple-fragment *in vivo* ligation was used to generate pGF646–pGF651. Gapped vector containing the *VMA16* promoter was transformed into yeast with two PCR fragments of the ring genes to be fused. For pGF646, the coding region of (1) *VMA16* (without codons 2–41) and (2) the coding region of Anc.11 (without codons 2–5 were amplified by PCR. The proteolipid on the C-terminal portion of the gene fusion also contained the *ADH* terminator and *Nat*^{r} cassette; the amplified products contained PCR tails with homology to link the genes to both the gapped vector and to each other. Gene fusions were modelled after the experimental design of Wang *et al.* (2007)[24] in which the lumenal protein sequence linking the two proteolipids was designed to be exactly 14 amino acids. To meet these criteria, additional amino acids were inserted into the following vectors linking the two subunits: pGF646 (Thr-Arg-Val-Asp), pGF648, pGF650 (Thr-Arg), pGF649, pGF651 (Gly-Ser).

Yeast strains that were used are listed in Supplementary Information, section 2. Strains containing deletion cassettes other than Kan^R [45] were constructed by PCR amplifying the *Hyg*^{R} or *Nat*^{r} cassette from pAG32 or pAG25, respectively, with primer tails with homology to flanking sequences to the *VMA11* or *VMA16* loci. *11Δ::Kan*^{R} and *16Δ::Kan*^{R} strains (SF838-1Dα) were transformed with the *Hyg*^{R} and *Nat*^{R} PCR fragments, respectively, and selected for drug resistance. The *11Δ::Hyg*^{R} locus was amplified and transformed into LGY113 (to create LGY125) and LGY115 (to create LGY124). This was repeated with the *16Δ::Nat*^{R} locus to create LGY139 and LGY143.

Yeast Growth Assays. Yeast were grown in liquid culture, diluted fivefold and spotted onto YEPD media buffered to pH 5.0 or yeast extract peptone dextrose media containing 25 mM (Figs 2, 3, 4) or 30 mM $CaCl_2$ (Fig. 2f).

Whole-cell extract preparation and immunoblotting. Yeast extracts and western blots were performed as previously described[31]. Antibodies that were used in this study included monoclonal primary anti-HA (Sigma-Aldrich), anti-Dpm1 (5C5; Invitrogen) and secondary horseradish-conjugated anti-mouse antibody (Jackson ImmunoResearch Laboratory, West Grove, Pennsylvania, USA).

Fluorescence microscopy. Staining with quinacrine was performed as previously described[31]. The cell wall (shown in red) was visualized using concanavalin A tetramethylrhodamine (Invitrogen). Microscopy images were obtained using an Axioplan 2 fluorescence microscope (Carl Zeiss). A ×100 objective, AxioVision software (Carl Zeiss) and Adobe Photoshop Creative Suite (v. 8.0) were used.

32. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA* **102,** 10557–10562 (2005).
33. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320,** 1632–1635 (2008).
34. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
35. Abascal, F., Zardoya, R. & Posada, D. Prottest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).
36. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).
37. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55,** 539–552 (2006).
38. Aguinaldo, A. M. A. *et al.* Evidence for a clade of nematodes, arthropods, and other moulting animals. *Nature* **387,** 489–493 (1997).
39. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
40. Fitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20,** 406–416 (1971).
41. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
42. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
43. Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15,** 330–340 (2005).
44. Fletcher, W. & Yang, Z. Indelible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26,** 1879–1888 (2009).
45. Sambrook, J. & Russel, D. W. *Molecular Cloning: A Laboratory Manual* 3rd edn (Cold Spring Harbor Laboratory Press, 2001).
46. Goldstein, A. L. & McCuster, J. H. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae. Yeast* **15,** 1541–1553 (1999).
47. Zheng, L., Baumann, U. & Reymond, J. L. An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res.* **32,** e115 (2004).

LETTER

doi:10.1038/nature10719

Global landscape of HIV–human protein complexes

Stefanie Jäger[1,2], Peter Cimermancic[2,3], Natali Gulbahce[1,2], Jeffrey R. Johnson[1,2,4], Kathryn E. McGovern[1,2], Starlynn C. Clarke[5], Michael Shales[1,2], Gaelle Mercenne[6], Lars Pache[7], Kathy Li[1,2,5], Hilda Hernandez[1,2,5], Gwendolyn M. Jang[1,2,8], Shoshannah L. Roth[9], Eyal Akiva[2,3], John Marlett[10], Melanie Stephens[9], Iván D'Orso[8]†, Jason Fernandes[8], Marie Fahey[1,2], Cathal Mahon[1,2,5], Anthony J. O'Donoghue[5], Aleksandar Todorovic[11], John H. Morris[5], David A. Maltby[5], Tom Alber[12], Gerard Cagney[13], Frederic D. Bushman[9], John A. Young[10], Sumit K. Chanda[7], Wesley I. Sundquist[6], Tanja Kortemme[2,3,14], Ryan D. Hernandez[2,3,14], Charles S. Craik[2,5], Alma Burlingame[2,5], Andrej Sali[2,3,5,14], Alan D. Frankel[2,8,14] & Nevan J. Krogan[1,2,4,14]

Human immunodeficiency virus (HIV) has a small genome and therefore relies heavily on the host cellular machinery to replicate. Identifying which host proteins and complexes come into physical contact with the viral proteins is crucial for a comprehensive understanding of how HIV rewires the host's cellular machinery during the course of infection. Here we report the use of affinity tagging and purification mass spectrometry[1–3] to determine systematically the physical interactions of all 18 HIV-1 proteins and polyproteins with host proteins in two different human cell lines (HEK293 and Jurkat). Using a quantitative scoring system that we call MiST, we identified with high confidence 497 HIV–human protein–protein interactions involving 435 individual human proteins, with ~40% of the interactions being identified in both cell types. We found that the host proteins hijacked by HIV, especially those found interacting in both cell types, are highly conserved across primates. We uncovered a number of host complexes targeted by viral proteins, including the finding that HIV protease cleaves eIF3d, a subunit of eukaryotic translation initiation factor 3. This host protein is one of eleven identified in this analysis that act to inhibit HIV replication. This data set facilitates a more comprehensive and detailed understanding of how the host machinery is manipulated during the course of HIV infection.

A map of the physical interactions between proteins within a particular system is necessary for studying the molecular mechanisms that underlie the system. The analysis of protein–protein interactions (PPIs) has been successfully accomplished in different organisms using a variety of technologies, including mass spectrometry approaches[1,3,4] and those designed to detect pairwise physical interactions, including the two-hybrid yeast system[5,6] and protein-fragment complementation assays[7]. Although two-hybrid methodologies have been used to systematically study host–pathogen interactions[8,9], so far no systematic affinity tagging/purification mass spectrometry (AP–MS) study has been carried out on any host–pathogen system. Here we have targeted HIV-1 for such an analysis, uncovering a wide variety of host proteins, complexes and pathways that are hijacked by the virus during the course of infection.

We aimed to identify host proteins associated with HIV-1 proteins systematically and quantitatively using an AP–MS approach[2,3]. To this end, we cloned the genes corresponding to all 18 HIV-1 proteins and polyproteins, including the accessory factors (Vif, Vpu, Vpr and Nef), Tat, Rev, the polyproteins (Gag, Pol and Gp160) and the corresponding processed products (MA, CA, NC and p6; PR, RT and IN; and Gp120 and Gp41, respectively) (Supplementary Fig. 1 and Supplementary Table 1). Each clone was fused to a purification tag (consisting of 2×Strep and 3×Flag) and transiently transfected into HEK293 cells; each also was used to generate stably expressed, tetracycline-inducible, affinity-tagged versions of the proteins in Jurkat cells (Fig. 1a and Supplementary Fig. 2). Following multiple purifications of each factor from both cell lines, the material on the anti-FLAG or Strep-Tactin beads, as well as the eluted material, was analysed by mass spectrometry (Fig. 1a and Supplementary Table 2). Finally, an aliquot of each purified factor was subjected to SDS–polyacrylamide gel electrophoresis, stained (Supplementary Fig. 3) and subjected to analysis by mass spectrometry.

For each HIV factor, we identified co-purifying host proteins that were reproducible regardless of the protocol used (Supplementary Figs 4, 5 and 7 and Supplementary Data 1). Several scoring systems can quantify PPIs from AP–MS proteomic data sets, including PE[10], CompPASS[4] and SAINT[11]. For this data set, we devised a scoring system particularly suited for identifying AP–MS-derived host–pathogen PPIs, which we call MiST (mass spectrometry interaction statistics). The MiST score is a weighted sum of three measures: protein abundance measured by peak intensities from the mass spectrum (abundance); invariability of abundance over replicated experiments (reproducibility); and uniqueness of an observed host–pathogen interaction across all viral purifications (specificity) (Fig. 1b and Supplementary Methods). These three metrics are summed by principal component analysis into a composite score (Fig. 1c and Supplementary Data 2). By comparing our dataset with a benchmark of well-characterized HIV–human PPIs (Supplementary Table 3), analysis of the MiST scoring system revealed superior performance on our data set when compared to CompPASS or SAINT (Supplementary Fig. 6) (and comparable performance using other data sets (Supplementary Fig. 8)) and allowed us to define a MiST cut-off of 0.75, corresponding to ~4% of all detected interactions. To estimate how many interactions would exceed this threshold by chance, we randomly shuffled our data set 1,000 times. A random MiST score of 0.75 or greater was assigned to an interaction ten times less frequently than we saw among the MiST scores for the real data, and the probability of an interaction assignment with a random MiST score greater than 0.75 was 2.5×10^{-4} (Fig. 1d).

At the MiST threshold of 0.75, the number of host proteins we found associated with each HIV protein ranged from 0 (CA and p6) to 63 (Gp160) (Fig. 1e). In total, we observed 497 different HIV–human PPIs (347 and 348 identified from HEK293 cells and Jurkat cells, respectively) (Supplementary Data 3). We detected 196 interactions (~40%) in both cell types; 150 and 151 were specific to the HEK293 cells and the Jurkat cells, respectively (Fig. 1e). Only some of these

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158, USA. [2]California Institute for Quantitative Biosciences, QB3, San Francisco, California 94158, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California 94158, USA. [4]J. David Gladstone Institutes, San Francisco, California 94158, USA. [5]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158, USA. [6]Department of Biochemistry, University of Utah, Salt Lake City, Utah 84112, USA. [7]Sanford-Burnham Medical Research Institute, La Jolla, California 92037, USA. [8]Department of Biochemistry and Biophysics, University of California, San Francisco, California 94158, USA. [9]Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [10]The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [11]Department of Chemistry, University of California, Berkeley, California 94720, USA. [12]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [13]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland. [14]Host Pathogen Circuitry Group, University of California, San Francisco, California 94158, USA. †Present address: Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, Texas, USA, 75390

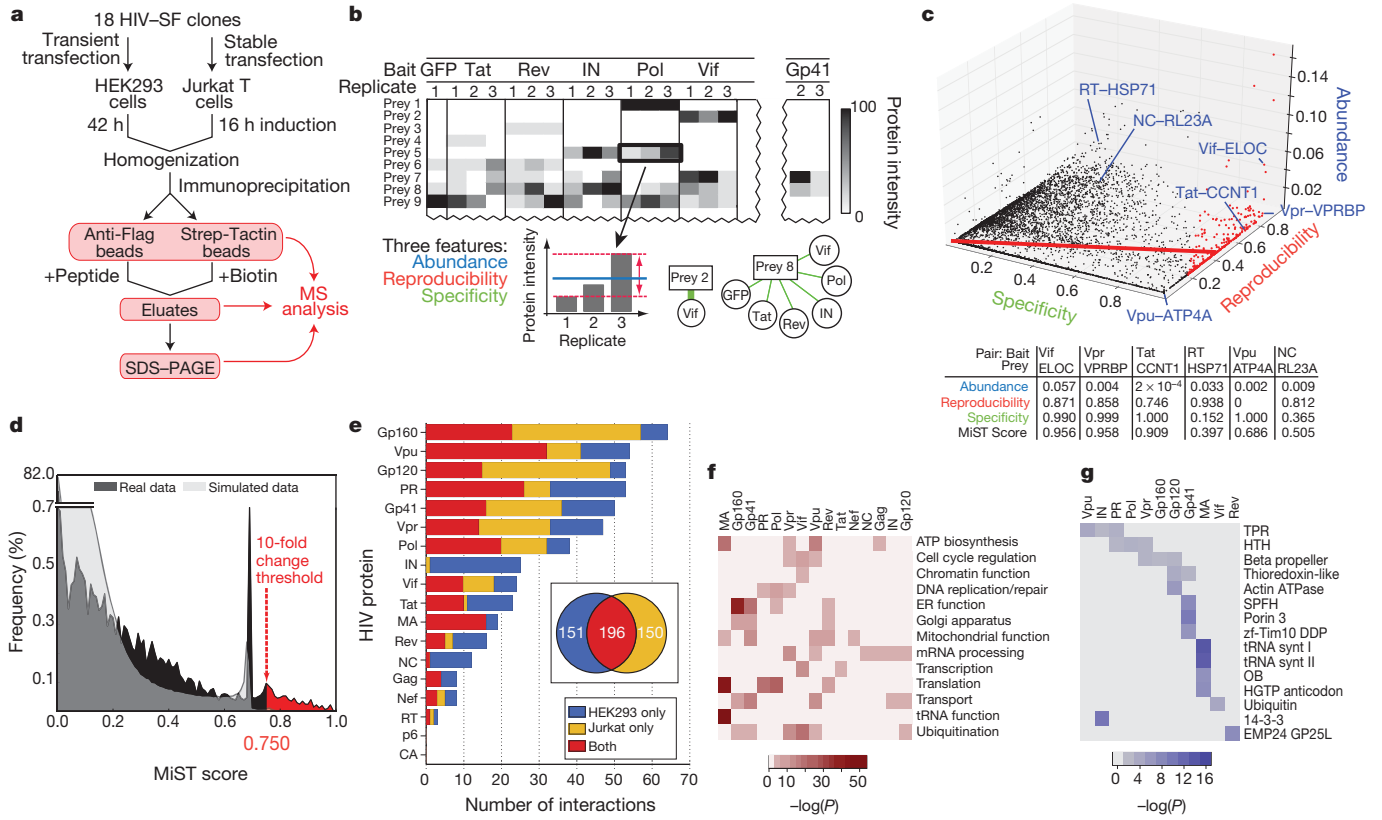


Figure 1 | Affinity purification of HIV-1 proteins, analysis and scoring of mass spectrometry data. **a**, Flowchart of the proteomic AP–MS used to define the HIV–host interactome. PAGE, polyacrylamide gel electrophoresis. SF, 2×Strep–3×Flag affinity tag. **b**, Data from AP–MS experiments are organized in an interaction table with cells representing amount of prey protein purified (for example spectral counts or peptide intensities). Three features are used to describe bait–prey relationships: abundance (blue), reproducibility (the invariability of bait–prey pair quantities; red) and specificity (green). **c**, All bait–prey pairs are mapped into the three-feature space (abundance, reproducibility and specificity). The MiST score is defined as a projection on the first principal component (red line). All interactions, represented as nodes, above the defined threshold (0.75) are shown in red. This procedure separates the interactions more likely to be biologically relevant (for example Vif–ELOC (ELOC also known as TCEB1), Vpr–VPRBP and Tat–CCNT1) from the interactions that are likely to be less relevant owing to low reproducibility (Vpu–ATP4A) or specificity (RT–HSP71 (HSP71 also known as HSPA1A) and NC–RL23A (RL23A also known as RPL23A)). **d**, The histogram of MiST scores (real data) is compared with a randomized set of scores obtained from randomly shuffling the bait–prey table (simulated data). The MiST score threshold (0.75) was defined using a benchmark (Supplementary Table 3) whereby the predictions are enriched for these interactions by a factor of at least ten relative to random predictions (as well as through ROC (receiver operating characteristic) and recall plots (Supplementary Fig. 6)). **e**, Bar graph of the number of host proteins we found interacting with each HIV factor (MiST score, >0.75). The cell type in which the interaction was found is represented in blue (HEK293 only), yellow (Jurkat only) or red (both). **f**, **g**, Heat maps representing enriched biological functions (**f**) and domains (**g**) from the human proteins identified as interacting with HIV proteins (Supplementary Methods). ER, endoplasmic reticulum; mRNA, messenger RNA; tRNA, transfer RNA. TPR, tetratricopeptide repeat; HTH, helix-turn-helix; SPFH, stomatin–prohibitin–flotillin–HflK/C.

specificities could be explained by differential gene expression in the two cell lines (Supplementary Fig. 9). Using antibodies against 26 of the human proteins, and affinity-tagged versions of an additional 101, we could confirm 97 of the 127 AP–MS derived HIV–human PPIs using co-immunoprecipitation/western blot analysis (76% success rate) (Supplementary Figs 10 and 11), suggesting that we derived a high-quality physical interaction data set.

We next analysed the functional categories of host proteins associated with each HIV protein, and in doing so uncovered many expected connections. These included an enrichment of host factors involved in transcription physically linked to the HIV transcription factor Tat and an enrichment of host machinery implicated in the regulation of ubiquitination associating with Vpu, Vpr and Vif, HIV accessory factors that hijack ubiquitin ligases[12] (Fig. 1f and Supplementary Data 4). When we considered domain types instead of whole proteins (Fig. 1g and Supplementary Table 4), we found that host proteins interacting with IN are enriched in 14-3-3 domains, which generally bind phosphorylated regions of proteins[13], and that proteins containing β-propellers have a higher propensity for binding to Vpr (for additional domain enrichment analysis, see Supplementary Fig. 12). These domain analyses could facilitate future structural modelling of HIV–human PPIs.

Next we compared our data to other HIV-related data sets, including previously published HIV–human PPIs and host factors implicated in HIV function from genome-wide RNA interference (RNAi) screens. For example, the VirusMint database[14] contains 587 HIV–human literature-curated PPIs (Supplementary Data 5), which are mostly derived from small-scale, targeted studies. Although the overlap between the 497 interactions identified in this work and those in VirusMint is statistically significant ($P = 8 \times 10^{-8}$), it corresponds to only 19 PPIs (Fig. 2a and Supplementary Table 5). However, a greater overlap exists, one that remains statistically significant, when interactions below the MiST threshold of 0.75 are considered using a sliding cut-off (for example, at a MiST score of 0.2 there exists an overlap of 67 PPIs ($P = 1 \times 10^{-3}$); Fig. 2c, red lines, and Supplementary Data 6). This overlap indicates that we have indeed identified many interactions that have been previously reported. However, it is likely that the higher scoring interactions identified here have a greater chance of being biologically relevant with respect to HIV function than do many of those in VirusMint.

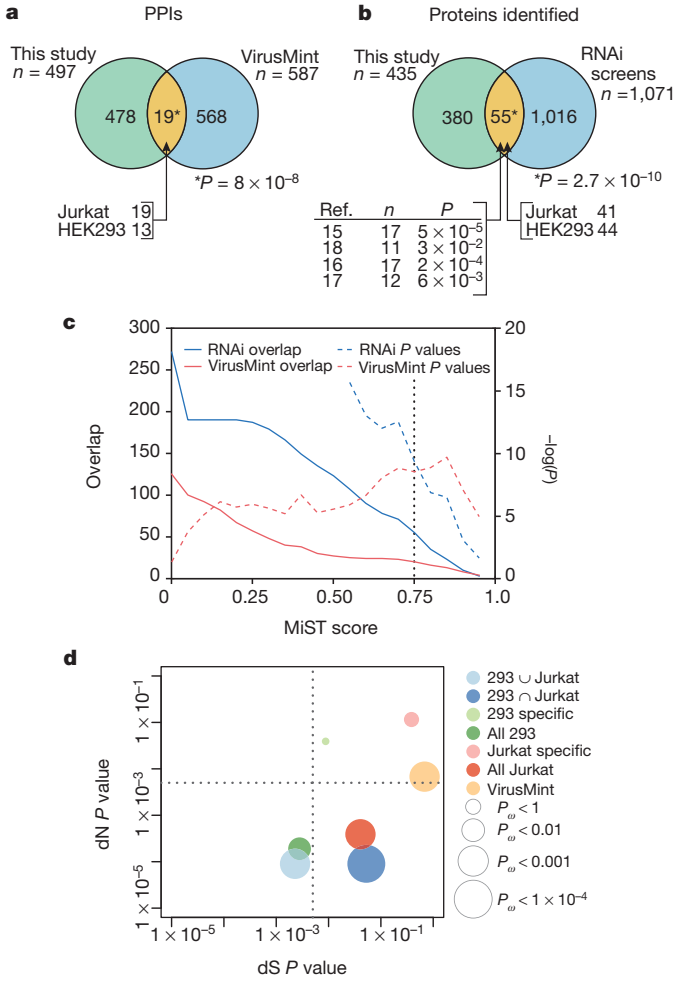


Figure 2 | Comparison of PPI data with other HIV data sets. **a**, Overlap of the 497 HIV–human PPIs with the 587 PPIs reported in VirusMint (Supplementary Table 5). **b**, Overlap of the 435 human proteins with the genes identified in four HIV-dependency RNAi screens[15–18] (Supplementary Table 6). **c**, Number of interactions overlapping with VirusMint (solid red line) and proteins with RNAi screens (solid blue line) as functions of the MiST cut-off. The *P* values of the overlap are represented as dashed lines using the same colours (Supplementary Data 6 and 8). **d**, Comparative genomics analysis of divergence patterns between human and rhesus macaque reveals strong evolutionary constraints on human proteins binding to HIV proteins. The *x* and *y* axes represent *P* values for the synonymous (dS) and non-synonymous (dN) rates of evolution (Supplementary Methods). Horizontal and vertical dotted lines are drawn at 0.5% to indicate the Bonferroni significance threshold for each axis. For the VirusMint data, the significance of ω (dN/dS) is primarily driven by higher rates of synonymous evolution. ∪, union; ∩, intersection; P_ω, bootstrap-based *P* value for ω.

Recently, four RNAi screens identified host factors that have an adverse effect on HIV-1 replication when knocked down[15–18]. In total, 1,071 human genes were identified in these four studies (Supplementary Data 7), 55 of which overlap with the 435 proteins ($P = 2.7 \times 10^{-10}$; Fig. 2b, Supplementary Fig. 12 and Supplementary Table 6). Again, this overlap increases (as does its statistical significance) if we consider proteins participating in HIV–human PPIs with MiST scores below 0.75 (Fig. 2c, blue lines, and Supplementary Data 8).

To identify the evolutionary forces operating on host proteins interacting with HIV-1, we performed a comparative genomics analysis of divergence patterns between human and rhesus macaque. The proteins identified in both HEK293 and Jurkat cell lines had stronger signatures of evolutionary constraint than those identified exclusively in one cell line or in VirusMint (Fig. 2d). Points in the lower-right quadrant of Fig. 2d show signatures of strong purifying selection, whereas the upper-right quadrant shows signatures more consistent with neutral evolution. This observation suggests that the PPIs identified in our study, especially the ones identified in both cell types, are more physiologically relevant to mammalian evolution than those reported in VirusMint.

We next plotted the 497 HIV–human interactions identified in this study in a network representation (Fig. 3) containing nodes corresponding to 16 HIV (yellow) and 435 human factors that were derived from the HEK293 cells (blue), Jurkat cells (red) or both. We also introduced 289 interactions between human proteins (black edges) derived from several databases[19,20] (Supplementary Data 9). These human–human interactions helped to identify many host complexes, including several that have been previously characterized (see Supplementary Information for a detailed discussion of the HIV–human interaction data sets). Ultimately, all data will be accessible for searching and comparison to other HIV-related data sets using the web-based software GPS-PROT[21] (http://www.gpsprot.org/).

Notably, we found that Pol and PR, which we needed to make catalytically inactive (Supplementary Fig. 1), bound the translational initiation complex eIF3, a 13-subunit complex (eIF3a to eIF3m). We detected 12 of the subunits bound to Pol and/or PR, except eIF3j, which is only loosely associated with the complex[22] (Fig. 4a). Even though PR is the smallest of the *pol*-encoded proteins, we find it associated with the greatest number of host factors (Fig. 4a). To determine whether components of the translation complex are substrates for PR, FLAG-tagged versions of ten eIF3 subunits were individually co-transfected, each with a small amount of active HIV-1 PR, into HEK293 cells. The cell lysates were analysed by western blotting and only eIF3d was found to be cleaved (Fig. 4b). Purification of tagged versions of the amino and carboxy termini of cleaved eIF3d revealed that only the N terminus of 114 amino-acid residues associates with the eIF3 complex (Supplementary Table 7). The cleavage occurred with an efficiency similar to that of the processing of the natural PR substrate Gag (Fig. 4c), whereas two cellular proteins previously described to be cleaved by HIV PR, PAPBC1[23] and BCL2[24], were cleaved only at higher PR concentrations or not at all, respectively. To confirm this result *in vitro*, we incubated purified human eIF3 with active PR, resulting in the removal of a 70-kDa band and the appearance of a ~60-kDa protein product (Fig. 4d). Analysis of the cleaved product by N-terminal sequencing revealed a cleavage of eIF3d between Met 114 and Leu 115, which corresponds to the consensus sequence for HIV-1 protease[25] and falls within the RNA-binding domain (RRM) of eIF3d (ref. 26; Fig. 4d).

Next we used four to six short interfering RNAs against different eIF3 subunits in HIV infectivity assays (Fig. 4e, f, Supplementary Fig. 14 and Supplementary Table 8). Using a fusion of HIV with vesicular stomatitis virus glycoprotein (VSV-G), which only allows for a single round of replication, knockdown of eIF3d, but not other eIF3 subunits, resulted in an increase in infectivity (Fig. 4e), suggesting that this factor acts in early stages of infection. In assays requiring multiple rounds of HIV infection, knockdown of eIF3d, eIF3e and eIF3f enhanced HIV NL4.3 infectivity by a factor of three to five, whereas inhibition of eIF3c, eIF3g and eIF3i had no promoting effect (Fig. 4f). Consistent with these results, a previous overexpression screen for factors that restrict HIV-1 replication identified eIF3f as the most potent inhibitory clone[27]. Furthermore, using assays monitoring both early and late products we found that knockdown of eIF3d results in an increase in accumulation of reverse transcription product (Fig. 4g and Supplementary Fig. 15). This suggests that eIF3 does in fact have a role in the early stages of infection, perhaps by binding to the viral RNA through the RNA-binding domain in eIF3d, and thus inhibiting RT, an effect that is overcome by PR cleavage of eIF3d (Supplementary Fig. 16). These results suggest that our data set will be enriched not only for host proteins the virus requires for efficient replication (Fig. 2b, c), but also those that have an inhibitory role during infection. Indeed, we have found that an additional ten factors from our list of interactors, when knocked down by RNAi, produce an increase in HIV infection (Supplementary Figs 17–19, Supplementary Tables 12

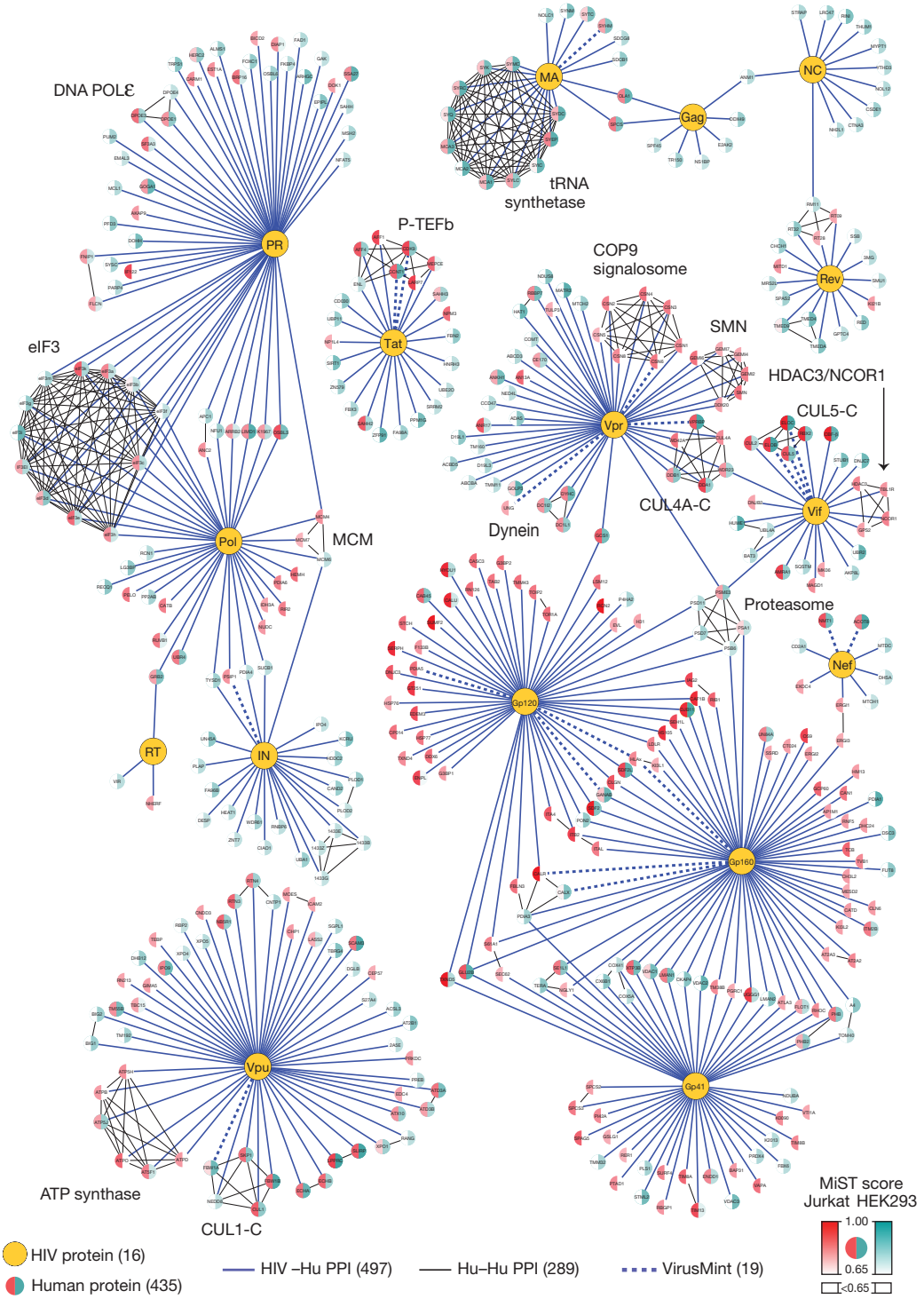


Figure 3 | Network representation of the HIV–human PPIs. In total, 497 HIV–human interactions (blue) are represented between 16 HIV proteins and 435 human factors. Each node representing a human protein is split into two colours and the intensity of each colour corresponds to the MiST score from interactions derived from HEK293 (blue) or Jurkat (red) cells. Black edges correspond to interactions between host factors (289) that were obtained from publicly available databases; dashed edges correspond to interactions also found in VirusMint[14].

and 13 and Supplementary Methods). Knockdown of two of these, DESP and HEAT1, also resulted in an increase in HIV integration (Supplementary Fig. 20 and Supplementary Table 14), consistent with their physical association with IN.

As well as performing the systematic AP–MS study reported here, we explored in further detail the biological significance of two newly identified HIV–human interactions: HIV protease targeting a component of eIF3 that is inhibitory to HIV replication; and CBF-β, a new component of the Vif–CUL5 ubiquitin ligase complex required for APOBEC3G stability and HIV infectivity[28]. Further work will be required to determine whether, how and at what stage of infection the remaining host factors impinge on HIV function. Ultimately, our analysis of the host factors co-opted by different viruses using the same proteomic pipeline will allow for the identification of protein complexes routinely targeted by different pathogens, which may represent better therapeutic targets for future studies.

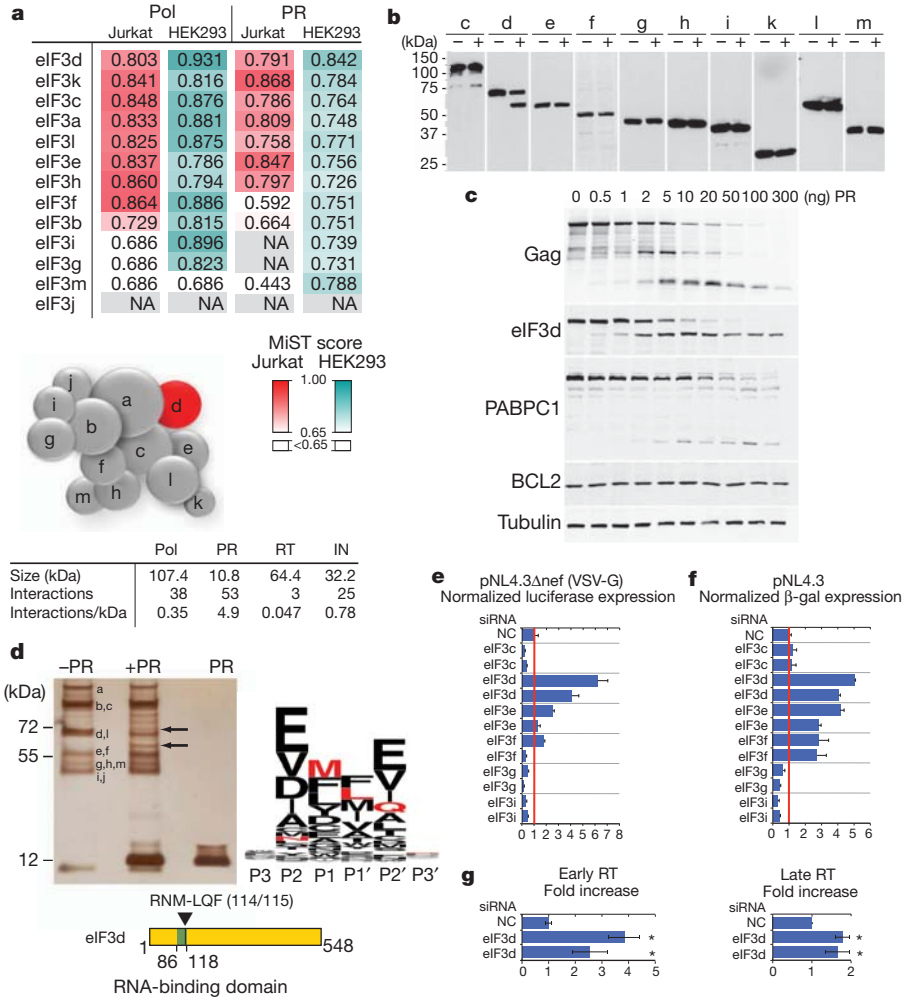


Figure 4 | eIF3d is cleaved by HIV-1 PR and inhibits infection. **a**, MiST scores for eIF3 subunits associated with PR (right) and Pol (left) in HEK293 and Jurkat cells. Sizes of the proteins and numbers of significant interactions (MiST score, >0.75) detected for Pol and its subunits are shown below, as is a modular representation of the eIF3 complex[29]. The cleaved subunit, eIF3d, is in red. **b**, Western blot of HEK293 cell lysate expressing FLAG-tagged eIF3 subunits in the absence (−) or presence (+) of active PR probed with an anti-FLAG antibody. **c**, HEK293 cells were co-transfected with Gag, or FLAG-tagged eIF3d, PABPC1, BCL2 and increasing amounts of PR. Cell lysates were probed against Gag (upper panel), FLAG-tagged eIF3d (middle panel) or tubulin as control (lower panel). **d**, Silver stain of purified eIF3 complex incubated with recombinant HIV-1 PR. The residues corresponding to the eIF3d cleavage site (red) is located within the RNA-binding domain[26]. **e**, **f**, HeLa-derived P4/R5 MAGI cells were transfected with two different short interfering RNAs (siRNAs) targeting individual subunits of the eIF3 complex (Supplementary Tables 7 and 9) and subsequently infected with either a pNL4-3-derived, VSV-G-pseudotyped, single-cycle virus (HIV–VSV-G) (**e**) or wild-type pNL4-3 (**f**). NC, negative control. **g**, Early (left) and late (right) HIV-1 DNA levels measured by quantitative PCR amplification in cells transfected with two independent eIF3d siRNAs or with control siRNAs. Samples were normalized by input DNA amount or by cellular gene (*HMBS*) copy number. The RT and replication assays were done three to five times and the standard deviations are shown (Supplementary Tables 7, 10 and 11). *$P < 0.05$ (Kruskal–Wallis test with Dunn's correction for multiple comparisons).

METHODS SUMMARY

More details on experimental assays, plasmid constructs, sequences, cell lines, antibodies and computational analysis are provided in Supplementary Methods. Briefly, affinity tagging and purification was carried out as previously described[2] and the protein samples were analysed on a Thermo Scientific LTQ Orbitrap XL mass spectrometer. For the evolutionary analysis, genome-wide alignments to rhesus macaque were downloaded from the University of California, Santa Cruz genome browser (http://genome.ucsc.edu/) and evolutionary rates for each group of genes considered were measured using the synonymous and non-synonymous rates of evolution. For the *in vitro* protease assay, maltose binding protein (MBP)-tagged PR was expressed in BL21 (Gold) DE3 cells in the presence of 100 μM Saquinavir and purified on an MBP trap column. Purified eIF3 was obtained from J. Cate (University of California, Berkeley). For the infection assays, HeLa P4.R5 cells were transfected with short interfering RNAs and after 48 h infected with pNL4-3 or a pNL4-3-derived VSV-G-pseudotyped reporter virus. Infection levels were determined by luminescence read-out.

Received 26 March; accepted 18 November 2011.
Published online 21 December 2011.

1. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636 (2006).
2. Jäger, S. *et al.* Purification and characterization of HIV-human protein complexes. *Methods* **53,** 13–19 (2011).
3. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643 (2006).
4. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138,** 389–403 (2009).
5. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322,** 104–110 (2008).
6. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122,** 957–968 (2005).
7. Tarassov, K. *et al.* An *in vivo* map of the yeast protein interactome. *Science* **320,** 1465–1470 (2008).
8. Calderwood, M. A. *et al.* Epstein-Barr virus and virus human protein interaction maps. *Proc. Natl Acad. Sci. USA* **104,** 7606–7611 (2007).

9. Shapira, S. D. *et al.* A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* **139,** 1255–1267 (2009).
10. Collins, S. R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6,** 439–450 (2007).
11. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods* **8,** 70–73 (2011).
12. Malim, M. H. & Emerman, M. HIV-1 accessory proteins–ensuring viral survival in a hostile environment. *Cell Host Microbe* **3,** 388–398 (2008).
13. Yaffe, M. B. *et al.* The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell* **91,** 961–971 (1997).
14. Chatr-aryamontri, A. *et al.* VirusMINT: a viral protein interaction database. *Nucleic Acids Res.* **37,** D669–D673 (2009).
15. Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319,** 921–926 (2008).
16. König, R. *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* **135,** 49–60 (2008).
17. Yeung, M. L., Houzet, L., Yedavalli, V. S. & Jeang, K. T. A genome-wide short hairpin RNA screening of Jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J. Biol. Chem.* **284,** 19463–19473 (2009).
18. Zhou, H. *et al.* Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe* **4,** 495–504 (2008).
19. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* **38,** D497–D501 (2010).
20. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–D539 (2006).
21. Fahey, M. E. *et al.* GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. *BMC Bioinformatics* **12,** 298 (2011).
22. Hinnebusch, A. G. eIF3: a versatile scaffold for translation initiation complexes. *Trends Biochem. Sci.* **31,** 553–562 (2006).
23. Alvarez, E., Castello, A., Menendez-Arias, L. & Carrasco, L. HIV protease cleaves poly(A)-binding protein. *Biochem. J.* **396,** 219–226 (2006).
24. Strack, P. R. *et al.* Apoptosis mediated by HIV protease is preceded by cleavage of Bcl-2. *Proc. Natl Acad. Sci. USA* **93,** 9571–9576 (1996).
25. Schilling, O. & Overall, C. M. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature Biotechnol.* **26,** 685–694 (2008).
26. Asano, K. *et al.* Structure of cDNAs encoding human eukaryotic initiation factor 3 subunits. Possible roles in RNA binding and macromolecular assembly. *J. Biol. Chem.* **272,** 27042–27052 (1997).
27. Valente, S. T., Gilmartin, G. M., Mott, C., Falkard, B. & Goff, S. P. Inhibition of HIV-1 replication by eIF3f. *Proc. Natl Acad. Sci. USA* **106,** 4071–4078 (2009).
28. Jäger, S. *et al.* Vif hijacks CBF-β to degrade APOBEC3G and promote HIV-1 infection. *Nature* http://dx.doi.org/10.1038/nature10693 (this issue).
29. Zhou, M. *et al.* Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. *Proc. Natl Acad. Sci. USA* **105,** 18139–18144 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Choi, Z. Rizvi and E. Kwon for cloning of human genes and J. Cate for purified eIF3. We also thank J. Gross, R. Andino, R. Harris, M. Daugherty and members of the Krogan lab for discussion. This research was funded by grants from QB3@UCSF and the National Institutes of Health (P50 GM082250 to N.J.K., A.D.F., C.S.C. and T.A.; P01 AI090935 to N.J.K., S.K.C., J.A.Y. and F.D.B.; P50 GM081879 to N.J.K. and A.B.; P50 GM082545 to W.I.S.; P41RR001614 to A.B.; U54 RR022220 to A.S.; P01 GM073732-05 to A.T.; CHRP-ID08-TBI-063 to S.K.C.; P41 RR001081 to J.H.M.) and from the Nomis Foundation (to J.A.Y.). N.J.K. is a Searle Scholar and a Keck Young Investigator.

Author Contributions S.J. generated the protein–protein interaction map; P.C. developed the MiST scoring system; N.G., M. Shales, E.A., M.F., J.H.M., J.R.J. and R.D.H. provided computational support; K.E.M., K.L., J.R.J., H.H., G.M.J., I.D., J.F. and D.A.M. provided experimental support; S.J., S.C.C., A.J.O. and A.T. characterized the PR–eIF3d interaction; S.J., G.M.J., C.M. and G.M. confirmed the interactions by immunoprecipitation/western blot; L.P., S.L.R., J.M. and M. Stephens used RNAi for functional verification; T.A., G.C., F.D.B., J.A.Y., S.K.C., W.I.S., T.K., R.D.H., C.S.C., A.B., A.S., A.D.F. and N.J.K. supervised the research; and S.J., P.C., A.S. and N.J.K. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.J.K. (krogan@cmp.ucsf.edu).

LETTER

doi:10.1038/nature10693

Vif hijacks CBF-β to degrade APOBEC3G and promote HIV-1 infection

Stefanie Jäger[1,2]*, Dong Young Kim[3]*, Judd F. Hultquist[4]*, Keisuke Shindo[4], Rebecca S. LaRue[4], Eunju Kwon[3], Ming Li[4], Brett D. Anderson[4], Linda Yen[3], David Stanley[3], Cathal Mahon[1,2,3], Joshua Kane[1,2], Kathy Franks-Skiba[1,2], Peter Cimermancic[2,5], Alma Burlingame[2,3], Andrej Sali[2,3,5,6], Charles S. Craik[2,3], Reuben S. Harris[4], John D. Gross[2,3,6] & Nevan J. Krogan[1,2,6,7]

Restriction factors, such as the retroviral complementary DNA deaminase APOBEC3G, are cellular proteins that dominantly block virus replication[1–3]. The AIDS virus, human immunodeficiency virus type 1 (HIV-1), produces the accessory factor Vif, which counteracts the host's antiviral defence by hijacking a ubiquitin ligase complex, containing CUL5, ELOC, ELOB and a RING-box protein, and targeting APOBEC3G for degradation[4–10]. Here we reveal, using an affinity tag/purification mass spectrometry approach, that Vif additionally recruits the transcription cofactor CBF-β to this ubiquitin ligase complex. CBF-β, which normally functions in concert with RUNX DNA binding proteins, allows the reconstitution of a recombinant six-protein assembly that elicits specific polyubiquitination activity with APOBEC3G, but not the related deaminase APOBEC3A. Using RNA knockdown and genetic complementation studies, we also demonstrate that CBF-β is required for Vif-mediated degradation of APOBEC3G and therefore for preserving HIV-1 infectivity. Finally, simian immunodeficiency virus (SIV) Vif also binds to and requires CBF-β to degrade rhesus macaque APOBEC3G, indicating functional conservation. Methods of disrupting the CBF-β–Vif interaction might enable HIV-1 restriction and provide a supplement to current antiviral therapies that primarily target viral proteins.

Mammals have evolved cellular proteins termed restriction factors that function to prevent the spread of mobile genetic elements including retroviruses[1–3]. As a counter-defence, most retroviruses, including the human pathogen HIV-1, have developed mechanisms to prevent restriction, often through subversion of the host's ubiquitin–proteasome system. In eukaryotic cells, 8.6-kDa ubiquitin moieties are added to a target protein by sequential action of one of two ubiquitin-activating enzymes (E1), which transfer ubiquitin to a pool of dozens of ubiquitin-conjugating enzymes (E2) that, in turn, collaborate with hundreds of ubiquitin ligases (E3) to catalyse transfer to specific substrates[11]. If more than four ubiquitins are joined together through K48 linkages, the target protein is usually degraded by the 26S proteasome[12]. At least three HIV-1 proteins, Vif, Vpu and Vpr, hijack cullin-RING E3 ligases consisting of CUL5, CUL1 and CUL4A to promote ubiquitination and degradation of APOBEC3 family members (for example, APOBEC3G, A3G), BST2/tetherin and an unknown, putative restriction factor, respectively[2]. Understanding the composition of cullin-RING E3 ligase complexes and the underlying cellular signalling components may provide therapeutic routes for treating a variety of human diseases, including infection by HIV-1.

HIV-1 Vif is recruited to CUL5 by virtue of its SOCS box, which contains an elongin C binding helix (the BC-box), a conserved HCCH Zn binding motif and a short Cullin Box[4–6]. Although a structure of the BC-box peptide in complex with the heterodimer of Elongin B and C (ELOBC) has been reported[13], the architecture of the full-length Vif in complex with host factors has remained elusive, in part because Vif complexes have poor solubility and activity. We therefore reasoned that Vif may bind an additional host factor and that such a factor may render it more tractable *in vitro*.

We took an unbiased proteomic approach to identify host factors that bind all 18 HIV processed and polyproteins using an affinity tag/purification mass spectrometry (AP–MS) approach[14,15]. To this end, 2×Strep and 3×Flag was fused to the carboxy (C) terminus of these factors, including Vif. The tagged Vif construct was both transiently transfected into HEK293 cells and used to make a stable, tetracycline-inducible Vif–Strep–Flag Jurkat T cell line (Fig. 1a). Epitope-tagged Vif was purified from both cell types using antibodies specific to either Strep or Flag and aliquots of the co-purifying proteins were subjected to SDS–polyacrylamide gel electrophoresis (SDS–PAGE) (Fig. 1b). Materials from each step were analysed by mass spectrometry[14].

Using a new scoring system for data derived from AP–MS studies, termed Mass Spectrometry Interaction Statistics (MiST)[15], we identified 24 Vif–human protein–protein interactions with seven of them found in both cell types (Fig. 1c). Seventeen of these were verified independently by co-immunoprecipitation (Supplementary Fig. 1). Among these were the components of the E3 ubiquitin ligase complex, CUL5, ELOB and ELOC, known to interact with Vif and trigger A3G degradation[4–6,8–10]. Although the RING-box protein RBX1 was originally reported as part of this complex[4], only RBX2 was above the MiST score threshold used[15] consistent with recent work showing that it binds CUL5 (refs 16, 17). We did not find endogenous A3G, probably because of its poor expression in HEK293 and Jurkat cell lines exacerbated by further depletion through Vif-mediated degradation. We did find Vif associating with two proteins that function in autophagy, AMRA1 and SQSTM, as well as with the transcriptional co-repressor complex NCOR1/HDAC3/GPS2/TBL1R (the last only in T cells) (Fig. 1c). Also, in both cell types, Vif was found to interact with the transcription cofactor CBF-β, which is known to heterodimerize with the RUNX family of transcription factors[18].

To determine if any of the newly defined Vif interactors belong to the Vif–CUL5 complex, we performed double affinity purifications using cells co-transfected with Vif–2×Strep and either A3G– or CUL5–3×Flag (Fig. 1d, e). After purification first with Strep-Tactin and second with anti-Flag beads, mass spectrometry analysis of the final elution revealed the presence of CUL5, ELOB, ELOC, RBX2 and invariably CBF-β, strongly suggesting that this last protein may be a new component of the Vif E3 ubiquitin ligase complex (other factors from single Vif purifications depicted in Fig. 1c were not present). To confirm this interaction and the composition of the complex, we performed an additional double affinity purification experiment using

[1]Department of Cellular and Molecular Pharmacology, University of California-San Francisco, San Francisco, California 94158, USA. [2]California Institute for Quantitative Biosciences, QB3, San Francisco, California, California 94158, USA. [3]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158, USA. [4]Department of Biochemistry, Molecular Biology and Biophysics, Institute for Molecular Virology, Center for Genome Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA. [5]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California 94158, USA. [6]HPC (Host Pathogen Circuitry) Group, University of California-San Francisco, San Francisco, California 94158, USA. [7]J. David Gladstone Institutes, San Francisco, California 94158, USA.
*These authors contributed equally to this work.

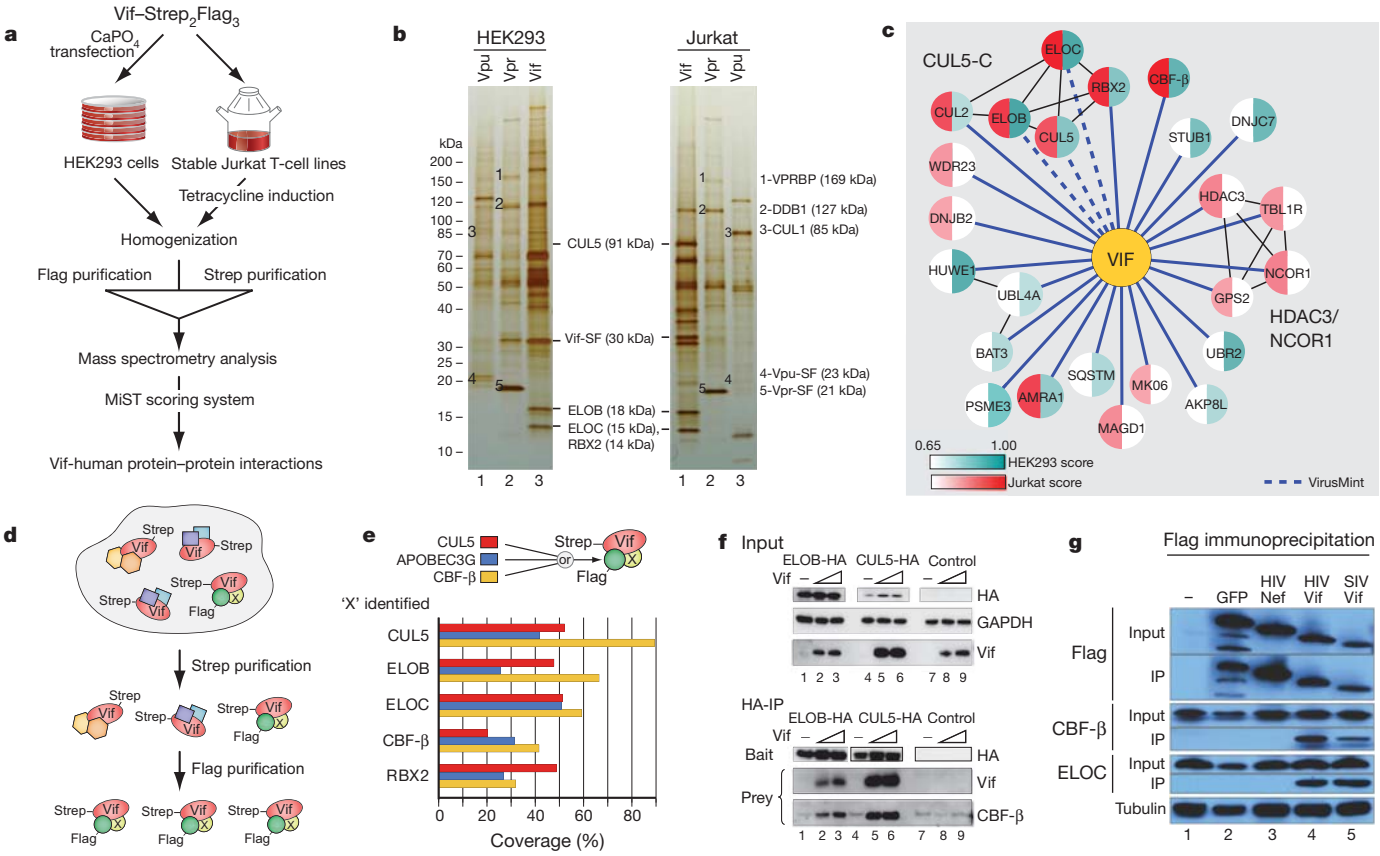


Figure 1 | AP–MS experiments identify CBF-β as a Vif-dependent component of the Vif–CUL5 ubiquitin ligase complex. a, Flow-chart of the proteomic analysis performed during the study. **b**, Affinity-tagged versions of Vif, Vpu and Vpr were purified using 3×Flag from HEK293 and Jurkat cells, subjected to SDS–PAGE and stained with silver. Visible bands corresponding to interactions that are known for each accessory factor are labelled. Note Vif and CBF-β run at a similar place on the gel. Tagged versions of Vpr and Vpu were used as specificity controls. **c**, A network representation of Vif–host protein–protein interactions from both HEK293 (blue) and Jurkat T cells (red) after subjecting the data derived from the AP–MS analysis to the MiST scoring system[15]. The intensity of the node colours corresponds to the quantitative MiST score. Blue edges represent interactions derived during this work; black edges are previously described interactions between host factors; dashed edges correspond to previously described Vif–host interactions present in the database VirusMint. **d**, The double purification approach, which allows for the identification of stable, stoichiometric protein complexes. **e**, Double purifications were performed in triplicate using 3×Flag-tagged CUL5, A3G or CBF-β with 2×Strep-tagged Vif in HEK293 cells. Proteins that were identified in all three double purifications, after trypsin digestion and analysis by mass spectrometry, are represented. The coverage corresponds to the percentage of protein identified by tryptic peptides. **f**, Immunoblots showing that Vif recruits CBF-β to the CUL5/ELOBC/RBX2 ubiquitin ligase complex. HA-tagged ELOB or CUL5 were immunoprecipitated in the absence or presence of increasing amounts of Vif, and endogenous CBF-β was monitored by immunoblot. **g**, HIV and SIV Vif co-immunoprecipitate CBF-β and ELOC. GFP and HIV Nef were analysed in parallel as specificity controls.

Vif–2×Strep and CBF-β–3×Flag. This strategy also yielded CUL5, ELOB, ELOC and RBX2, in addition to the epitope-tagged bait proteins (Fig. 1d, e).

To determine if the association of CBF-β with the CUL5 ligase complex was dependent on Vif, we immunoprecipitated CUL5-haemagglutinin (HA) or ELOB-HA in the presence or absence of Vif in HEK293 cells and blotted for endogenous CBF-β. Only in the presence of Vif did CBF-β co-immunoprecipitate with tagged CUL5 or ELOB, indicating that recruitment of CBF-β to the CUL5 ligase is dependent on Vif (Fig. 1f). SIV Vif also associated with CBF-β by immunoprecipitation, suggesting the interaction is conserved (Fig. 1g).

We next asked if the Vif–CUL5 ligase could be reconstituted with CBF-β using recombinant proteins purified from *Escherichia coli*. Initial purification attempts without CBF-β yielded aggregated and inactive complexes, assayed by size-exclusion chromatography and autoubiquitination activity, suggesting that CBF-β may be required for complex formation (data not shown). Therefore, full-length Vif, ELOB, ELOC and CBF-β were co-expressed, purified to homogeneity and found to form a stable, monodisperse complex with recombinant CUL5/RBX2, as shown by size-exclusion chromatography and SDS–PAGE analysis (Fig. 2a, b). Pull-down experiments performed with purified, His-tagged APOBEC3 enzymes immobilized on cobalt-chelating resin showed that the four protein complex containing Vif, CBF-β and ELOBC binds A3G, but not the related Vif-resistant deaminase, A3A (Fig. 2c). These observations suggested that Vif, CBF-β and ELOBC form a substrate adaptor for CUL5/RBX2 that enables specific interaction with susceptible A3 proteins.

To test the activity of the reconstituted six protein complex, CUL5/RBX2/ELOB/ELOC/Vif/CBF-β (CRL5–Vif–CBF-β), we assayed substrate and Vif ubiquitination activities using two distinct and well characterized ubiquitin conjugating enzymes, UBE2R1 (hCDC34a) and UBCH5b, which are capable of forming specific K48 and heterogenous ubiquitin chain linkages, respectively[19,20]. With UBE2R1, CRL5–Vif–CBF-β catalysed formation of high-molecular mass K48 chains on A3G, but not A3A (Fig. 2d, e), mirroring the chain linkage and substrate specificity observed in cells[4,6,21–23]. As with most ubiquitin ligase assemblies, the CRL5–Vif–CBF-β complex also possessed autoubiquitination activity that was only marginally affected by substrate A3s (Supplementary Fig. 2). These experiments were done with NEDD8-modified CUL5, because NEDD8ylation is required for CUL5 to degrade A3G *in vivo*[4] (Supplementary Fig. 3). Similarly, with UBCH5b, CRL5–Vif–CBF-β was able to promote the specific polyubiquitination of A3G and elicit Vif autoubiquitination activity (Supplementary Fig. 4). We conclude that the reconstituted Vif E3 ligase is specific for A3G,

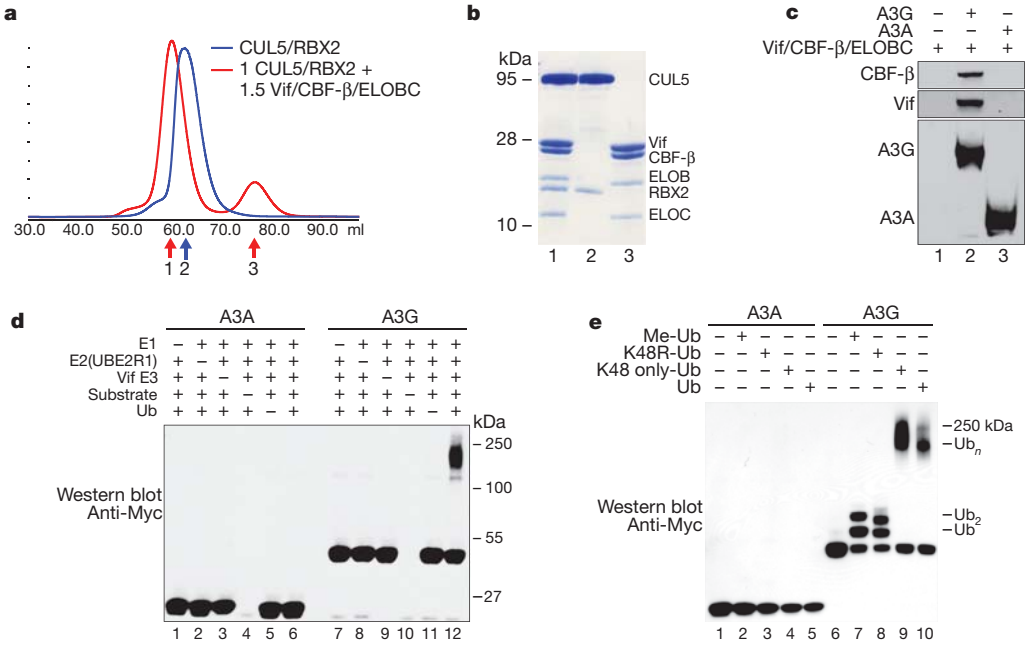


Figure 2 | CBF-β is a stoichiometric component of the Vif E3 ubiquitin ligase. **a**, Size exclusion chromatography of recombinant purified CUL5/RBX2 (blue) overlaid with CUL5/RBX2 mixed with 1.5 equivalents of purified Vif substrate adaptor containing Vif, ELOBC and CBF-β (red). **b**, Coomassie-stained SDS–PAGE of fractions labelled 1–3 in **a** indicating the Vif substrate adaptor and a six-protein assembly (CRL5–Vif–CBF-β) co-purify as stable monodisperse species. **c**, A3G, but not A3A, directly binds the tetrameric Vif substrate adaptor in pull-down experiments *in vitro*. **d**, CRL5–Vif–CBF-β is an E3 ligase that promotes polyubiquitination of A3G, but not A3A (detected using an anti-c-Myc antibody to the C-terminal tag on the deaminases). Ub, ubiquitin. **e**, CRL5–Vif–CBF-β and UBE2R1 catalyse formation of K48-linked chains on A3G. Immunoblots showing substrate in ubiquitination reactions containing UBE2R1 as E2, no ubiquitin, Me-ubiquitin, K48R-ubiquitin, K48-only ubiquitin or wild-type ubiquitin. Reactions with Me-ubiquitin indicate at least two distinct sites are modified on A3G; K48R recapitulates the pattern observed with Me-ubiquitin, whereas both wild type and K48R-only ubiquitin result in extensive polyubiquitin chains.

supports K48 chain formation and can function with at least two ubiquitin conjugating enzymes *in vitro*. It is conceivable that these two ubiquitin-conjugating enzymes work together in cells to promote multi-monoubiquitination of A3G followed by specific chain elongation, as described for other RING E3s[24,25], but additional work will be necessary to rule out other E2s *in vivo*.

To determine if CBF-β is required for Vif folding and/or stability in living cells, we transfected a constant amount of Vif into HEK293T cells expressing either a scrambled short hairpin (sh)RNA or a CBF-β-specific shRNA. The levels of steady-state Vif were threefold lower in CBF-β-depleted cells than in the scrambled control cells (Fig. 3a). Proteasome inhibitor MG132 reversed this effect, suggesting that Vif

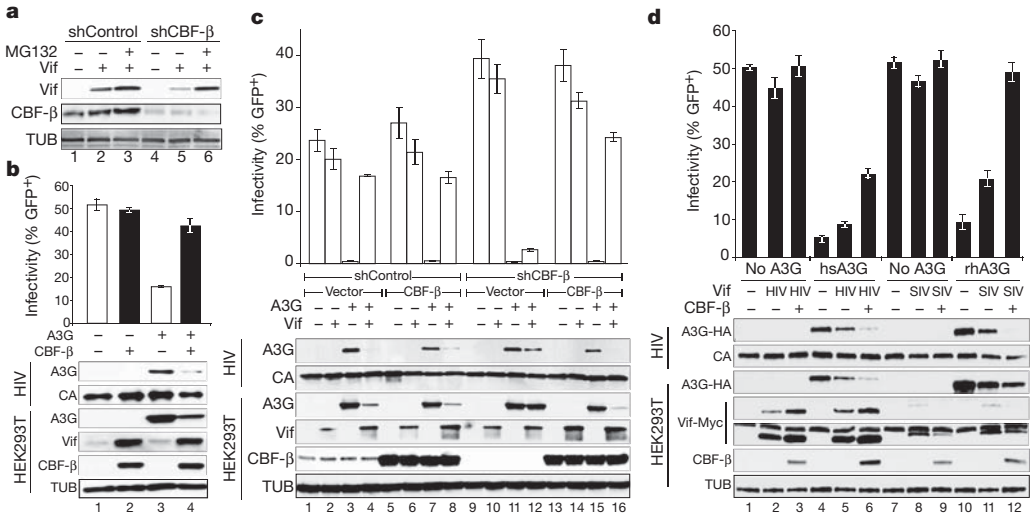


Figure 3 | CBF-β and Vif collaborate to degrade APOBEC3G and enable HIV-1 infectivity. **a**, CBF-β-depleted HEK293T cells have lower steady-state Vif levels, which recover upon treatment with 2.5 μM MG132. **b**, Infectivity of replication-competent, Vif-proficient HIV-1 in the presence and absence of CBF-β and A3G ($n = 3$; mean, s.d.). Immunoblots are shown for the indicated proteins in virus-producing cells and viral particles. **c**, Infectivity of HIV-GFP produced using HEK293T-shCBF-β or HEK293T-shControl clones transfected with the single-cycle virus cocktail, A3G, Vif and CBF-β as indicated ($n = 3$; mean, s.d.). The corresponding immunoblots are shown below. **d**, Infectivity of a Vif-deficient HIV-1 molecular clone produced in the presence or absence of human or rhesus A3G-HA, HIV or SIV Vif–Myc, and CBF-β as indicated ($n = 3$; mean, s.d.). Immunoblots are shown for the indicated proteins in virus-producing cells and viral particles with two exposures of the anti-Myc (Vif) blot shown to clarify the SIV Vif signal (the longer exposure also shows endogenous c-Myc).

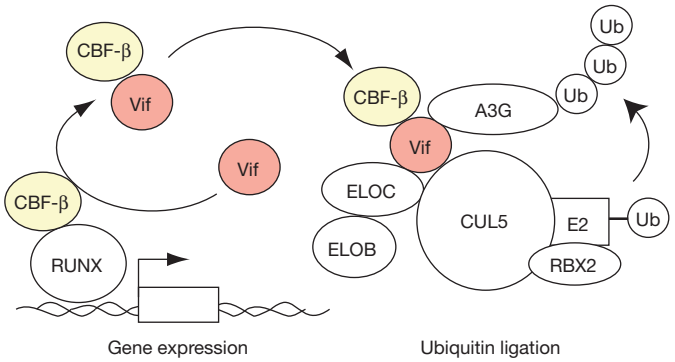


Figure 4 | Model for Vif–CBF-β E3 ligase formation and APOBEC3G polyubiquitination and degradation. Vif is depicted hijacking cellular CBF-β to the E3 ubiquitin ligase complex required for A3G polyubiquitination and degradation. Vif may recruit newly translated CBF-β (not shown) and/or hijack existing CBF-β from RUNX transcription complexes.

degradation is accelerated without CBF-β. Analogous data were obtained when Vif was expressed from a proviral plasmid in CBF-β-depleted cells and complemented with a CBF-β expression plasmid (Supplementary Fig. 5).

Based on these observations, we predicted that CBF-β knockdown should result in less functional Vif and less infectious HIV-1 particles when produced in the presence of A3G. To test this prediction, shRNA was used to deplete CBF-β stably in HEK293T cells, and a knockdown clone was used to produce replication competent Vif-proficient HIV-1 in the presence or absence of A3G and CBF-β expressed from plasmids (Fig. 3b). In CBF-β depleted cells, steady-state Vif levels were very low despite equivalent levels of virus production as indicated by capsid. Moreover, Vif levels increased when CBF-β was replenished by complementation, and this correlated with decreases in cellular and viral A3G levels and corresponding increases in viral infectivity. In the absence of A3G, no difference in infectivity was observed regardless of cellular CBF-β or Vif levels. Titration experiments showed that CBF-β complementation is dose-responsive (Supplementary Fig. 6). Analogous results were obtained with a multi-vector HIV/green fluorescent protein (GFP) system (Fig. 3c and Supplementary Fig. 7). The Vif/CBF-β interaction was confirmed in virus-producing cells by co-immunoprecipitation experiments (Supplementary Fig. 8). Furthermore, SIVmac239 Vif requires CBF-β to degrade rhesus macaque A3G and promote viral infectivity (Fig. 3d). Interestingly, in contrast to HIV Vif, lower steady-state levels of SIV Vif were observed in the presence of CBF-β, which may be functionally significant or may be a consequence of the heterologous assay system (that is, expressing SIV/rhesus proteins in human cells). Nevertheless, these results demonstrate the essential and conserved nature of CBF-β for Vif function in promoting A3G degradation and efficient virus replication.

Our proteomic, biochemical and genetic studies combine to suggest a model in which HIV-1 Vif hijacks the cellular transcription factor CBF-β to facilitate Vif folding and/or stability as well as nucleation of the rest of the E3 ubiquitin ligase complex (Fig. 4). CBF-β is required for A3G substrate binding and, ultimately, for polyubiquitination and degradation, thereby enabling the production of infectious viral particles. Because genetic studies have shown that Vif is also capable of degrading APOBEC3F and several other human APOBEC3 proteins[2,3,23] most of which are expressed in primary CD4+ T lymphocytes[26,27], it is quite likely that CBF-β is required for counteracting multiple endogenous APOBEC3s and thus for rendering T lymphocytes permissive for HIV-1 replication. We anticipate that the development of antiviral therapies that antagonize the CBF-β–Vif interaction will be more powerful than those that specifically target the A3G–Vif interaction, because they have the potential to unleash the simultaneous restriction potential of multiple APOBEC3s analogous to current combinatorial therapies.

METHODS SUMMARY

Affinity tagging, purification[14] and ubiquitination assays[28] were performed as described. Recombinant A3A and A3G were purified as Myc-His tagged proteins from HEK293 cells. HIV-1 infectivity studies used an HIV-1_{IIIB} proviral DNA construct (with or without Vif) or an HIV-GFP reporter plasmid set. Control (RHS4346) and CBF-β (RHS4430-99161432) shRNA constructs were obtained from Open Biosystems. A CBF-β complementary DNA matching NM_001755.2 was cloned from the CEM T cell line by RT–PCR. Immunoblots used antibodies to A3G (National Institutes of Health (NIH) ARRRP 10201 courtesy of J. Lingappa), CBF-β (Santa Cruz Biotechnology), HA (HA.11; Covance), TUB (tubulin; Covance), c-Myc (Sigma), Vif (NIH ARRRP 2221 courtesy of D. Gabuzda) and p24/capsid (NIH ARRRP 3537 courtesy of B. Chesebro and K. Wehrly). Details are provided in the Supplementary Methods.

Received 26 March; accepted 1 November 2011.
Published online 21 December 2011.

1. Goff, S. P. Retrovirus restriction factors. *Mol. Cell* **16,** 849–859 (2004).
2. Malim, M. H. & Emerman, M. HIV-1 accessory proteins—ensuring viral survival in a hostile environment. *Cell Host Microbe* **3,** 388–398 (2008).
3. Albin, J. S. & Harris, R. S. Interactions of host APOBEC3 restriction factors with HIV-1 *in vivo*: implications for therapeutics. *Expert Rev. Mol. Med.* **12,** e4 (2010).
4. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302,** 1056–1060 (2003).
5. Mehle, A., Goncalves, J., Santa-Marta, M., McPike, M. & Gabuzda, D. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation. *Genes Dev.* **18,** 2861–2866 (2004).
6. Mehle, A. *et al.* Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *J. Biol. Chem.* **279,** 7792–7798 (2004).
7. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418,** 646–650 (2002).
8. Sheehy, A. M., Gaddis, N. C. & Malim, M. H. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nature Med.* **9,** 1404–1407 (2003).
9. Conticello, S. G., Harris, R. S. & Neuberger, M. S. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Curr. Biol.* **13,** 2009–2013 (2003).
10. Stopak, K., de Noronha, C., Yonemoto, W. & Greene, W. C. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* **12,** 591–601 (2003).
11. Passmore, L. A. & Barford, D. Getting into position: the catalytic mechanisms of protein ubiquitylation. *Biochem. J.* **379,** 513–525 (2004).
12. Thrower, J. S., Hoffman, L., Rechsteiner, M. & Pickart, C. M. Recognition of the polyubiquitin proteolytic signal. *EMBO J.* **19,** 94–102 (2000).
13. Stanley, B. J. *et al.* Structural insight into the human immunodeficiency virus Vif SOCS box and its role in human E3 ubiquitin ligase assembly. *J. Virol.* **82,** 8656–8663 (2008).
14. Jäger, S. *et al.* Purification and characterization of HIV-human protein complexes. *Methods* **53,** 13–19 (2011).
15. Jäger, S. *et al.* Global landscape of HIV–human protein complexes. *Nature* doi:10.1038/nature10719 (this issue).
16. Kamura, T. *et al.* VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases. *Genes Dev.* **18,** 3055–3065 (2004).
17. Huang, D. T. *et al.* E2-RING expansion of the NEDD8 cascade confers specificity to cullin modification. *Mol. Cell* **33,** 483–495 (2009).
18. Wong, W. F., Kohu, K., Chiba, T., Sato, T. & Satake, M. Interplay of transcription factors in T-cell differentiation and function: the role of Runx. *Immunology* **132,** 157–164 (2011).
19. Petroski, M. D. & Deshaies, R. J. Mechanism of lysine 48-linked ubiquitin-chain synthesis by the cullin-RING ubiquitin-ligase complex SCF-Cdc34. *Cell* **123,** 1107–1120 (2005).
20. Kim, H. T. *et al.* Certain pairs of ubiquitin-conjugating enzymes (E2s) and ubiquitin-protein ligases (E3s) synthesize nondegradable forked ubiquitin chains containing all possible isopeptide linkages. *J. Biol. Chem.* **282,** 17375–17386 (2007).
21. DeHart, J. L., Bosque, A., Harris, R. S. & Planelles, V. Human immunodeficiency virus type 1 Vif induces cell cycle delay via recruitment of the same E3 ubiquitin ligase complex that targets APOBEC3 proteins for degradation. *J. Virol.* **82,** 9265–9272 (2008).
22. Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nature Struct. Mol. Biol.* **17,** 222–229 (2010).
23. Hultquist, J. F. *et al.* Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *J. Virol.* **85,** 11220–11234 (2011).
24. Rodrigo-Brenni, M. C. & Morgan, D. O. Sequential E2s drive polyubiquitin chain assembly on APC targets. *Cell* **130,** 127–139 (2007).
25. Wu, K., Kovacev, J. & Pan, Z. Q. Priming and extending: a UbcH5/Cdc34 E2 handoff mechanism for polyubiquitination on a SCF substrate. *Mol. Cell* **37,** 784–796 (2010).

26. Koning, F. A. *et al.* Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *J. Virol.* **83,** 9474–9485 (2009).
27. Refsland, E. W. *et al.* Quantitative profiling of the full *APOBEC3* mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res.* **38,** 4274–4284 (2010).
28. Saha, A. & Deshaies, R. J. Multimodal activation of the ubiquitin ligase SCF by Nedd8 conjugation. *Mol. Cell* **32,** 21–31 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Krogan, Gross and Harris laboratories for comments, J. R. Johnson for mass spectrometry, B. Leonard for sharing unpublished data, M. Shales for help with figures, and B. Chesebro, D. Gabuzda, J. Lingappa, M. Malim, K. Wehrly, X. F. Yu, A. Bullock, B. Schulman and the AIDS Research and Reference Reagent Program for reagents. This research was funded by grants from QB3 at University of California, San Francisco, and the National Institutes of Health (P50 GM082250, P01 AI090935 and P50 GM081879 to N.J.K.; U54 RR022220 to A.S.; R01 AI064046 and P01 GM091743 to R.S.H.; P50 GM082250 to J.D.G. and C.S.C.; P41RR001614 and P50GM081879 to A.B.). N.J.K. is a Searle Scholar and a Keck Young Investigator.

Author Contributions K.S., R.S.L. and E.K. made equal secondary contributions to this work. S.J. and K.F.S. generated the Vif protein–protein interaction map (Fig. 1); S.J. performed the co-immunoprecipitation confirmation assays (Supplementary Fig. 1); P.C. developed and implemented the MiST scoring system (Fig. 1c); S.J. and E.K. performed double purification analyses (Fig. 1c, d); C.M. and J.K. performed immunoprecipitation analyses (Fig. 1f, g); D.Y.K., L.Y. and D.S. reconstituted the Vif E3 ligase from recombinant components and performed ubiquitination assays (Fig. 2 and Supplementary Figs 2–4); M.L. expressed and purified A3 proteins and did *in vitro* pulldowns (Fig. 2); K.S., J.F.H. and R.S.L. performed CBF-β knockdown, complementation and virus infectivity experiments (Fig. 3 and Supplementary Figs 5–7); and B.D.A. and J.F.H. did the CBF-β–Vif co-immunoprecipitation experiments (Supplementary Fig. 8). A.B., A.S., C.S.C., R.S.H., J.D.G. and N.J.K. supervised the research; S.J., D.Y.K., J.F.H., K.S., R.S.H., J.D.G. and N.J.K. wrote and revised the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.S.H. (rsh@umn.edu), J.D.G. (jdgross@cgl.ucsf.edu) or N.J.K. (krogan@cmp.ucsf.edu).

LETTER

doi:10.1038/nature10718

T-cell differentiation factor CBF-β regulates HIV-1 Vif-mediated evasion of host restriction

Wenyan Zhang[1]*, Juan Du[1,2]*, Sean L. Evans[2], Yunkai Yu[2] & Xiao-Fang Yu[1,2]

The human APOBEC3 cytidine deaminases are potent inhibitors of diverse retroviruses, including human immunodeficiency virus-1 (HIV-1)[1–6]. HIV-1 Vif forms an E3 ubiquitin ligase complex with cullin 5 (CUL5), elongin B and elongin C [7–9], which promotes the polyubiquitination and degradation of APOBEC3 substrates[7,10–14]. Here we demonstrate in human T cells that core binding factor β (CBF-β) is a key regulator of the evasion of HIV-1 from the host defence mediated by APOBEC3. CBF-β, the non-DNA-binding subunit of a heterodimeric transcription factor, regulates the folding and DNA-binding activity of partner RUNX family proteins, which have important roles in the development and differentiation of diverse cell types, including T lymphocytes[15,16]. In our study, knockdown of endogenous CBF-β blocked Vif-induced APOBEC3G polyubiquitination and degradation. CBF-β was not required for the interaction between Vif and APOBEC3G, yet was essential for the assembly of the Vif–CUL5 E3-ubiquitin-ligase complex. CBF-β proved to be a unique regulator of primate lentiviral Vif and not a general component of the CUL5 E3 ubiquitin ligase. We show that Vif and CBF-β physically interact, and that the amino-terminal region of Vif is required for this interaction. Furthermore, interactions with Vif required regions in CBF-β that are not involved in RUNX protein binding[17–19]. Considering the importance of the interaction between Vif and CBF-β, disrupting this interaction represents an attractive pharmacological intervention against HIV-1.

To identify novel cellular proteins involved in HIV-1 Vif function, we infected human $CD4^+$ T cells (H9) with HIV-1 Vif–haemagglutinin (HA) or HIV-1 (control) virus[7] and then characterized the interaction of cellular proteins with HIV-1 Vif–HA in the infected cells by co-immunoprecipitation analysis. HIV-1 Vif forms an E3 ubiquitin ligase with CUL5 and elongin B/elongin C (ELOB/C) to promote the polyubiquitination and degradation of various substrates, including APOBEC3 and Vif itself[8,20]. To limit the potential degradation of these co-factors, we treated the cells with the proteasome inhibitor MG132 before co-immunoprecipitation. HA-tagged Vif was immunoprecipitated with anti-HA antibody from HIV-1 Vif–HA-infected H9 cell lysates but not HIV-1 (control)-infected H9 cell lysates (Fig. 1a). Immunoprecipitation of Vif–HA was confirmed by immunoblotting with a HA-specific antibody (Fig. 1b). Untagged Vif was not immunoprecipitated from control-infected H9 cells (Fig. 1b, lane 2). CUL5 (84 kDa) was co-precipitated with Vif–HA (Fig. 1a, lane 1) and confirmed by immunoblotting with a CUL5-specific antibody (Fig. 1b, lane 1); ELOB/C was also co-precipitated with Vif–HA (Fig. 1b, lane 1). These proteins were not co-precipitated from control HIV-1-infected H9 cells (Fig. 1b, lane 2).

A previously uncharacterized protein of ~22 kDa also co-precipitated with Vif–HA (Fig. 1a, lane 1). This protein was identified by mass spectrometry analysis as CBF-β, a ubiquitously expressed host protein[15,16]. The identification of the protein was confirmed by immunoblotting with a CBF-β-specific antibody (Fig. 1b, lane 1). We also detected an interaction between HIV-1 Vif–HA and CBF-β in transfected 293T (human embryonic kidney) cells (Fig. 1c, lane 4), indicating that the interaction between CBF-β and HIV-1 Vif can occur in the absence of

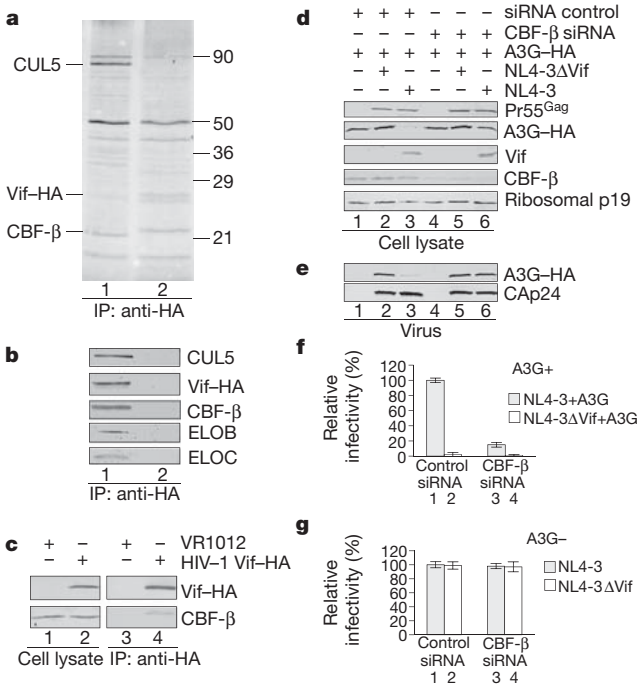


Figure 1 | Characterization of the functional interaction between HIV-1 Vif and CBF-β. a, Co-immunoprecipitation of cellular proteins with Vif–HA. Cell lysates from H9 cells infected with HIV-1 Vif–HA (HXB2Neo-Vif-HA, lane 1) or with control HIV-1 (that is, HXB2Neo containing untagged Vif, lane 2) were immunoprecipitated (IP) with anti-HA antibody conjugated to agarose beads, followed by SDS–PAGE and silver staining. The identification of CBF-β co-precipitated with Vif–HA (lane 1) was achieved by mass spectrometry. **b**, Immunoblot of precipitated samples from HIV-1- or HIV-1 Vif–HA-infected H9 cells using antibodies against CUL5, CBF-β, ELOB, ELOC or HA (to detect Vif–HA). **c**, Co-immunoprecipitation of endogenous CBF-β with Vif–HA from transfected 293T-cell samples. **d**, Vif-mediated downregulation of APOBEC3G (A3G) requires CBF-β. 293T cells were co-transfected with expression vectors for APOBEC3G–HA plus NL4-3, NL4-3ΔVif, or a control vector (pcDNA3.1) in the presence of control siRNA (lanes 1–3) or siRNA against CBF-β (lanes 4–6). **e**, CBF-β is required for Vif-mediated exclusion of APOBEC3G from virions. **e**, **f**, Supernatants from transfected 293T cells as described in **d** were collected and analysed for APOBEC3G–HA virion packaging (**e**) and viral infectivity (**f**). **f**, Virus infectivity was assessed by multinuclear activation of a galactosidase indicator (MAGI) assay. Virus input was normalized by the amount of CAp24. Virus infectivity with NL4-3 (Vif-positive) produced from transfected 293T cells in the presence of control siRNA was set to 100%. Error bars in **f** and **g** represent the standard deviations from triplicate wells. **g**, Silencing CBF-β expression had no effect on viral infectivity in the absence of APOBEC3G.

[1]First Hospital of Jilin University, Institute of Virology and AIDS Research, Changchun, Jilin Province 130021, China. [2]Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, Maryland 21205, USA.
*These authors contributed equally to this work.

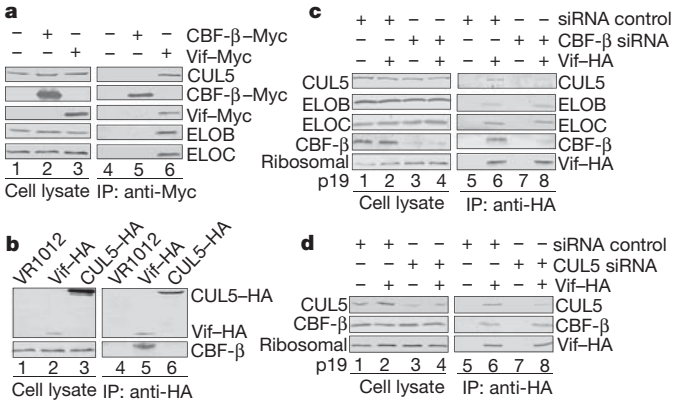


Figure 2 | CBF-β is critical for the interaction between Vif and CUL5. **a**–**d**, 293T cells were transfected with the indicated expression vectors. Cell lysates were prepared and immunoprecipitated followed by immunoblot analysis as described in Fig. 1. **a**, Co- immunoprecipitation of CUL5, ELOB and ELOC with Vif–Myc but not CBF-β–Myc. **b**, Co-immunoprecipitation of endogenous CBF-β with Vif–HA but not CUL5–HA. **c**, Silencing CBF-β inhibits the interaction between Vif and CUL5. **d**, Silencing CUL5 does not affect the Vif–CBF-β interaction.

other viral proteins. MG132 treatment increased the stability of HIV-1 Vif but did not alter its interaction with CBF-β (Supplementary Fig. 1); RNase treatment also did not block its interaction with CBF-β. Using a live cell imaging approach, we observed co-localization of yellow fluorescent protein (YFP)–Vif and cyan fluorescent protein (CFP)–CBF-β in transfected 293T cells along with a positive fluorescence resonance energy transfer (FRET) signal, further validating a physical interaction (Supplementary Fig. 2).

To determine whether CBF-β is required for the function of HIV-1 Vif, we knocked down endogenous CBF-β expression by CBF-β-specific short interfering RNA (siRNA). As compared to control siRNA-treated cells (Fig. 1d, lanes 1–3), CBF-β was reduced by ~70% in 293T cells treated with CBF-β-specific siRNA (lanes 4–6). In the control siRNA-treated cells, HIV-1 Vif efficiently reduced the expression of APOBEC3G (lane 3) when compared to HIV-1 NL4-3 strain Vif-negative (NL4-3ΔVif) virus-expressing cells (lane 2). Silencing of endogenous CBF-β (lanes 4–6) impeded the ability of Vif to reduce the expression of APOBEC3G (lane 6). Because HIV-1 Vif mainly reduces the expression of APOBEC3G by inducing its degradation, these data indicated that CBF-β is required for the HIV-1 Vif-mediated degradation of APOBEC3G (Fig. 1d, compare lanes 2 and 3 to lanes 5 and 6). Interestingly, expression of exogenous CFP–CBF-β enhanced both wild-type, untagged Vif (pcDNA-hVif)- and YFP–Vif-induced APOBEC3G degradation in 293T cells (Supplementary Fig. 2).

To exert its antiviral activity, APOBEC3G must be packaged into HIV-1 virions so that it can then inhibit virus infectivity in new target cells; Vif suppresses this antiviral activity by limiting the packaging of APOBEC3G into virions. Indeed, we found that when Vif was present, APOBEC3G was efficiently excluded from released HIV-1 virions in control siRNA-treated cells (Fig. 1e, lane 3); when Vif was absent, APOBEC3G could be packaged into HIV-1 virions (Fig. 1e, lane 2). However, the ability of Vif to exclude APOBEC3G from released HIV-1 virions was markedly compromised when CBF-β expression was knocked down in virus-producing cells (Fig. 1e, compare lanes 3 and 6). Reducing the expression of CBF-β specifically blocked the ability of HIV-1 Vif to suppress the antiviral activity of APOBEC3G (Fig. 1f). In contrast, silencing CBF-β expression had no effect on the viral infectivity of NL4-3 or NL4-3ΔVif when APOBEC3G was absent (Fig. 1g), indicating that CBF-β regulates HIV-1 infectivity only in the presence of APOBEC3G.

CBF-β apparently does not interact directly with APOBEC3G and is not required for the Vif–APOBEC3G interaction (Supplementary Fig. 3).

Human APOBEC3F is another human cytidine deaminase that has potent anti-HIV-1 activity; however, APOBEC3F and APOBEC3G are recognized by HIV-1 Vif through distinct mechanisms[21–27]. We postulated that if CBF-β is not involved in Vif-substrate recognition, it most likely regulates Vif mediated suppression of APOBEC3F as well as APOBEC3G. Indeed, HIV-1 Vif-mediated degradation, virion exclusion and suppression of APOBEC3F also required CBF-β (Supplementary Fig. 4). Thus, CBF-β is apparently required for Vif activity at a critical step(s) other than substrate recognition.

Because CBF-β does not affect the interaction between HIV-1 Vif and APOBEC3G, it is possible that CBF-β influences the ubiquitin ligase activity of Vif–CUL5 E3 and could in fact be a previously unrecognized component of the CUL5–ELOB/C E3 ubiquitin ligase. To determine whether CBF-β is an integral component of this complex, we examined the interaction of CBF-β with CUL5 and ELOB/C. Co-immunoprecipitation of CUL5 and ELOB/C with HIV-1 Vif–Myc (Fig. 2a, lane 6) was readily detected, but there was no co-immunoprecipitation of CUL5 or ELOB/C with CBF-β–Myc under the same conditions (lane 5). We also saw no interaction between endogenous CBF-β and CUL5–HA in a reciprocal co-immunoprecipitation analysis (Fig. 2b, lane 6), despite a readily detectable interaction between CBF-β and Vif–HA (lane 5). Therefore, CBF-β is unlikely to be an integral component of the CUL5–ELOB/C E3 complex under physiological conditions.

To determine whether CBF-β might have a role in the assembly of the Vif–CUL5 E3 ubiquitin ligase, we examined the interaction of Vif–HA with endogenous CUL5 and ELOB/C by co-immunoprecipitation analysis. Reducing CBF-β expression with specific siRNA had no effect on the intracellular expression of CUL5 or ELOB/C (Fig. 2c, compare lanes 1 and 2 to lanes 3 and 4). CUL5 and ELOB/C were co-immunoprecipitated with HIV-1 Vif–HA by the anti-HA antibody in control siRNA-treated cells (lane 6), but not in control cells that did not express HIV-1 Vif–HA (lanes 5 and 7). Co-immunoprecipitation of CUL5 with HIV-1 Vif–HA was significantly inhibited when CBF-β was silenced (Fig. 2c, lane 8) but co-precipitation of ELOB/C with HIV-1 Vif–HA was not (lane 8). Although silencing of CBF-β inhibited the interaction of Vif with CUL5 (Fig. 2c), reciprocal silencing of CUL5 (Fig. 2d, lane 4) had no

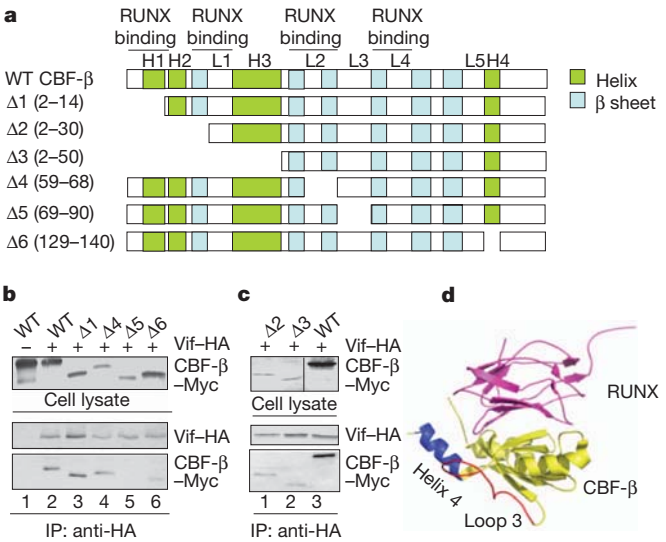


Figure 3 | Identification of regions in CBF-β is important for HIV-1 Vif interaction. **a**, Diagrams of mutant CBF-β constructs. The numbers in parentheses indicate amino acids that were deleted in the mutant CBF-β constructs. **b**, **c**, Co-precipitation of full-length (wild-type; WT) CBF-β–Myc and CBF-β mutants with HIV-1 Vif–HA. Cell lysates and precipitated samples were analysed by immunoblotting with an anti-Myc or anti-HA antibody. **d**, The structures of CBF-β (yellow) complexed with RUNX1 (pink) have been well characterized[17–19]. Loop 3 (red) and helix 4 (blue) of CBF-β, which are located away from the interfaces between CBF-β and RUNX1, are important for the Vif–CBF-β interaction.

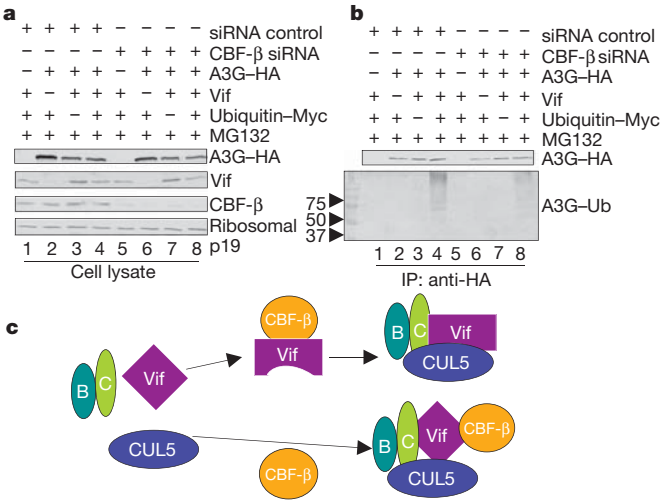


Figure 4 | Vif-induced polyubiquitination of APOBEC3G requires CBF-β. **a**–**c**, 293T cells were transfected with an expression vector encoding APOBEC3G–HA (A3G–HA), Vif, Myc-tagged ubiquitin (Ub) or control vector as indicated. All experiments were performed in the presence of MG132. **a**, Cells were lysed and analysed by immunoblot analysis using an anti-Vif, anti-CBF-β or anti-HA antibody. **b**, Cell lysates were immunoprecipitated with the anti-HA antibody and analysed by immunoblotting with an anti-Myc antibody to detect ubiquitinated APOBEC3G. Unmodified APOBEC3G–HA was detected with the anti-HA antibody. **c**, Proposed models for the role of CBF-β in HIV-1 Vif function.

detectable effect on the CBF-β–Vif interaction (Fig. 2d, lane 8). Again, these data indicate that CBF-β is not a general co-factor of CUL5, and CBF-β uniquely regulates the interaction of HIV-1 Vif with CUL5.

We also observed that the activity of the adenoviral E4orf6-containing CUL5–ELOB/C ubiquitin ligase was not affected by CBF-β silencing (Supplementary Fig. 5). In addition, cellular SOCS-box-containing proteins (that is, ASB6 and SOCS3) (Supplementary Fig. 5), which are known to assemble with the CUL5–ELOB/C E3 ubiquitin ligase, did not interact with CBF-β. Also, although primate lentiviral Vif interacted with CBF-β, Vif from bovine immunodeficiency virus (BIV) did not (Supplementary Fig. 6). BIV Vif-mediated bovine APOBEC3 depletion was not affected by CBF-β silencing.

CBF-β is a ubiquitously expressed component of a heterodimeric transcription factor including RUNX family proteins[15,16]. The interface between CBF-β and RUNX proteins is well defined[17–19]. To explore the regions of CBF-β and HIV-1 Vif that are critical for their interaction, we first examined the Vif-binding activity of a series of in-frame CBF-β deletion mutants (Fig. 3a). The regions of CBF-β that were important for the interaction with HIV-1 Vif mapped to loop 3 and helix 4 (Fig. 3b, c), regions that do not interact with RUNX proteins (Fig. 3d).

Consistent with the idea that HIV-1 Vif does not compete with RUNX for binding to CBF-β, we observed that the interaction of RUNX1 with CBF-β was not affected by the presence of Vif (Supplementary Fig. 7); in fact, a possible co-precipitation of Vif–HA with RUNX1–Myc was detected (Supplementary Fig. 7). RUNX1-mediated transcriptional activity[28] was also not significantly affected by Vif, as measured in a CBF promoter-driven luciferase reporter assay (Supplementary Fig. 7). We determined that the N-terminal regions of Vif, including amino acids W21 and W38, were important for CBF-β binding (Supplementary Fig. 8).

To determine whether CBF-β can influence the HIV-1 Vif-induced polyubiquitination of APOBEC3G, we transfected 293T cells with various combinations of expression vectors for HIV-1 Vif–Myc, APOBEC3G–HA and Myc-tagged ubiquitin in the presence of control siRNA or siRNA against CBF-β (Fig. 4a). To inhibit APOBEC3G degradation, all transfected cells were treated with MG132. HA-tagged APOBEC3G was immunoprecipitated by the anti-HA antibody, followed by the detection of ubiquitinated APOBEC3G–HA using the antibody against Myc-tagged ubiquitin. Polyubiquitinated APOBEC3G–HA was detected in cells expressing Vif (Fig. 4b, lane 4); however, the Vif-induced polyubiquitination of APOBEC3G was inhibited by siRNA against CBF-β (lane 8).

Taken together, our results demonstrate that CBF-β is a key regulator of HIV-1 Vif function. We have further shown that CBF-β is critical for Vif-induced endogenous APOBEC3G degradation and HIV-1 replication in H9 $CD4^+$ T cells (Supplementary Fig. 9). CBF-β allosterically regulates the conformation and DNA-binding activity of RUNX proteins[17–19]. It remains to be determined, however, whether CBF-β can induce conformational changes in HIV-1 Vif that allow CUL5 binding to occur. Alternatively, CBF-β may stabilize the interaction between Vif and CUL5 in the E3 ubiquitin ligase complex (Fig. 4c). Vif–CBF-β binding is critical for HIV-1 infection but does not seem to interfere with normal CBF-β–RUNX interaction. Thus, the interface between Vif and CBF-β may constitute a promising target for new anti-HIV drug discovery.

METHODS SUMMARY

Live-cell imaging and FRET assay. 293T cells were transfected with YFP–Vif (2 μg) and CFP–CBF-β (0.25 μg) using Lipofectamine 2000 (Invitrogen) in 6-well coverslip glass-bottom cell culture dishes (MatTek) at 37 °C with 5% CO_2. Eightfold excess YFP–Vif was used due to low expression of YFP–Vif compared to CFP–CBF-β (Supplementary Fig. 2d). YFP–CFP dimer (0.25 μg), co-transfection of YFP (0.25 μg) plus CFP (0.25 μg), and CFP–CBF-β alone were used for a positive and negative FRET signal as imaging controls. Cells were visualized 24–36 h after transfection using a Zeiss LSM510-Meta confocal imaging system equipped with four argon lasers (458, 477, 488 and 514 nm lines), two HeNe lasers (542 and 633 nm), and one diode laser (405 nm). YFP and CFP were excited using a 514 nm and 458 nm laser, respectively. After excitation, channel mode images were collected using a long-pass filter for CFP emission (LP 420) or band-pass filter for YFP emission (BP 530-600). In addition, the channel mode detector gain was set so that neither CFP nor YFP images contained saturated pixels before bleaching. A total of ten images (five pre-bleach and five post-bleach) were acquired for each cell with a ×63 objective with ×2 zoom, and image analysis and manipulation was performed using Zeiss LSM imaging software. FRET analysis was performed using the acceptor photo-bleaching technique: YFP or YFP–Vif was photo-bleached for approximately 5 min at 514 nm (100% laser power), and the average donor signal was evaluated before and after acceptor photobleaching. As a control, cells were also imaged under similar conditions, except the 514 nm laser was set at 0% transmission to prevent acceptor photobleaching. Two experiments were performed for each bleached sample for a total of ten cells each. One experiment was performed for each unbleached sample for a total of four cells each. The FRET efficiency percentage (E) was calculated using the FRET macro of Zeiss LSM software. The following equation was used: $E = (F_{D\Delta}/F_{Dpost}) \times 100$, where $F_{D\Delta}$ is the change in donor (CFP) fluorescence intensity before and after acceptor (YFP) bleaching and F_{Dpost} is donor fluorescence intensity after acceptor bleaching. The FRET macro of Zeiss LSM software also subtracts background when calculating E.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 May; accepted 17 November 2011.
Published online 21 December 2011.

1. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418,** 646–650 (2002).
2. Navarro, F. & Landau, N. R. Recent insights into HIV-1 Vif. *Curr. Opin. Immunol.* **16,** 477–482 (2004).
3. Cullen, B. R. Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. *J. Virol.* **80,** 1067–1076 (2006).
4. Chiu, Y. L. & Greene, W. C. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu. Rev. Immunol.* **26,** 317–353 (2008).
5. Malim, M. H. & Emerman, M. HIV-1 accessory proteins—ensuring viral survival in a hostile environment. *Cell Host Microbe* **3,** 388–398 (2008).
6. Goila-Gaur, R. & Strebel, K. HIV-1 Vif, APOBEC, and intrinsic immunity. *Retrovirology* **5,** 51 (2008).
7. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-CUL5-SCF complex. *Science* **302,** 1056–1060 (2003).

8. Mehle, A., Goncalves, J., Santa-Marta, M., McPike, M. & Gabuzda, D. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-CUL5 complex that promotes APOBEC3G degradation. *Genes Dev.* **18,** 2861–2866 (2004).
9. Yu, Y., Xiao, Z., Ehrlich, E. S., Yu, X. & Yu, X. F. Selective assembly of HIV-1 Vif-CUL5-ElonginB-ElonginC E3 ubiquitin ligase complex through a novel SOCS box and upstream cysteines. *Genes Dev.* **18,** 2867–2872 (2004).
10. Stopak, K., de Noronha, C., Yonemoto, W. & Greene, W. C. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* **12,** 591–601 (2003).
11. Marin, M., Rose, K. M., Kozak, S. L. & Kabat, D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nature Med.* **9,** 1398–1403 (2003).
12. Conticello, S. G., Harris, R. S. & Neuberger, M. S. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Curr. Biol.* **13,** 2009–2013 (2003).
13. Sheehy, A. M., Gaddis, N. C. & Malim, M. H. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nature Med.* **9,** 1404–1407 (2003).
14. Mehle, A. *et al.* Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *J. Biol. Chem.* **279,** 7792–7798 (2003).
15. de Bruijn, M. F. & Speck, N. A. Core-binding factors in hematopoiesis and immune function. *Oncogene* **23,** 4238–4248 (2004).
16. Ito, Y. RUNX genes in development and cancer: regulation of viral gene expression and the discovery of RUNX family genes. *Adv. Cancer Res.* **99,** 33–76 (2008).
17. Tahirov, T. H. *et al.* Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFβ. *Cell* **104,** 755–767 (2001).
18. Bravo, J., Li, Z., Speck, N. A. & Warren, A. J. The leukemia-associated AML1 (Runx1)–CBFβ complex functions as a DNA-induced molecular clamp. *Nature Struct. Biol.* **8,** 371–378 (2001).
19. Yan, J., Liu, Y., Lukasik, S. M., Speck, N. A. & Bushweller, J. H. CBFβ allosterically regulates the Runx1 Runt domain via a dynamic conformational equilibrium. *Nature Struct. Mol. Biol.* **11,** 901–906 (2004).
20. Liu, B., Sarkis, P. T., Luo, K., Yu, Y. & Yu, X. F. Regulation of Apobec3F and human immunodeficiency virus type 1 Vif by Vif-CUL5-ElonB/C E3 ubiquitin ligase. *J. Virol.* **79,** 9579–9587 (2005).
21. Simon, V. *et al.* Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. *PLoS Pathog.* **1,** e6 (2005).
22. Tian, C. *et al.* Differential requirement for conserved tryptophans in human immunodeficiency virus type 1 Vif for the selective suppression of APOBEC3G and APOBEC3F. *J. Virol.* **80,** 3112–3115 (2006).
23. Schröfelbauer, B., Senger, T., Manning, G. & Landau, N. R. Mutational alteration of human immunodeficiency virus type 1 Vif allows for functional interaction with nonhuman primate APOBEC3G. *J. Virol.* **80,** 5984–5991 (2006).
24. Russell, R. A. & Pathak, V. K. Identification of two distinct human immunodeficiency virus type 1 Vif determinants critical for interactions with human APOBEC3G and APOBEC3F. *J. Virol.* **81,** 8201–8210 (2007).
25. Mehle, A. *et al.* Identification of an APOBEC3G binding site in human immunodeficiency virus type 1 Vif and inhibitors of Vif-APOBEC3G binding. *J. Virol.* **81,** 13235–13241 (2007).
26. He, Z., Zhang, W., Chen, G., Xu, R. & Yu, X. F. Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction. *J. Mol. Biol.* **381,** 1000–1011 (2008).
27. Zhang, W., Chen, G., Niewiadomska, A. M., Xu, R. & Yu, X. F. Distinct determinants in HIV-1 Vif and human APOBEC3 proteins are required for the suppression of diverse host anti-viral proteins. *PLoS ONE* **3,** e3963 (2008).
28. Cao, W., Adya, N., Britos-Bray, M., Liu, P. P. & Friedman, A. D. The core binding factor (CBF) α interaction domain and the smooth muscle myosin heavy chain (SMMHC) segment of CBFβ-SMMHC are both required to slow cell proliferation. *J. Biol. Chem.* **273,** 31534–31540 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Strebel, A. Friedman, N. Speck, D. Yue, R. Siliciano, M. Malim, R. Harris, T. Inoue and D. Gabuzda for critical reagents; T. Wang, K. Zhao, X. Zhou and A. Zhen for technical assistance; R. Markham, J. Margolick and J. Bream for thoughtful discussions; and D. McClellan for editorial assistance. We also wish to thank staff within the Mass Spectrometry Core and the Institute for Basic Biomedical Sciences Microscope Facility at Johns Hopkins School of Medicine for their technical assistance. The following reagents were obtained through the AIDS Research and Reference Reagents Program, Division of AIDS, NIAID, NIH: monoclonal antibodies against HIV-1 p24 (B. Chesebro and H. Chen), pNL4-3 (M. Martin), pcDNA-hVif (S. Bour and K. Strebel), antiserum to HIV-1 Vif (D. Gabuzda) and MAGI-CCR5 cells (J. Overbaugh). This work was supported in part by funding from the Chinese Ministry of Science and Technology (2012CB911100) and Chinese Ministry of Education (IRT1016), the Key Laboratory of Molecular Virology, Jilin Province (20102209), China, and a grant (2R56AI62644-6) from the NIAID.

Author Contributions W.Z., J.D., S.L.E. and Y.Y. performed experiments and analysed the data. X.-F.Y. directed the project, analysed the data, and wrote the paper with help from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to X-F.Y. (xfyu@jhsph.edu).

METHODS

Plasmid construction. CBF-β coding sequences were amplified by reverse transcription and PCR using mRNA samples from H9 cells with the following primers: forward 5′-GCTAGCAAGATGCCGCGCGTCGTG-3′, reverse 5′-AAGCTTACTACAGATCTTCTTCTGATATGAGTTTTTGTTCGGGTCTTGTTGTCTTCTTGCC-3′, containing NheI and HindIII sites. The PCR product was cloned into pcDNA3.1 to generate pCBF-β-Myc. pCBF-βΔN1, CBF-βΔN2, CBF-βΔN3, CBF-βΔN4, CBF-βΔN5 and CBF-βΔN6 were made from pCBF-β-Myc by site-directed mutagenesis and confirmed by DNA sequencing. The following expression vectors have been previously described[7,20,29,30]: VR1012, HIV-1 Vif expression vectors pVif-Myc and pVif-HA, pAPOBEC3G-HA, pAPOBEC3G-Myc, pUbiquitin-Myc, pCUL5-HA, pAPOBEC3F-V5, E4orf6-Myc, SOCS3-HA, human p53, SIV_{MAC}Vif-HA, SIV_{AGM}Vif-Myc, BIVVif-HA, HXB2Neo and HXB2NeoVif-HA. pVifΔN-HA was constructed with the following primers: forward 5′-GTCGACATGGACCCTGAACTAGC-3′, reverse 5′-GGAATTCCTACGCGTAATCTGGGACGTCGTAAGGGTAGTGTTCCATTCATTGTGTGGCT-3′, containing SalI and NotI sites, respectively, and a C-terminal HA tag. pNL4-3ΔVif and pcDNA-hVif were gifts from K. Strebel. pRUNX1-Myc and $p(CBF)_4$TKLuc were gifts from A. Friedman. NL43-ΔE-EGFP was a gift from R. Silicano. Plasmids pVifW11A, W21A, W38A, DR14/15AA, K22E, Y40A, RH41/42AA, T74A, E76A, R77A and W79A were made from pVif-Myc by site-directed mutagenesis and confirmed by DNA sequencing. To generate YFP N-terminal epitope-tagged NL4-3 Vif, pcDNA-hVif was used to PCR amplify the Vif coding region and cloned into the NotI and XbaI sites of pcDNA3-YFP (D. T. Yue). To generate CFP N-terminal epitope-tagged CBF-β, the CBF-β coding region was cloned into the NotI and XbaI sites of pcDNA3-CFP (D. T. Yue). The YFP–CFP dimer construct was gift from D. T. Yue and was generated by fusing CFP and YFP with a 21 amino acid linker, SRAQASNSAVDGTAGPGSIAT. The following CBF-β shRNA clones were obtained from Open Biosystems: TRCN0000016643, with mature sense sequence 5′-CGAGAGTATGTCGACTTAGAA-3′; TRCN0000016644, 5′-CCGCGAGTGTGAGATTAAGTA-3′; TRCN0000016645, 5′-GAAGATAGAGACAGGTCTCAT-3′; TRCN0000016646, 5′-GCTGGCAGTAACTGGCAAGAA-3′; TRCN0000016647, 5′-TGAGATTAAGTACACGGGCT-3′.

Antibodies and cell lines. The H9 human $CD4^+$ T-cell line was maintained in RPMI-1640 medium (Invitrogen) with 10% fetal bovine serum and penicillin/streptomycin (R-10 medium). 293T and MAGI-CCR5 cells (AIDS Research and Reference Reagents Program, catalogue no. 3522) were maintained in Dulbecco's modified Eagle's medium (DMEM; Invitrogen) with 10% fetal bovine serum and penicillin/streptomycin (D-10 medium) and passaged when confluent. The following antibodies were used: anti-HA (Covance, MMS-101R-1000), anti-Myc (Upstate, 05-724), anti-V5 (Invitrogen, R960-25), anti-β-actin (Sigma, A3853), anti-histone H3 (Genscript, A01502), anti-human ribosomal P antigens (Immunovision, HP0-0100), anti-HA antibody–agarose conjugate (Roche, 11815016001), anti-ELOB (Santa Cruz, sc-1558), anti-ELOC (BD Transduction Lab, SIII/P15, 610760), anti-GAPDH (Sigma, G8795), anti-CBF-β (Abcam, ab11921), anti-CUL5 (Santa Cruz, sc-13014), anti-p53 (Oncogene Research Products, OP03) and anti-GFP (Genscript, A01704); anti-CAp24 (catalogue no. 1513) and anti-Vif antibodies (catalogue no. 2221) were obtained from the AIDS Research and Reference Reagents Program.

RNA interference. RNAi against CBF-β was carried out using a pool of two duplexed siRNAs (Dharmacon): duplex 1, sense, 5′-CCAGCAGGAGGAUGCAUUAUU; antisense, 5′-PUAAUGCAUCCUCCUGCUGGUU; duplex 2, sense, 5′-GCAGGCAAGGUAUAUUUGAUU; antisense, 5′-PUCAAAUAUACCUUGCCUGCUU. RNAi against CUL5 was carried out using a pool of four duplexed siRNAs (Dharmacon): duplex 1, sense, 5′-GACACGACGUCUUAUAUUAUU; antisense, 5′-PUAAUAUAAGACGUCGUGUCUU; duplex 2, sense, 5′-CGUCUAAUCUGUUAAAGAAUU; antisense, 5′-PUUCUUUAACAGAUUAGACGUU; duplex 3, sense, 5′-GAUGAUACGGCUUUGCUAAUU; antisense, 5′-PUUAGCAAAGCCGUAUCAUCUU; duplex 4, sense, 5′-GUUCAACUACGAAUACUAAUU; antisense, 5′-PUUAGUAUUCGUAGUUGAACUU. 293T cells were transfected with the CBF-β or CUL5 siRNA pool at a total final concentration of 100 nM using Lipofectamine 2000 (Invitrogen). The non-targeting siRNA no. 2 (Dharmacon) was used as a control. Protein expression was monitored by immunoblotting 2–3 days after transfection.

Identification of Vif-binding proteins. Vif-containing complexes were purified from HIV-1 Vif–HA and control HIV-1 (HXB2Neo) infected H9 cells by immunoprecipitation and analysed by SDS–PAGE. Gels were fixed in 50% methanol/10% acetic acid for 10 min, stained with mass spectrometry-compatible colloidal-Coomassie brilliant blue G-250 (Bio-Rad 1610406) staining solution (20% methanol, 8% ammonium sulphate, 1.6% phosphoric acid, 0.08% Coomassie blue G-250) to detect protein bands, and de-stained with distilled water. Protein standards from Bio-Rad (catalogue no. 1610314) were used to estimate protein size. Protein bands of interest were cut out of the gel and rinsed twice with 50% methanol (HPLC grade). In-gel digestion was performed on protein bands cut out of colloidal-Coomassie blue-stained SDS–polyacrylamide gels using sequencing-grade modified trypsin (Promega). Extracted peptides were co-crystallized in 2,5-dihydroxybenzoic acid (DHB) or α-cyano-4-hydroxycinnamic acid (CHCA) (10 mg ml^{-1} in 50% acetonitrile/0.3% TFA) and analysed by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) spectrometry on a Voyager DE STR (Applied Biosystems, http://www.appliedbiosystems.com) using Voyager Instrument Control Panel (v. 5.1) and Data Explorer (v. 4.0) software. Data were acquired in reflector mode, and masses were externally calibrated using a standard peptide mixture to $<$50 p.p.m. error. Proteins were identified by searching the acquired monoisotopic masses against the NCBI non-redundant or SwissProt databases using the MS-Fit search engine of ProteinProspector (http://prospector.ucsf.edu).

APOBEC3G ubiquitination assay. 293T cells were transfected with expression vectors encoding APOBEC3G–HA, HIV-1 Vif and ubiquitin–Myc, either individually or in combination. Cells were also transfected with control or CBF-β-specific siRNA. Transfected cells were treated with 10 μM MG132 (proteasome inhibitor) for 16 h, beginning at 24 h after transfection, and then lysed in lysis buffer (50 mM Tris, pH 7.5, with 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, 10 μM MG132 and complete protease inhibitor cocktail tablets), followed by centrifugation at 10,000*g* for 30 min. Lysates were applied to anti-HA antibody-conjugated agarose beads (Roche) and incubated for 3 h at 4 °C. After incubation, the beads were washed six times with washing buffer (20 mM Tris, pH 7.5, with 0.1 M NaCl, 0.1 mM EDTA and 0.05% Tween 20), then eluted with elution buffer (0.1 M glycine-HCl, pH 2.0) followed by SDS–PAGE and immunoblotting with anti-Myc and anti-HA antibodies.

Transfection, virus purification, immunoblot analysis, immunoprecipitation and viral infectivity (MAGI) assays were performed as previously described[7,29].

29. Luo, K. *et al.* Primate lentiviral virion infectivity factors are substrate receptors that assemble with cullin 5–E3 ligase through a HCCH motif to suppress APOBEC3G. *Proc. Natl Acad. Sci. USA* **102,** 11444–11449 (2005).
30. Luo, K. *et al.* Adenovirus E4orf6 assembles with Cullin5-ElonginB-ElonginC E3 ubiquitin ligase through an HIV/SIV Vif-like BC-box to regulate p53. *FASEB J.* **21,** 1742–1750 (2007).

LETTER

doi:10.1038/nature10602

Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia

Christian M. Metallo[1]†, Paulo A. Gameiro[1,2,3,4], Eric L. Bell[5], Katherine R. Mattaini[5,6], Juanjuan Yang[3,4], Karsten Hiller[1]†, Christopher M. Jewell[6], Zachary R. Johnson[6], Darrell J. Irvine[6,7], Leonard Guarente[5], Joanne K. Kelleher[1], Matthew G. Vander Heiden[5,6,8], Othon Iliopoulos[3,4] & Gregory Stephanopoulos[1]

Acetyl coenzyme A (AcCoA) is the central biosynthetic precursor for fatty-acid synthesis and protein acetylation. In the conventional view of mammalian cell metabolism, AcCoA is primarily generated from glucose-derived pyruvate through the citrate shuttle and ATP citrate lyase in the cytosol[1–3]. However, proliferating cells that exhibit aerobic glycolysis and those exposed to hypoxia convert glucose to lactate at near-stoichiometric levels, directing glucose carbon away from the tricarboxylic acid cycle and fatty-acid synthesis[4]. Although glutamine is consumed at levels exceeding that required for nitrogen biosynthesis[5], the regulation and use of glutamine metabolism in hypoxic cells is not well understood. Here we show that human cells use reductive metabolism of α-ketoglutarate to synthesize AcCoA for lipid synthesis. This isocitrate dehydrogenase-1 (IDH1)-dependent pathway is active in most cell lines under normal culture conditions, but cells grown under hypoxia rely almost exclusively on the reductive carboxylation of glutamine-derived α-ketoglutarate for *de novo* lipogenesis. Furthermore, renal cell lines deficient in the von Hippel–Lindau tumour suppressor protein preferentially use reductive glutamine metabolism for lipid biosynthesis even at normal oxygen levels. These results identify a critical role for oxygen in regulating carbon use to produce AcCoA and support lipid synthesis in mammalian cells.

Although hypoxic cells exhibit a shift towards glycolytic metabolism[4], a functional electron transport chain and glutamine-derived carbon are required for proliferation of most transformed cells[6]. In line with these studies, we observed increased glutamine consumption when A549 cells were cultured at approximately 1% oxygen while glutamate secretion remained unchanged (Supplementary Fig. 1a), indicating that net glutamine consumption was elevated and suggesting that glutamine carbon is used for biosynthesis. Notably, glucose consumption and lactate secretion also increased in these experiments. Consistent with this observation, we found that proliferating cells under both normoxia and hypoxia incorporate glutamine-derived carbon into lipids (Supplementary Fig. 1b). Glutamine can contribute carbon to lipogenic AcCoA through two distinct pathways. Cells can oxidatively metabolize glutamine-derived α-ketoglutarate (αKG) in the tricarboxylic acid (TCA) cycle and generate pyruvate from malate by glutaminolysis[5]. Alternatively, some tissues can reductively carboxylate αKG to generate citrate[7,8], and recent studies have indicated that the IDH reaction is highly reversible[9–11]. To determine which pathway cells use to incorporate glutamine carbon into lipids we used stable isotopic tracers[12–14]. We first cultured several cancer cell lines with [1-^{13}C]glutamine under normoxia and quantified the isotopic label present in metabolite pools along this pathway (Fig. 1a, red carbon atoms). All cells tested with this tracer retained significant label from [1-^{13}C]glutamine in citrate and metabolites downstream of the irreversible ATP citrate lyase reaction, indicating that the reductive flux contributes to the cytosolic AcCoA pool (Supplementary Fig. 2). Additional evidence for activity along this pathway was obtained with a uniformly ^{13}C-labelled ([U-$^{13}C_5$]) glutamine tracer (Supplementary Fig. 3).

To quantify the specific contributions of oxidative and reductive glutamine metabolism to fatty-acid synthesis, we cultured cells in the presence of tracers for several days and performed isotopomer spectral analysis (ISA; Supplementary Fig. 4)[15]. Use of [5-^{13}C]glutamine specifically allows estimation of the flux of glutamine to lipids through reductive carboxylation (Fig. 1a, blue carbon atoms). Virtually all cell lines cultured with this tracer generated labelled fatty acids, metabolizing glutamine reductively in the TCA cycle to supply 10–25% of their lipogenic AcCoA (Fig. 1b and Supplementary Fig. 5). Consistent with these data, we were able to detect 98 ± 5 c.p.m. per 10^6 cells in hexane extracts of A549 cells cultured with [5-^{14}C]glutamine. Next we directly compared the contribution of glutamine to fatty acids through the reductive flux as a fraction of the total, the latter determined by using [U-$^{13}C_5$]glutamine. In all cell lines tested, including those derived from lung, mammary, colon and squamous cell carcinoma as well as melanoma, glioblastoma and leukaemia, [5-$^{13}C_5$]glutamine labelled most of the AcCoA derived from glutamine (Fig. 1b), highlighting the general use of reductive carboxylation as the primary route through which glutamine, glutamate and αKG carbon are converted to lipids in cultured cells (Fig. 1b and Supplementary Figs 2, 3 and 5). The glutaminolysis pathway is also an important means of glutamine catabolism and can be characterized by quantifying the contribution of glutamine carbon to lactate. Consistent with published reports[5], glutamine-derived ^{13}C label was also detected in lactate, and the amount of ^{13}C-labelled lactate produced was highest in glioblastoma-derived cells compared with other cells cultured with [U-$^{13}C_5$]glutamine (Supplementary Fig. 6).

Mammalian cells express three IDH enzymes encoded by separate genes: IDH1 (cytosolic, $NADP^+$-dependent), IDH2 (mitochondrial, $NADP^+$-dependent) and the multi-subunit enzyme IDH3 (mitochondrial, NAD^+-dependent). IDH3 is allosterically regulated and is thought to operate in the oxidative direction. The $NADP^+$-dependent isozymes are capable of catalysing the reductive reaction; however, the specific enzyme responsible for this flux is not definitively known[16]. As measurements of metabolite pools in subcellular compartments and labelling therein cannot yet be reliably obtained, we used RNA interference to knockdown expression of IDH1 and IDH2 selectively in A549 cells. Using labelling from [1-^{13}C]glutamine as a readout, we measured a significant and robust decrease in reductive carboxylation when IDH1 messenger RNA (mRNA) was targeted using short hairpin RNAs (shRNAs) (Fig. 1c, d). These changes were consistent with results using [U-$^{13}C_5$]glutamine (Supplementary Fig. 7) and reproduced

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Department of Life Sciences, University of Coimbra, 3004-517 Coimbra, Portugal. [3]Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. [4]Massachusetts General Hospital Center for Cancer Research, Charlestown, Massachusetts 02129, USA. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Koch Institute for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [7]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. [8]Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. †Present addresses: Department of Bioengineering, University of California at San Diego, La Jolla, California 92093, USA (C.M.M.); Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-1511, Luxembourg (K.H.).

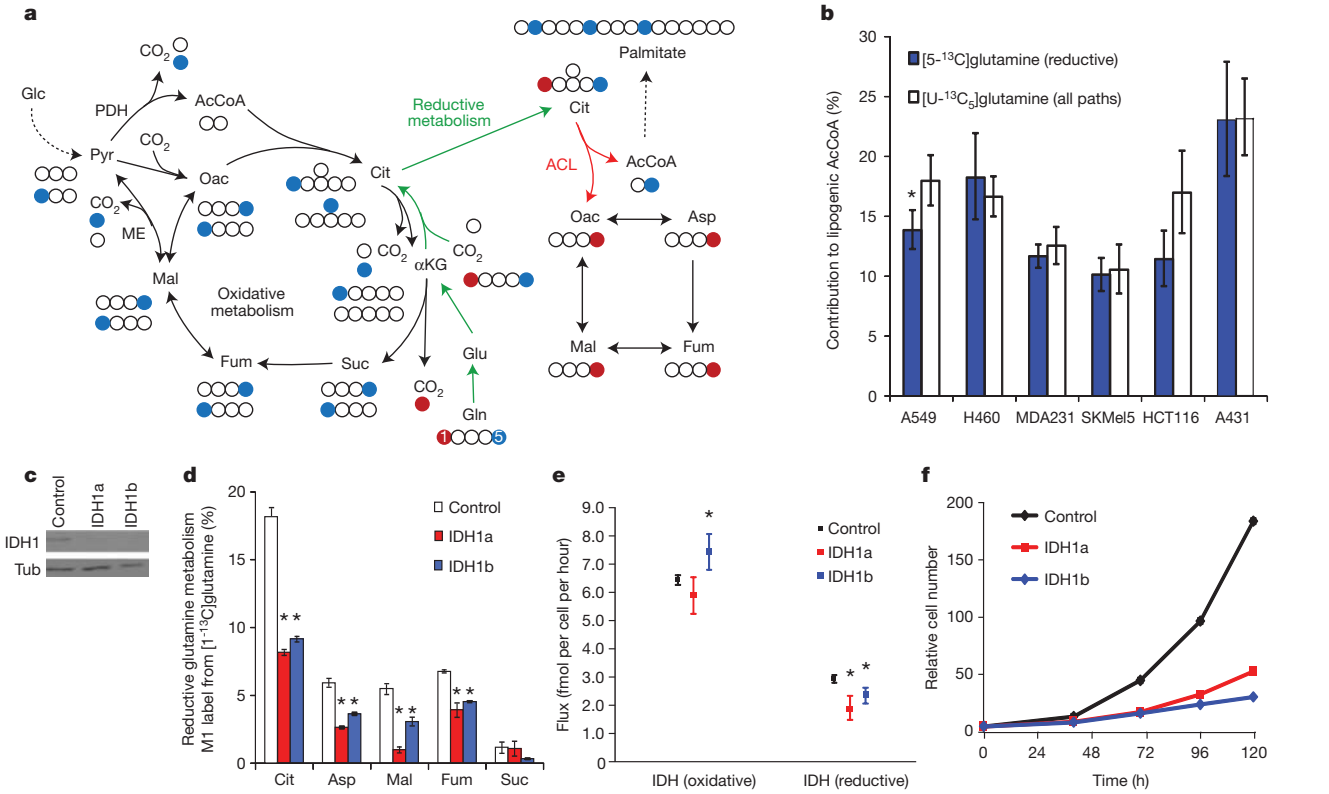


Figure 1 | Reductive carboxylation is the primary route of glutamine to lipids. **a**, Schematic of carbon atom (circles) transitions and tracers used to detect reductive glutamine metabolism. Isotopic label from [1-^{13}C]glutamine (red) is lost during oxidative conversion to succinate (Suc) but retained on citrate (Cit), oxaloacetate (Oac), aspartate (Asp), malate (Mal) and fumarate (Fum) in the reductive pathway (green arrows). [5-^{13}C]glutamine (blue) transfers label to AcCoA through reductive metabolism only. Molecular symmetry is shown for oxidative metabolism. Glc, glucose; Pyr, pyruvate; ME, malic enzyme; PDH, pyruvate dehydrogenase; Glu, glutamate; Gln, glutamine. **b**, Contribution of [5-^{13}C]glutamine and [U-^{13}C$_5$]glutamine to lipogenic AcCoA in cell lines. **c**, IDH1 levels in A549 cells expressing IDH1-specific (IDH1a and IDH1b) or control shRNAs. **d**, Metabolite labelling from [1-^{13}C]glutamine from cells in **c**. **e**, IDH flux estimates from ^{13}C metabolic flux analysis model in control or IDH1-knockdown A549 cells cultured with [U-^{13}C$_5$]glutamine. **f**, Cell proliferation of A549 cells expressing control or IDH1-shRNAs. Error bars, 95% confidence intervals (**b**, **e**) and s.e.m. ($n = 3$) (**d**, **f**). *$P < 0.05$.

using several cell lines (Supplementary Fig. 8). Finally, we used ^{13}C metabolic flux analysis[17,18] to quantify intracellular fluxes using [U-^{13}C$_5$]glutamine in A549 cells. The fitted data suggested that reductive IDH flux significantly decreased when IDH1 protein levels were decreased, and this change was the primary alteration observed in the network (Fig. 1e; see Supplementary Tables 1–5 for complete results and metabolic flux analysis model description).

Our results suggest that IDH1 can convert NADPH, αKG and CO_2 to isocitrate and $NADP^+$ in the cytosol. Enzymatic analysis using recombinant protein indicated that IDH1 is indeed capable of consuming NADPH and is responsive to physiological levels of CO_2 (Supplementary Fig. 9). Importantly, the proliferation rate of all cell lines with IDH1 knockdown was impaired (Fig. 1f and Supplementary Fig. 8), indicating that reductive metabolism of αKG in the cytosol may be necessary for robust growth. In contrast to our results with IDH1 shRNAs, we detected no significant change in reductive flux when targeting IDH2 mRNA in A549, MDA-MB-231 and HCT116 cells (Supplementary Fig. 10). Although IDH2 may promote reductive carboxylation in some tissues or conditions, these results are consistent with the proposed role of IDH2 as an oxidative TCA cycle enzyme[19].

Intriguingly, we detected a significant increase in reductive carboxylation activity when culturing cells under hypoxia (Supplementary Fig. 11). As glucose is usually the primary carbon source for mammalian tissues[1–3], we next compared the contributions of reductive glutamine metabolism and glucose oxidation to fatty-acid synthesis by culturing cells with either [5-^{13}C]glutamine or uniformly labelled [U-^{13}C$_6$]glucose under normal tissue culture conditions or hypoxia. Strikingly, oxygen levels influenced fatty-acid labelling from both tracers (Fig. 2a, b). Cells preferentially used glucose carbon for palmitate synthesis under normoxic conditions; however, fatty acids produced under hypoxia were primarily synthesized from glutamine carbon through the reductive pathway. In fact, the reductive carboxylation of glutamine-derived αKG accounted for approximately 80% of the carbon used for *de novo* lipogenesis in A549 cells growing under hypoxia (Fig. 2c). Conversely, we detected a concomitant decrease in the contribution of [U-^{13}C$_6$]glucose to fatty-acid synthesis under this condition. Significant increases in the relative use of this pathway were observed in all cell lines tested, including non-transformed cells (Supplementary Fig. 12). In addition, T lymphocytes freshly isolated from a mouse spleen preferentially used reductive glutamine metabolism over glucose oxidation for fatty-acid synthesis when activated under hypoxia (Supplementary Fig. 13). Although proliferation rates and relative *de novo* lipogenesis were lower under hypoxia (Supplementary Fig. 14a, b), the net flux of reductive glutamine metabolism to palmitate synthesis was significantly increased in hypoxic cultures (Fig. 2d). Knockdown of IDH1 protein mitigated the use of reductive glutamine metabolism for lipogenesis under hypoxia (Supplementary Fig. 14c). Although most human cells require glutamine for nucleotide and hexosamine biosynthesis, some cell lines can grow in the absence of exogenous sources of glutamine[20]. Remarkably, we found that hypoxia increases the dependence of such cells on glutamine, as evidenced by decreased proliferation in the absence of glutamine and increased reductive glutamine metabolism under hypoxia when glutamine is present (Fig. 2e and Supplementary Fig. 15).

To gain insight into the mechanisms controlling this switch to reductive glutamine metabolism, we analysed changes in the labelling

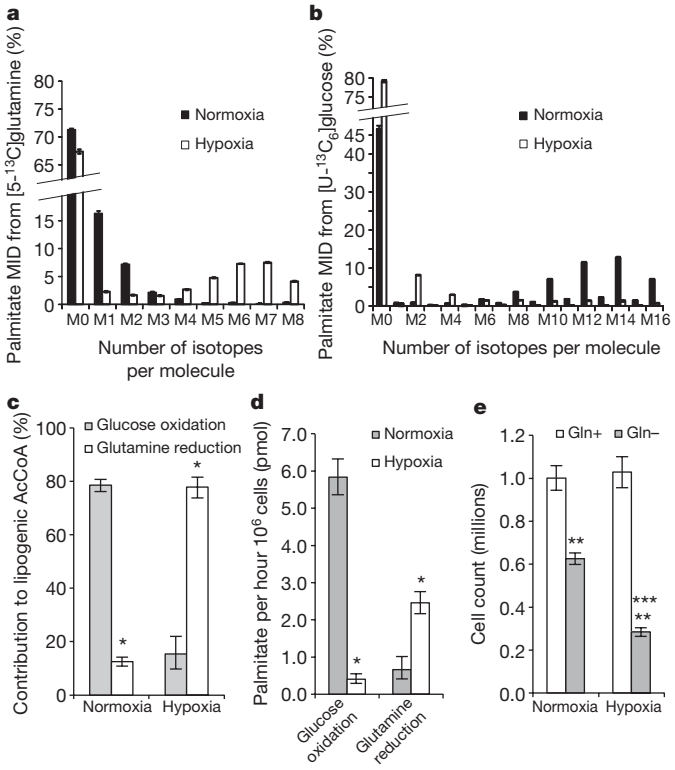


Figure 2 | Hypoxia reprograms cells to rely on reductive glutamine metabolism for lipid synthesis. **a**, **b**, Labelling of palmitate extracts from A549 cells cultured under normoxia or hypoxia with [5-^{13}C]glutamine (**a**) or [U-$^{13}C_6$]glucose (**b**). Similar results were observed in myristate, oleate and stearate pools (not shown). **c**, Relative contribution of glucose oxidation ([U-$^{13}C_6$]glucose) or glutamine reduction ([5-^{13}C]glutamine) to lipogenic AcCoA in A549 cells under normoxia and hypoxia. **d**, Absolute fluxes of [U-$^{13}C_6$]glucose and [5-^{13}C]glutamine to palmitate in A549 cells. Error bars, 95% confidence intervals from model (**c**, **d**); $*P < 0.05$. **e**, Huh7 cell proliferation after 4 days in the presence or absence of glutamine. Error bars, s.e.m. ($n = 3$) (**a**, **b**, **e**). $**P < 0.005$ comparing glutamine-free cultures. $***P < 0.001$ comparing normoxia and hypoxia.

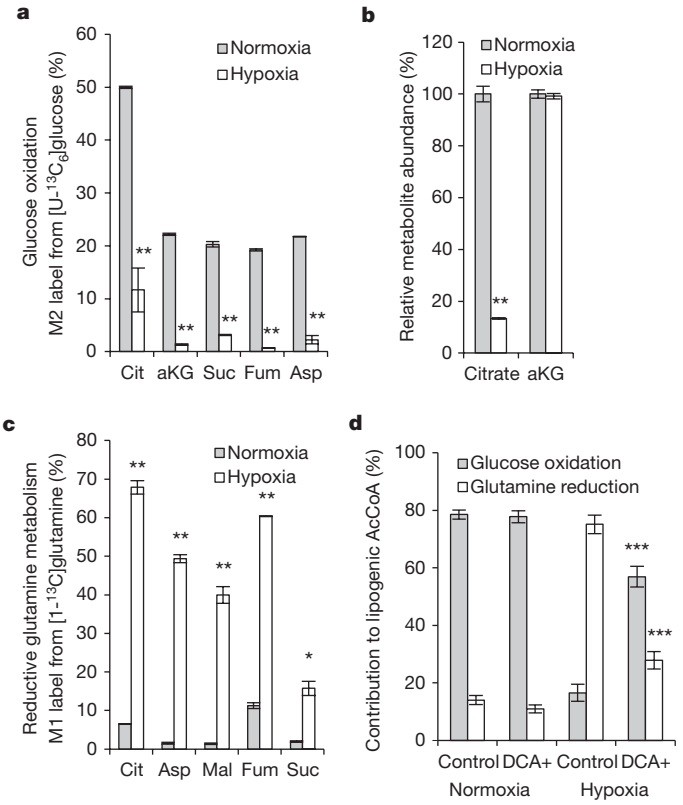


Figure 3 | Reductive TCA metabolism increases under hypoxia. MRC5 cells were cultured under normoxia or hypoxia for 3 days in the presence of tracer. **a**, Relative level of glucose oxidation as determined by M2 labelling from [U-$^{13}C_6$]glucose (see Supplementary Fig. 16a for atom transition map). M2 isotopologues were the most abundant labelled metabolites in mass spectra. **b**, Relative abundance of citrate and αKG. **c**, Relative contribution of reductive glutamine metabolism to TCA metabolites, determined by M1 labelling from [1-^{13}C]glutamine. **d**, Contribution of glucose oxidation and glutamine reduction to lipogenesis in A549 cells cultured with or without 5 mM DCA. Error bars, s.e.m. (**a**–**c**) ($n = 3$) and 95% confidence intervals (**d**). $*P < 0.01$ and $**P < 0.001$ comparing normoxia to hypoxia. $***P < 0.05$ comparing control with DCA in hypoxia.

and abundances of TCA cycle metabolites in cells grown under hypoxia. Using [U-$^{13}C_6$]glucose, we observed a significant decrease in relative flux through the pyruvate dehydrogenase (PDH) complex (Fig. 3a and Supplementary Fig. 16a, b). In addition, the citrate pool became depleted, which would be expected to increase reductive carboxylation flux through mass action (Fig. 3b and Supplementary Fig. 16c). On the other hand, we detected increased amounts of isotopic label in TCA cycle intermediates when using labelled glutamine tracers under hypoxia (Fig. 3c and Supplementary Fig. 16d, e). Reductively metabolized glutamine accounted for as much as 40–70% of the intracellular citrate, aspartate, malate and fumarate pools when cells were cultured in low oxygen (Fig. 3c). Given the marked reduction in PDH flux observed in hypoxia, we tested the ability of dichloroacetate (DCA) to restore PDH activity and mitigate the contribution of reductive metabolism to lipogenesis. DCA inhibits pyruvate dehydrogenase kinases (PDKs)[21], and PDK1 is a known target of HIF-1α that inhibits the activity of PDH through phosphorylation[22,23]. Although DCA treatment had no observable effect on carbon use under normoxia, reductive glutamine metabolism was inhibited and glucose oxidation was partly restored in A549 cells cultured with DCA under hypoxia (Fig. 3d and Supplementary Fig. 16f), suggesting that hypoxia-induced PDK1 expression contributes to the use of reductive carboxylation for fatty-acid synthesis.

The von Hippel–Lindau (VHL) tumour suppressor protein is frequently lost in renal cell carcinoma (RCC) and results in a state of 'pseudohypoxia' by activating HIF signalling[24–26]. To understand further the role of this pathway in promoting the switch to reductive TCA metabolism we tested RCC cells deficient in VHL using ISA. Remarkably, VHL-deficient RCC cell lines preferentially used reductive glutamine metabolism for lipogenesis, even when cultured under normal oxygen levels, whereas those expressing wild-type (WT) VHL behaved similarly to other carcinoma cell lines (Fig. 4a and Supplementary Fig. 17). Re-expression of WT VHL in previously VHL-deficient cell lines resulted in a switch back to oxidative glucose metabolism as the source of carbon for lipid synthesis (Fig. 4b), reduced extracellular fluxes of glucose, lactate and glutamine (Supplementary Fig. 18a), and increased the pool of intracellular citrate relative to αKG (Supplementary Fig. 18b) under normoxia. Furthermore, shRNA-mediated knockdown of HIF-2α partly restored glucose-mediated lipogenesis in 786-O cells (Fig. 4c, d). Consistent with VHL and HIF influencing the switch to reductive glutamine metabolism during hypoxia, glucose entry into the TCA cycle by PDH was increased under normoxia upon introduction of WT VHL or knockdown of HIF-2α in 786-O cells (Supplementary Fig. 18c). Similar changes were observed after knockdown of the HIF-α dimerization partner ARNT (aryl hydrocarbon receptor nuclear translocator) in VHL-deficient normoxic UMRC2 cells, which express both HIF-1α and HIF-2α (ref. 27), or after ARNT knockdown in hypoxic A549 and 143B cells (Supplementary Fig. 19).

Our results highlight an important role for reductive TCA metabolism of glutamine in cell proliferation at physiological oxygen levels

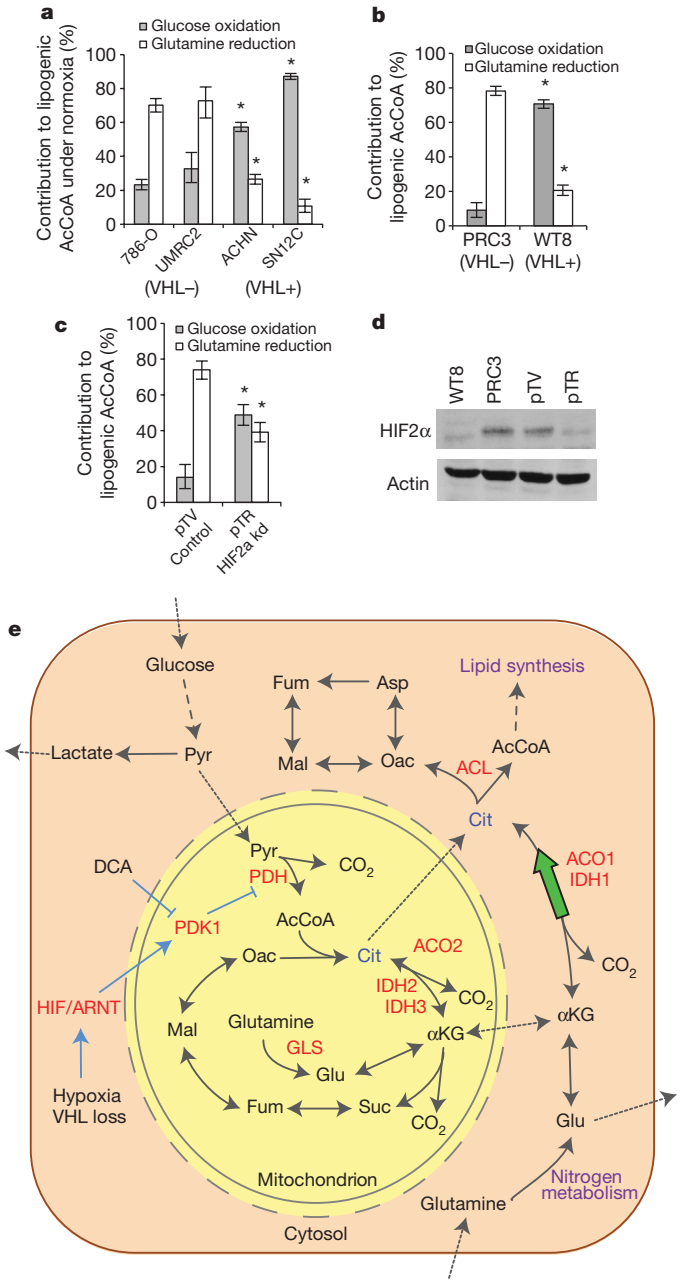


Figure 4 | HIF/ARNT/VHL signalling regulate carbon use for lipogenesis. **a**–**c**, Contribution of glucose oxidation ([U-$^{13}C_6$]glucose) and glutamine reduction ([5-^{13}C]glutamine) to lipogenesis in RCC lines (**a**), parental control (PRC3) and VHL+ (WT8) cells derived from 786-O line (**b**) or vector control (pTV) or HIF2α shRNA (pTR) cells derived from 786-O line (**c**). **d**, Western blot to determine HIF-2α levels for cells in **b**, **c**. Error bars, 95% confidence intervals obtained from ISA model. *$P < 0.05$. **e**, Model depicting the metabolic reprograming of mammalian cells by hypoxia or VHL loss to use reductive glutamine metabolism for lipogenesis. HIF stabilization drives transcription of PDK1, which decreases PDH activity and subsequently intracellular citrate levels. IDH1 and ACO1 reductively generate lipogenic citrate from glutamine-derived αKG. DCA can inhibit PDKs, forcing increased glucose oxidation in hypoxic cells.

(Fig. 4e). Given the almost exclusive use of reductive carboxylation for lipogenesis under hypoxia, a redundant or contributing role of mitochondrial IDH2 in this pathway is probable; however, our data provide evidence that the reductive pathway involves IDH1-mediated catalysis in the cytoplasm. Although the carbon source that cells use for lipid synthesis appears to be determined, at least partly, by HIF-mediated regulation of PDK1, additional hypoxia-associated changes may also promote reductive glutamine metabolism. For example, HIF-2α enhances c-MYC activity[28], which in turn drives glutamine catabolism through the regulation of numerous genes including glutaminase[29]. This metabolic reprogramming provides an effective, glucose-independent means of generating AcCoA for biosynthesis. Because glucose is also delivered to cells through the vasculature, it may be limited in microenvironments with decreased oxygen availability[30]. Reductively metabolizing amino acids for lipid synthesis under these conditions would allow cells to conserve glucose for production of ribose and other biosynthetic precursors (for example, one carbon pool, hexosamines) that are not typically generated from other nutrients. Thus, reductive metabolism may allow cells to distribute available nutrients more efficiently in poorly vascularized microenvironments. These results add a new dimension to our understanding of cell metabolism and suggest potential therapeutic targets along the reductive carboxylation and glutamine catabolic pathways that could mitigate hypoxic tumour growth.

METHODS SUMMARY

For determination of steady-state labelling of polar metabolites, cells were cultured for approximately 24 h in the presence of ^{13}C-labelled glutamine or glucose before extraction. For experiments involving stable isotopic labelling of lipid biomass, cells were grown for approximately 3–5 days in the presence of tracer before extraction. Details of the extraction and derivatization methods are described in the Supplementary Methods. Computational determination of metabolic fluxes, confidence intervals, *de novo* lipogenesis and the contribution of tracers to fatty-acid carbon was accomplished using an in-house software package, Metran[18]. Details of the metabolic networks and gas chromatography/mass spectrometry (GC/MS) measurements used for modelling and complete results are described in Supplementary Information. The generation of cells stably expressing control shRNAs or those targeted IDH1 or IDH2 is described in Supplementary Methods; all experiments were conducted within four passages of initial selection. Hypoxic microenvironments were generated by feeding incubators with a pre-mixed gas composed of 1% O_2, 5% CO_2 and 94% N_2, and O_2 levels were confirmed to range between 1 and 3% using a Fyrite combustion analyser. For details of recombinant IDH1 production and enzyme assays, T-cell activation, medium analysis, [5-^{14}C]glutamine experiment and western blotting, see Supplementary Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 March; accepted 28 September 2011.
Published online 20 November 2011.

1. Hatzivassiliou, G. *et al.* ATP citrate lyase inhibition can suppress tumor cell growth. *Cancer Cell* **8,** 311–321 (2005).
2. Wellen, K. E. *et al.* ATP-citrate lyase links cellular metabolism to histone acetylation. *Science* **324,** 1076–1080 (2009).
3. Migita, T. *et al.* ATP citrate lyase: activation and therapeutic implications in non-small cell lung cancer. *Cancer Res.* **68,** 8547–8554 (2008).
4. Semenza, G. L. HIF-1: upstream and downstream of cancer metabolism. *Curr. Opin. Genet. Dev.* **20,** 51–56 (2010).
5. DeBerardinis, R. J. *et al.* Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc. Natl Acad. Sci. USA* **104,** 19345–19350 (2007).
6. Weinberg, F. *et al.* Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl Acad. Sci. USA* **107,** 8788–8793 (2010).
7. Des Rosiers, C. *et al.* Isotopomer analysis of citric acid cycle and gluconeogenesis in rat liver. Reversibility of isocitrate dehydrogenase and involvement of ATP-citrate lyase in gluconeogenesis. *J. Biol. Chem.* **270,** 10027–10036 (1995).
8. Yoo, H., Antoniewicz, M. R., Stephanopoulos, G. & Kelleher, J. K. Quantifying reductive carboxylation flux of glutamine to lipid in a brown adipocyte cell line. *J. Biol. Chem.* **283,** 20621–20627 (2008).
9. Metallo, C. M., Walther, J. L. & Stephanopoulos, G. Evaluation of ^{13}C isotopic tracers for metabolic flux analysis in mammalian cells. *J. Biotechnol.* **144,** 167–174 (2009).
10. Lemons, J. M. *et al.* Quiescent fibroblasts exhibit high metabolic activity. *PLoS Biol.* **8,** e1000514 (2010).
11. Ward, P. S. *et al.* The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting α-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17,** 225–234 (2010).
12. Boros, L. G. *et al.* Defective RNA ribose synthesis in fibroblasts from patients with thiamine-responsive megaloblastic anemia (TRMA). *Blood* **102,** 3556–3561 (2003).

13. Munger, J. *et al.* Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. *Nature Biotechnol.* **26,** 1179–1186 (2008).
14. Maier, K., Hofmann, U., Reuss, M. & Mauch, K. Identification of metabolic fluxes in hepatic cells from transient ^{13}C-labeling experiments: part II. Flux estimation. *Biotechnol. Bioeng.* **100,** 355–370 (2008).
15. Kharroubi, A. T., Masterson, T. M., Aldaghlas, T. A., Kennedy, K. A. & Kelleher, J. K. Isotopomer spectral analysis of triglyceride fatty acid synthesis in 3T3–L1 cells. *Am. J. Physiol.* **263,** E667–E675 (1992).
16. Siebert, G., Carsiotis, M. & Plaut, G. W. The enzymatic properties of isocitric dehydrogenase. *J. Biol. Chem.* **226,** 977–991 (1957).
17. Sauer, U. Metabolic networks in motion: ^{13}C-based flux analysis. *Mol. Syst. Biol.* **2,** 62 (2006).
18. Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H. & Stephanopoulos, G. An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol. Bioeng.* **99,** 686–699 (2008).
19. Hartong, D. T. *et al.* Insights from retinitis pigmentosa into the roles of isocitrate dehydrogenases in the Krebs cycle. *Nature Genet.* **40,** 1230–1234 (2008).
20. Cheng, T. *et al.* Pyruvate carboxylase is required for glutamine-independent growth of tumor cells. *Proc. Natl Acad. Sci. USA* **108,** 8674–8679 (2011).
21. Bonnet, S. *et al.* A mitochondria-K^+ channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth. *Cancer Cell* **11,** 37–51 (2007).
22. Kim, J. W., Tchernyshyov, I., Semenza, G. L. & Dang, C. V. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.* **3,** 177–185 (2006).
23. Papandreou, I., Cairns, R. A., Fontana, L., Lim, A. L. & Denko, N. C. HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metab.* **3,** 187–197 (2006).
24. Kaelin, W. G. Jr. Molecular basis of the VHL hereditary cancer syndrome. *Nature Rev. Cancer* **2,** 673–682 (2002).
25. Ivan, M. *et al.* HIFα targeted for VHL-mediated destruction by proline hydroxylation: implications for O_2 sensing. *Science* **292,** 464–468 (2001).
26. Jaakkola, P. *et al.* Targeting of HIF-α to the von Hippel-Lindau ubiquitylation complex by O_2-regulated prolyl hydroxylation. *Science* **292,** 468–472 (2001).
27. Zimmer, M. *et al.* Small-molecule inhibitors of HIF-2a translation link its 5′UTR iron-responsive element to oxygen sensing. *Mol. Cell* **32,** 838–848 (2008).
28. Gordan, J. D., Bertout, J. A., Hu, C. J., Diehl, J. A. & Simon, M. C. HIF-2α promotes hypoxic cell proliferation by enhancing c-myc transcriptional activity. *Cancer Cell* **11,** 335–347 (2007).
29. Gao, P. *et al.* c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* **458,** 762–765 (2009).
30. Sutherland, R. M. Cell and environment interactions in tumor microregions: the multicell spheroid model. *Science* **240,** 177–184 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Vokes and P. Ward for discussions. We also thank S. Gross and Agios Pharmaceuticals for providing the IDH1 construct. We acknowledge support from National Institutes of Health grant R01 DK075850-01. C.M.M. is supported by a postdoctoral fellowship from the American Cancer Society. K.H. is supported by the German Research Foundation (DFG) grant HI1400. L.G. is supported by the NIH and the Glenn Foundation for Medical Research. M.G.V.H. is supported by the Burrough's Wellcome Fund, the Smith Family, the Damon Runyon Cancer Research Foundation and the National Cancer Institute. D.J.I. is an investigator of the Howard Hughes Medical Institute. O.I. is supported by R01 CA122591 and the Dana Farber/Harvard Cancer Center Kidney SPORE Grant Developmental Award.

Author Contributions C.M.M., P.A.G., E.L.B., K.R.M., J.Y., K.H. and C.M.J. performed cellular experiments and isotope tracing. C.M.M. and P.A.G. performed metabolite profiling and analysed data. K.R.M. and M.G.V.H. performed enzyme assays and ^{14}C experiments. E.L.B., J.Y. and Z.R.J. generated western blots. D.J.I. and L.G. provided support and reagents. J.K.K., M.G.V.H., O.I. and G.S. provided conceptual advice. C.M.M., J.K.K., M.G.V.H., O.I. and G.S. wrote and edited the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.S. (gregstep@mit.edu) or O.I. (iliopoul@helix.mgh.harvard.edu).

METHODS

Cell culture, isotopic labelling and hypoxia. All cell lines were cultured in Dulbecco's modified Eagle medium (DMEM; Mediatech) containing 10% fetal bovine serum (FBS; Invitrogen) and 100 U ml^{-1} penicillin/streptomycin (Mediatech) unless otherwise mentioned. Cell lines were obtained from ATCC unless otherwise noted. The VC3 glioma cell line was provided by T. Lautenschlaeger, MDA-MB-231 and HCT116 cell lines were provided by F. Chiaradonna, MRC5 cells were provided by S. Lippard, SN12C, ACHN and 786-O cells were provided by K. Courtney and L. Cantley, MCF10A cells were provided by J. Brugge and Huh7 cells were provided by M. Hemann. MCF10A cells were cultured in custom DMEM/F12 (Hyclone) containing 5% horse serum, 20 ng ml^{-1} EGF, 10 μg ml^{-1} insulin, 100 ng ml^{-1} cholera toxin, 0.5 mg ml^{-1} hydrocortisone, 100 U ml^{-1} penicillin/streptomycin, and labelled/unlabelled glucose and glutamine at 18 mM and 2.5 mM, respectively. PRC3, WT8, pTV and pTR cell lines were subcloned from the 786-O cell line as previously described[31,32]. For isotopic labelling experiments, cells were cultured in six-well plates in glucose- and glutamine-free DMEM (Sigma) containing 10% dialysed FBS (Invitrogen), 100 U ml^{-1} penicillin/streptomycin, naturally labelled glucose or glutamine, and the appropriate tracer, including [U-$^{13}C_5$]glutamine (Isotec), [5-^{13}C]glutamine (C/D/N Isotopes), [1-^{13}C]glutamine, [U-$^{13}C_6$]glucose and [1,2-$^{13}C_2$]glucose (all from Cambridge Isotope Labs). Steady-state labelling of organic and amino acids was accomplished by culturing sub-confluent cells in tracer medium for 24 h (Supplementary Fig. 20). Labelling of fatty acids for ISA was conducted over 3–5 days of culture in an excess of tracer medium (3–4 ml per well in a six-well plate) to prevent nutrient depletion. Hypoxic culture was conducted by feeding a custom mixture of 1% O_2, 5% CO_2 and 94% N_2 to a standard incubator controlled at 5% CO_2. HEPES (Mediatech) was added to the culture medium at 20 mM to maintain pH between normoxic and hypoxic cultures. The internal gas content was monitored using Fyrite gas analysers (Bacharach) for CO_2 and O_2. O_2 levels were confirmed at 1–3% during hypoxic culture.

Isolation and culture of primary CD8$^+$ T lymphocytes. Whole spleens were isolated from OT-1 mice (Jackson Laboratories) expressing ovalbumin (SIINFEKL)-specific T-cell receptors on CD8$^+$ T cells[33]. Spleens were passed through 40-μm cell strainers, and red blood cells were removed by lysis (ACK buffer, Gibco). Splenocytes were magnetically enriched for CD8$^+$ T cells by negative selection (Stem Cell Technologies) then suspended in DMEM containing 10% dialysed serum, 100 U ml^{-1} penicillin/streptomycin and 1× MEM non-essential amino-acid supplement, and labelled/unlabelled glucose (25 mM) and glutamine (4 mM). Cells were cultured for 6 days in 12-well plates under hypoxia and expanded by addition of 10 μg ml^{-1} SIINFEKL peptide. Media was supplemented with 30 U ml^{-1} of recombinant interleukin-2 every 48 h, and cells were maintained at a concentration between 1×10^6 and 2.5×10^6 cells per millilitre by adding additional medium. To verify proliferation, cells were incubated with 2 μM carboxyfluorescein succinimidyl ester (CSFE) in PBS + 2% FBS for 5 min, washed three times in buffer, re-suspended in media and cultured as above. Viability and T-cell enrichment were determined by flow cytometry after staining with 4′,6-diamidino-2-phenylindole (DAPI) and fluorescent anti-CD8 antibody. T-cell proliferation was quantified by flow cytometry using a FACSCanto II equipped with an High Throughput system (Becton Dickinson) and analysed using FlowJo 7.5.5 (TreeStar).

Lentiviral-mediated generation of cells with knockdown of IDH1/2. Stable cell cultures with decreased IDH1 and IDH2 expression were generated by lentiviral-mediated shRNA expression. pLKO.1 lentiviral vectors targeting IDH1 had shRNA sequences of CCGGGCTGCTTGCATTAAAGGTTTACTCGAGTAAACCTTTAATGCAAGCAGCTTTTT (IDH1a; TRCN0000027298) and CCGGCGAATCATTTGGGAATTGATTCTCGAGAATCAATTCCCAAATGATTCGTTTTT (IDH1b; TRCN0000027289), IDH2 had shRNA sequence CCGGGTGGACATCCAGCTAAAGTATCTCGAGATACTTTAGCTGGATGTCCACTTTTT (TRCN0000027225). For controls, either non-targeting control shRNA (SHC002; Sigma) or pLKO.1 scrambled control vector[34] (Addgene) were used. pLKO.1 vector targeting ARNT with shRNA sequence CCGGGAGAAGTCAGATGGTTTATTTCTCGAGAAATAAACCATCTGACTTCTCTTTTT (TRCN0000003819) was obtained from Open Biosystems. HEK293T cells were co-transfected with pLKO.1 vectors and packaging plasmids to produce lentivirus. Filtered supernatants were used for infection, and cells were selected with puromycin (2 μg ml^{-1}) for at least two passages before initiating tracer and flux experiments.

Metabolite extraction and GC/MS analysis. At the conclusion of culture, cells were rinsed with 1 ml ice-cold PBS and quenched with 0.4 ml ice-cold methanol. An equal volume of water was added, and cells were collected in tubes by scraping with a pipette. One volume of ice-cold chloroform was added to each tube, and the extracts were vortexed at 4 °C for 30 min. Samples were centrifuged at 14,000*g* for 5 min at room temperature. For analysis of polar metabolites, the aqueous phase was transferred to a new tube for evaporation under airflow. For ISA experiments, the non-polar fraction was collected and evaporated under airflow. In some ISA experiments, cells were trypsinized, counted and pelleted before lipid extraction as described above.

Dried polar metabolites were dissolved in 20 μl of 2% methoxyamine hydrochloride in pyridine (Pierce) and held at 37 °C for 1.5 h. After dissolution and reaction, tert-butyldimethylsilyl (TBDMS) derivatization was initiated by adding 30 μl *N*-methyl-*N*-(tert-butyldimethylsilyl)trifluoroacetamide (MBTSTFA) + 1% tert-butyldimethylchlorosilane (TBDMCS; Pierce) and incubating at 55 °C for 1 h. Fatty-acid methyl esters were generated by dissolving and reacting dried chloroform fractions in 50–100 μl of Methyl-8 reagent (Pierce) and incubating at 60 °C for 1 h. GC/MS analysis was performed using an Agilent 6890 GC equipped with a 30 m DB-35MS capillary column connected to an Agilent 5975B MS operating under electron impact ionization at 70 eV. One microlitre of sample was injected in splitless mode at 270 °C, using helium as the carrier gas at a flow rate of 1 ml min^{-1}. For measurement of organic and amino acids, the GC oven temperature was held at 100 °C for 3 min and increased to 300 °C at 3.5° min^{-1}. For analysis of fatty-acid methyl esters, the GC temperature was held at 100 °C for 5 min after injection, increased to 200 °C at 15° min^{-1}, then to 250 °C at 5° min^{-1} and finally to 300 °C at 15° min^{-1}. The MS source and quadrupole were held at 230 °C and 150 °C, respectively, and the detector was run in scanning mode, recording ion abundance in the range of 100–605 *m*/*z*. Mass isotopomer distributions were determined by integrating the appropriate ion fragments listed in Supplementary Table 1. When required, mass isotopomer distributions were corrected for natural isotope abundance using in-house algorithms adapted from ref. 35. Relative metabolite abundances were measured using a norvaline internal standard and integrating all potentially labelled ions in the metabolite fragments listed in Supplementary Table 1.

ISA and metabolic flux analysis. Computational estimation of fluxes or percentage enrichment in the AcCoA pool and their associated 95% confidence intervals were accomplished using the elementary metabolite unit (EMU)-based software Metran, executed within Matlab (Mathworks) as previously described[36–40]. Briefly, fluxes were determined iteratively by simulating MS measurements from a given flux vector and comparing with mass isotopomer distribution measurements (three biological replicates). Upon obtaining an acceptable fit, confidence intervals were determined for each flux using parameter continuation. Assumptions, network (including atom transitions), raw data and model fits are presented in Supplementary Information. ISA was performed in a similar manner using the simple network described in Supplementary Fig. 4 and Supplementary Table 6 to determine the tracer enrichment in lipogenic AcCoA (*D* value) and percentage of newly synthesized lipids and *de novo* lipogenesis, $g(t)$ (ref. 15). Uncorrected mass isotopomer distributions, fitted parameters and confidence intervals used for ISA are listed in Supplementary Information.

Calculation of absolute flux of tracers to palmitate in biomass. The quantity of newly synthesized palmitate was determined by multiplying the fractional newly synthesized palmitate value ($g(t)$ value from ISA) by the total cellular palmitate. Total cellular palmitate was quantified by GC/MS using a triheptadecanoin internal standard. Flux of a given tracer to palmitate was calculated by multiplying the tracer contribution (*D* value from ISA) by the amount of newly synthesized palmitate and dividing by the integral viable cell density over the course of the experiment.

Detection of ^{14}C incorporation into lipids. One hundred and fifty thousand A549 cells were plated in DMEM. After attachment, [5-^{14}C]glutamine was added to the medium at a final concentration of 1 μCi ml^{-1}. After 72 h of growth, lipids were extracted from plates using two rinses of 500 μl hexane:isopropanol (3:2). Three hundred microlitres of PBS was added to induce phase separation and collect the non-polar phase, and an additional 300 μl was added to the hexane fraction to rinse the non-polar fraction. The hexane fraction was dried under nitrogen gas, and the residue was dissolved in 200 μl chloroform before quantifying c.p.m. by liquid scintillation counting. Matched plates of cells were counted to determine cell number.

Metabolite analysis of spent medium. Glucose, lactate, glutamine and glutamate concentrations were measured in fresh and spent medium samples using a Yellow Springs Instruments 7100. Cell number was determined using a hemocytometer. Extracellular flux measurements were calculated by assuming exponential growth over the culture period to determine integral viable cell density.

Purification of recombinant IDH1. His-tagged IDH1 in pET41a was transformed into *Escherichia coli* (BL21 plysS DE3) and cells were grown with kanamycin selection to an $D_{600\,nm}$ of 0.6. The cells were then moved to 18 °C and induced with 1 mM IPTG for 16–18 h, pelleted and subjected to freeze/thaw before re-suspension in 60 ml lysis buffer (20 mM Tris, pH 7.4, 0.1% Triton

X-100, 500 mM NaCl, 5 mM β-mercaptoethanol, 10% glycerol, supplemented with protease inhibitors). Cells were lysed by sonication and protein bound to Ni-NTA agarose. The beads were batch washed three or four times with wash buffer (20 mM Tris, pH 7.4, 500 mM NaCl, 5 mM β-mercaptoethanol, 10% glycerol), then eluted from a column in 1-ml fractions with elution buffer (20 mM Tris, pH 7.4, 500 mM NaCl, 5 mM β-mercaptoethanol, 500 mM imidazole, 10% glycerol). The first and second fractions containing most of the protein were dialysed into 50 mM Tris pH 7.5, 200 mM NaCl, 5 mM β-mercaptoethanol, 2 mM $MnCl_2$, 10% glycerol and the recombinant enzyme was stored at −80 °C.

Recombinant IDH1 enzyme assays. All reactions were performed in reaction buffer (100 mM Tris pH 7.5, 1.3 mM $MnCl_2$, 200 μM NADPH and 2 mM α-ketoglutarate), which was equilibrated overnight at 0%, 5% or 10% CO_2 as indicated. One hundred microlitres of reaction buffer for each CO_2 condition was added to 10 μg rIDH1 and activity was measured by following NADPH fluorescence (excitation at 340 nm, emission at 460 nm).

SDS–PAGE and western blotting. Cells were rinsed with ice-cold PBS and lysed using RIPA buffer. Proteins were separated by SDS–PAGE and transferred to a nitrocellulose membrane. After blocking, membranes were probed with goat anti-IDH1 polyclonal antibody (Santa Cruz Biotechnology, sc49996), mouse anti-IDH2 monoclonal antibody (Abcam, ab55271), rabbit anti-HIF2α polyclonal antibody (Novus Biologicals, NB100-122), mouse anti-ARNT1 monoclonal antibody (BD Biosciences, 611079), mouse anti-β-actin mouse monoclonal antibody (Novus Biologicals, ab8226), mouse anti-tubulin monoclonal antibody (Sigma) or rabbit anti-tubulin antibody (Sigma). Protein was detected using horseradish-peroxidase-conjugated secondary antibodies and chemiluminescence.

31. Zimmer, M., Doucette, D., Siddiqui, N. & Iliopoulos, O. Inhibition of hypoxia-inducible factor is sufficient for growth suppression of $VHL^{-/-}$ tumors. *Mol. Cancer Res.* **2,** 89–95 (2004).
32. Iliopoulos, O., Kibel, A., Gray, S. & Kaelin, W. G. Jr. Tumour suppression by the human von Hippel-Lindau gene product. *Nature Med.* **1,** 822–826 (1995).
33. Hogquist, K. A. *et al.* T cell receptor antagonist peptides induce positive selection. *Cell* **76,** 17–27 (1994).
34. Sarbassov, D. D., Guertin, D. A., Ali, S. M. & Sabatini, D. M. Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex. *Science* **307,** 1098–1101 (2005).
35. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of ^{13}C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom.* **31,** 255–262 (1996).
36. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metab. Eng.* **8,** 324–337 (2006).
37. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab. Eng.* **9,** 68–86 (2007).
38. Noguchi, Y. *et al.* Effect of anaplerotic fluxes and amino acid availability on hepatic lipoapoptosis. *J. Biol. Chem.* **284,** 33425–33436 (2009).
39. Gaglio, D. *et al.* Oncogenic K-Ras decouples glucose and glutamine metabolism to support cancer cell growth. *Mol. Syst. Biol.* **7,** 523 (2011).
40. Grassian, A. R., Metallo, C. M., Coloff, J. L., Stephanopoulos, G. & Brugge, J. S. Erk regulation of pyruvate dehydrogenase flux through PDK4 modulates cell proliferation. *Genes Dev.* **25,** 1716–1733 (2011).

LETTER

doi:10.1038/nature10642

Reductive carboxylation supports growth in tumour cells with defective mitochondria

Andrew R. Mullen[1], William W. Wheaton[2,3], Eunsook S. Jin[4,5], Pei-Hsuan Chen[1], Lucas B. Sullivan[2,3], Tzuling Cheng[1], Youfeng Yang[6], W. Marston Linehan[6], Navdeep S. Chandel[2,3] & Ralph J. DeBerardinis[1,7,8]

Mitochondrial metabolism provides precursors to build macromolecules in growing cancer cells[1,2]. In normally functioning tumour cell mitochondria, oxidative metabolism of glucose- and glutamine-derived carbon produces citrate and acetyl-coenzyme A for lipid synthesis, which is required for tumorigenesis[3]. Yet some tumours harbour mutations in the citric acid cycle (CAC) or electron transport chain (ETC) that disable normal oxidative mitochondrial function[4–7], and it is unknown how cells from such tumours generate precursors for macromolecular synthesis. Here we show that tumour cells with defective mitochondria use glutamine-dependent reductive carboxylation rather than oxidative metabolism as the major pathway of citrate formation. This pathway uses mitochondrial and cytosolic isoforms of $NADP^+$/NADPH-dependent isocitrate dehydrogenase, and subsequent metabolism of glutamine-derived citrate provides both the acetyl-coenzyme A for lipid synthesis and the four-carbon intermediates needed to produce the remaining CAC metabolites and related macromolecular precursors. This reductive, glutamine-dependent pathway is the dominant mode of metabolism in rapidly growing malignant cells containing mutations in complex I or complex III of the ETC, in patient-derived renal carcinoma cells with mutations in fumarate hydratase, and in cells with normal mitochondria subjected to acute pharmacological ETC inhibition. Our findings reveal the novel induction of a versatile glutamine-dependent pathway that reverses many of the reactions of the canonical CAC, supports tumour cell growth, and explains how cells generate pools of CAC intermediates in the face of impaired mitochondrial metabolism.

We first studied the metabolism of isogenic 143B human osteosarcoma cells that contained or lacked a loss-of-function mutation in ETC complex III (cytochrome *b*–c1 complex). These cell lines were generated by depleting 143B mitochondrial DNA (mtDNA) and repopulating with either wild-type mtDNA or mtDNA containing a frameshift mutation in the gene encoding cytochrome *b* (*CYTB*: also known as *MT-CYB*), an essential complex III component[8]. Despite lack of respiration and complex III function in the mutants[8], both wild-type (WT 143B) and *CYTB*-mutant (*CYTB* 143B) cells form colonies in soft agar[9] and proliferate at comparable rates (Supplementary Fig. 1a), making these cells a good model in which to study growth during mitochondrial dysfunction. Both cell lines had detectable CAC intermediates, although citrate was less abundant and succinate was significantly more abundant in the *CYTB* 143B cells (Supplementary Fig. 1b). As expected for cells with defective oxidative phosphorylation, *CYTB* 143B cells had higher glucose consumption and lactate production than WT 143B cells, indicating a metabolic shift towards aerobic glycolysis (Fig. 1a). To determine the effects of *CYTB* mutation on the metabolic fates of glucose, we cultured both cell lines in medium containing D[U-^{13}C]glucose (here U indicates uniformly labelled) and measured ^{13}C enrichment of intracellular metabolites by mass spectrometry. In WT 143B cells, most citrate molecules contained glucose-derived ^{13}C (Fig. 1b). Citrate m+2 (citrate containing two additional mass units from ^{13}C) results from oxidative decarboxylation of glucose-derived pyruvate by pyruvate dehydrogenase (PDH) to form [1,2-^{13}C]acetyl-coA (coA, coenzyme A), followed by condensation with an unlabelled

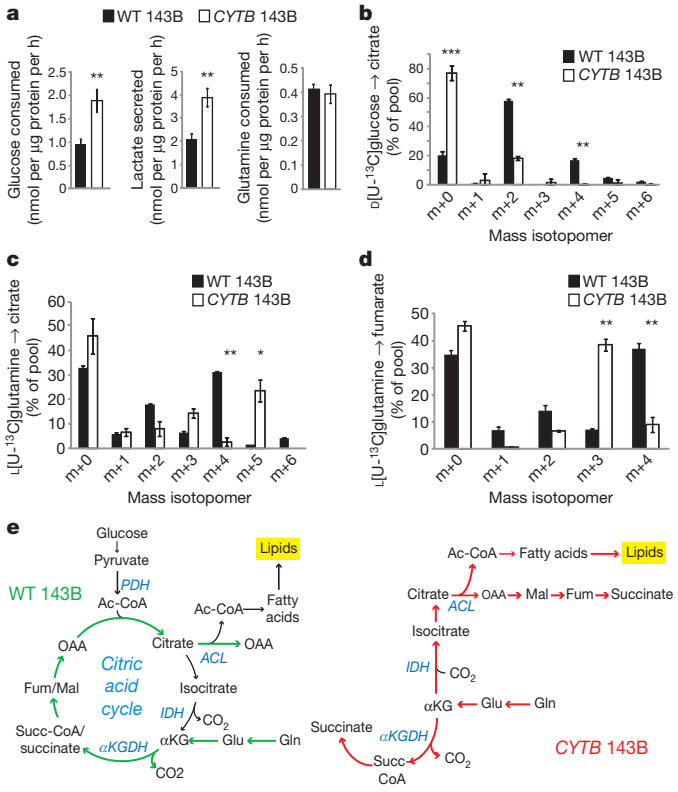


Figure 1 | A reductive pathway of glutamine metabolism in cancer cells lacking activity of ETC complex III. **a**, Glucose utilization, lactate secretion and glutamine utilization in WT 143B and *CYTB* 143B cells. **b**, Mass isotopomer analysis of citrate in cells cultured with D[U-^{13}C]glucose and unlabelled glutamine. **c**, **d**, Mass isotopomer analysis of citrate and fumarate in cells cultured with L[U-^{13}C]glutamine and unlabelled glucose. Data are the average ± s.d. for three independent cultures. $*P < 0.05$; $**P < 0.005$, Student's *t*-test. **e**, Schematic of glutamine metabolism in WT 143B and *CYTB* 143B cells. Coloured arrows follow the paths of glutamine-derived carbon. Abbreviations: Ac-CoA, acetyl-CoA; OAA, oxaloacetate; Gln, glutamine; Glu, glutamate; αKG, α-ketoglutarate; Succ-CoA, succinyl-CoA; Fum, fumarate; Mal, malate; PDH, pyruvate dehydrogenase; ACL, ATP-citrate lyase; IDH, isocitrate dehydrogenase; αKGDH, α-ketoglutarate dehydrogenase.

[1]Department of Pediatrics, University of Texas – Southwestern Medical Center at Dallas, Dallas, Texas 75390-9063, USA. [2]Department of Medicine, Northwestern University, Chicago, Illinois 60611-3008, USA. [3]Department of Cell and Molecular Biology, Northwestern University, Chicago, Illinois 60611-3008, USA. [4]Department of Internal Medicine, University of Texas – Southwestern Medical Center at Dallas, Dallas, Texas 75390-9063, USA. [5]Advanced Imaging Research Center, University of Texas – Southwestern Medical Center at Dallas, Dallas, Texas 75390-8568, USA. [6]Urologic Oncology Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. [7]McDermott Center for Human Growth and Development, University of Texas – Southwestern Medical Center at Dallas, Dallas, Texas 75390-8591, USA. [8]Harold C. Simmons Comprehensive Cancer Center, University of Texas – Southwestern Medical Center at Dallas, Dallas, Texas 75235-5303, USA.

oxaloacetate (OAA). Processing of citrate m+2 around one turn of the CAC produces citrate m+4 (Supplementary Fig. 2). In *CYTB* 143B cells, most citrate contained no glucose carbon (m+0), indicating suppressed PDH contribution to acetyl-CoA (Fig. 1b). Fumarate and malate m+2 were also decreased (Supplementary Fig. 1c, d).

Glutamine is a major respiratory substrate in cancer cells, providing energy and carbon for growth[3,10]. Such cells use glutamine to supply anaplerosis, the replenishment of CAC intermediates withdrawn from the cycle to feed biosynthetic pathways. Both WT and *CYTB* 143B cells require glutamine for colony formation, implying a respiration-independent function for glutamine in cell growth[9], and both cell types consumed glutamine at similar rates (Fig. 1a). We cultured both cell lines with L[U-^{13}C]glutamine to define the metabolic fates of glutamine. Like other glutamine-dependent cancer cells[11], WT 143B cells used glutamine as the major anaplerotic precursor, resulting in a large amount of fumarate, malate and citrate m+4 (Fig. 1c, d; Supplementary Figs 1e and 3, top). In contrast, *CYTB* 143B cells produced only trace quantities of citrate m+4. Instead, they produced citrate m+5 through reductive carboxylation of glutamine-derived α-ketoglutarate (Fig. 1c). This reaction involves addition of an unlabelled carbon by isocitrate dehydrogenase (IDH) acting in reverse relative to the canonical oxidative CAC[12]. Cleavage of citrate m+5 by ATP-citrate lyase then produces acetyl-CoA m+2 and OAA m+3, with the unlabelled carbon retained on OAA (Supplementary Fig. 3, bottom). Examination of other CAC metabolites in *CYTB* 143B cells revealed that they were formed downstream of citrate m+5 and OAA m+3. These cells contained abundant malate and fumarate m+3, which was essentially absent from WT 143B cells (Fig. 1d and Supplementary Fig. 1e). Thus these metabolites are generated from reductive, rather than oxidative, glutamine metabolism (Fig. 1e). Succinate was formed through both reductive and oxidative glutamine metabolism in *CYTB* 143B cells (Supplementary Fig. 1f). Glutamine-dependent reductive carboxylation was also observed in *CYTB* 143B cells cultured in medium containing a full complement of amino acids (Supplementary Fig. 4), and in cells with a genetic defect in ETC complex I (Supplementary Fig. 5).

To determine which of the three IDH isoforms contributed to reductive carboxylation, we used RNA interference (RNAi) to silence expression of *IDH1* (isocitrate dehydrogenase 1 (NADP$^+$), soluble), *IDH2* and *IDH3* in *CYTB* 143B cells and measured glutamine-derived ^{13}C labelling in citrate. Silencing either of the NADP$^+$/NADPH-dependent IDH isoforms, *IDH1* or *IDH2*, reduced citrate m+5 in cells cultured with L[U-^{13}C]glutamine and silencing both at once further reduced citrate m+5 (Fig. 2a, b). Conversion of glutamine to glutamate was unaffected (Fig. 2b). Silencing *IDH3*, the mitochondrial NAD$^+$/NADH-dependent isoform, had no effect on labelling (Fig. 2a, b). Thus reductive carboxylation in *CYTB* 143B cells involves *IDH1* and *IDH2*, but not *IDH3*. Because IDH1 and IDH2 proteins are localized primarily in the cytoplasm and mitochondria, respectively, of both cell lines (Supplementary Fig. 6), the data suggest that reductive carboxylation in *CYTB* 143B cells may occur in either compartment.

To determine whether IDH1 and IDH2 are required for growth, we stably silenced each enzyme using lentiviral RNAi. In WT 143B cells, silencing either isoform reduced cell growth (Supplementary Fig. 7), consistent with findings in other cancer cells with normal mitochondria[13]. In *CYTB* 143B cells, a similar growth reduction occurred when either isoform was silenced (Supplementary Fig. 7). Thus both IDH1 and IDH2 support growth in cells that use glutamine-dependent reductive carboxylation.

We next determined whether reductive glutamine metabolism contributed to lipogenesis, a biosynthetic activity required for cancer cell growth. Both WT and *CYTB* 143B cells produced fatty acids *de novo*, as indicated by incorporation of ^3H_2O into lipids (Fig. 3a). The two cell lines were then cultured with ^{14}C-labelled lipogenic substrates to examine carbon preferences for lipogenesis (Fig. 3a). Both cell lines had an intact pathway of lipogenesis from exogenously supplied acetate. The enhanced labelling in *CYTB* 143B cells probably resulted from an increased specific activity of the intracellular two-carbon pool because of suppressed PDH. Lipids derived from *CYTB* 143B cells contained little radioactivity from glucose, but abundant radioactivity from glutamine, indicating a switch in the extent to which these nutrients supplied lipogenesis. To quantify the contribution of glucose and glutamine to lipogenic acetyl-CoA, cells were cultured in medium

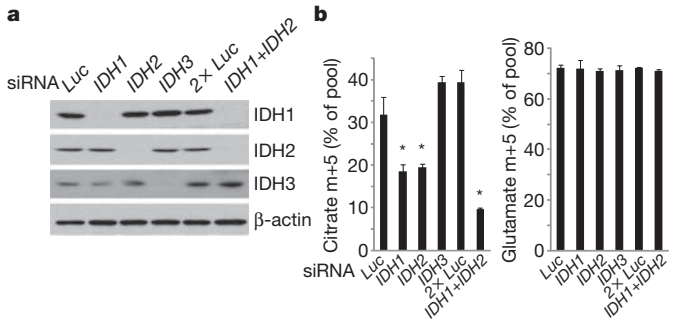


Figure 2 | NADP$^+$/NADPH-dependent isoforms of isocitrate dehydrogenase contribute to reductive carboxylation. **a**, Transient silencing of isocitrate dehydrogenase (IDH) proteins in *CYTB* 143B cells with short interfering RNAs (siRNAs) directed against *IDH1*, *IDH2* or *IDH3*. An siRNA directed against luciferase (*Luc*) was used as a negative control. When *IDH1* and *IDH2* siRNAs were transfected concurrently (*IDH1*+*IDH2*), a transfection with the same nanomolar amount of Luc siRNA (*2×Luc*) was used as a negative control. **b**, Mass isotopomer analysis of citrate and glutamate in *CYTB* 143B cells cultured with L[U-^{13}C]glutamine after silencing of IDH isoforms. Data are the average ± s.d. of three independent cultures. *$P < 0.05$, Student's *t*-test.

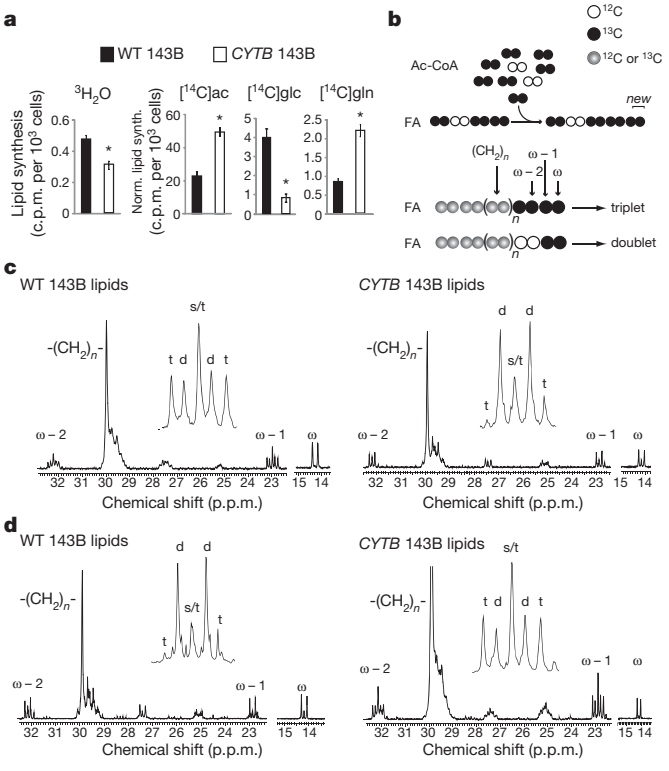


Figure 3 | Glutamine is the major lipogenic precursor in cells lacking oxidative CAC function. **a**, Cells were cultured in medium containing ^3H_2O (21 vol.%), or ^{14}C tracers of acetate (ac, 5 μCi), glucose (glc, 10 μCi) or glutamine (gln, 10 μCi). Lipids were analysed for ^{3}H or ^{14}C content by scintillation counting. In the ^{14}C experiments, raw counts per 10^3 cells were normalized to the absolute rate of fatty acid synthesis established using ^3H_2O. *$P < 0.05$, Student's *t*-test. **b**, Schematic outlining synthesis of fatty acids (FA) from acetyl-CoA (Ac-CoA) and the source of the triplet and doublet at ω − 1 in ^{13}C NMR spectroscopy. **c**, **d**, ^{13}C NMR of lipids labelled with D[U-^{13}C]glucose (**c**) or L[U-^{13}C]glutamine (**d**). Insets are expansions of the ω − 1 resonance. Abbreviations: s, singlet; d, doublet; t, triplet.

containing either D[U-^{13}C]glucose and unlabelled glutamine, or L[U-^{13}C]glutamine and unlabelled glucose. Extracted lipids were analysed by ^{13}C NMR. Because fatty acids are synthesized by the sequential addition of two-carbon units, analysing ^{13}C–^{13}C coupling between the $\omega-1$ and $\omega-2$ carbons—which enter the fatty acyl chain on different two-carbon units—enables calculation of ^{13}C enrichment in lipogenic acetyl-CoA[3]. In the $\omega-1$ multiplet, ^{13}C–^{13}C coupling produces a triplet when two ^{13}C-labelled two-carbon units are added in succession, whereas one labelled and one unlabelled unit produce a doublet (Fig. 3b). Therefore ^{13}C labelling in lipogenic acetyl-CoA is proportional to the contribution of the triplet to the overall area of the $\omega-1$ multiplet. In WT 143B cells cultured with D[U-^{13}C]glucose, the $\omega-1$ resonance was dominated by the triplet, which accounted for 65% of the total area (Fig. 3c, left). Glutamine was a minor source of fatty acyl carbon, as the triplet accounted for only 15% of the area when the cells were cultured with L[U-^{13}C]glutamine (Fig. 3d, left). This pattern was reversed in the *CYTB* 143B cells, with glutamine providing 67% of the lipogenic carbon (Fig. 3c, d, right; Supplementary Fig. 8a). Thus, reductive glutamine metabolism was the major pathway of fatty acid production in newly synthesized lipids during *CYTB* 143B growth.

Although most cancer cells have functional mitochondria, a subset of human tumours harbours mutations that impair mitochondrial metabolism. Two major classes of mutations occur in genes required for function of the CAC enzymes succinate dehydrogenase (*SDHA*, *SDHB*, *SDHC*, *SDHD*, *SDHAF2*) or fumarate hydratase (*FH*)[14]. To test whether glutamine-dependent reductive carboxylation occurs in tumour cells harbouring these mutations, we studied UOK262 cells derived from a renal tumour in a patient with hereditary leiomyomatosis renal cell carcinoma syndrome. In this syndrome, affected individuals inherit one loss-of-function *FH* allele and their tumours display loss of heterozygosity for the *FH* locus[15]. UOK262 cells are defective in respiration and devoid of FH activity[15]. They failed to survive culture in glucose-free medium containing galactose and glutamine (Fig. 4a), a condition that forces cells to use mitochondrial metabolism to generate ATP[16]. Lack of FH activity precludes the use of glutamine as a carbon source in the conventional oxidative CAC. Yet the UOK262 cells required glutamine to proliferate (Fig. 4b). Culture of these cells with D[U-^{13}C]glucose produced mass isotopomer data similar to *CYTB* 143B cells, with minimal formation of citrate m+2 (Fig. 4c, left). In cells cultured with L[U-^{13}C]glutamine, the most abundant mass isotopomer of citrate was m+5 formed by reductive carboxylation (Fig. 4c, left). Fumarate was predominantly formed through oxidative glutamine metabolism, producing fumarate m+4 (Fig. 4c, middle). However, a large fraction of the malate pool (m+3) was formed via citrate cleavage and subsequent reductive metabolism (Fig. 4c, right). Lipid labelling using D[U-^{13}C]glucose or L[U-^{13}C]glutamine revealed that, like *CYTB* 143B cells, glutamine is the major carbon source for fatty acid synthesis (Supplementary Fig. 8b). Thus these cancer cells produce CAC intermediates using two distinct pathways to compensate for the loss of FH enzyme activity: oxidative metabolism of glutamine to fumarate, and glutamine-dependent reductive carboxylation to produce citrate, acetyl-CoA for lipogenesis and remaining four-carbon CAC intermediates.

Finally, we tested whether glutamine-dependent reductive carboxylation was unique to cells with permanent abnormalities in mitochondrial metabolism. WT 143B cells and transformed mouse embryonic fibroblasts (MEFs) were subjected to acute ETC inhibition. Exposure to inhibitors of complex III (antimycin) or complex I (rotenone) induced a switch from oxidative to reductive metabolism of glutamine (Fig. 4d, Supplementary Fig. 9a). Enhanced reductive carboxylation was evident one hour after ETC inhibition and was not eliminated by pre-treatment with the protein synthesis inhibitor cycloheximide (Supplementary Fig. 9b, c). Metformin, an antidiabetic agent that inhibits respiration through effects on complex I[17], also induced glutamine-dependent reductive carboxylation at high doses (Fig. 4d, Supplementary

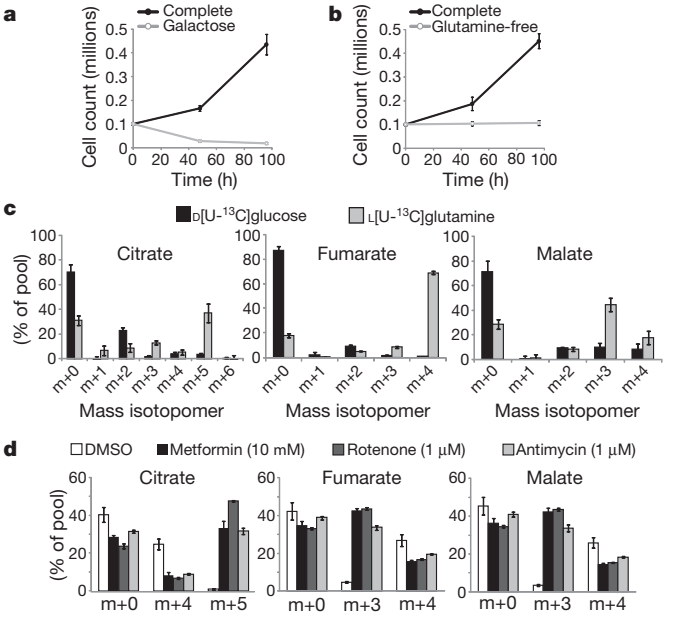


Figure 4 | Glutamine-dependent reductive carboxylation in renal carcinoma cells lacking fumarate hydratase and during ETC inhibition in tumour cells with normal mitochondria. Growth of UOK262 renal carcinoma cells in complete medium, in medium in which glucose was replaced with galactose (**a**), and in medium lacking glutamine (**b**). **c**, Mass isotopomer analysis of citrate, fumarate and malate from UOK262 cells cultured in medium containing D[U-^{13}C]glucose and unlabelled glutamine, or L[U-^{13}C]glutamine and unlabelled glucose. **d**, Mass isotopomer analysis for citrate, fumarate and malate from WT 143B cells cultured in medium containing unlabelled glucose and L[U-^{13}C]glutamine. The labelled glutamine was introduced at the same time the cells were treated with vehicle (DMSO), metformin, rotenone or antimycin. Metabolites were extracted 6 h later. In all experiments, data are the average ± s.d. of three independent cultures.

Fig. 9a). Interestingly, doses that mimicked plasma concentrations in patients with acute metformin toxicity and lactic acidosis induced reductive glutamine metabolism (Supplementary Fig. 10). These data demonstrate that glutamine-dependent reductive carboxylation is a common cellular response to impaired mitochondrial metabolism.

Glutamine-dependent reductive carboxylation has been described previously as a minor source of isocitrate/citrate and lipogenic carbon in mammalian cells[12,13,18–20]. In *Plasmodium falciparum*, a unicellular protozoan with limited respiration, the pathway is a major source of citrate and acetyl units[21]. We show that this pathway can also function as the predominant metabolic strategy to produce lipids for cancer cell growth, and can be induced acutely during ETC inhibition. We speculate that reductive carboxylation is stimulated by a disturbance in the redox ratio of the mitochondria caused by ETC impairment, decreasing the NAD^+/NADH ratio and rendering oxidative function of the CAC less efficient. This ratio could be dissipated in part by NAD(P)-transhydrogenases, which transfer reducing equivalents from NADH to NADPH, and this in turn may drive NADPH-dependent reductive carboxylation by IDH1 and IDH2. This redirection of the CAC extends the versatility of glutamine metabolism in cancer cell growth. In cells with normal mitochondria, glutamine oxidation is frequently the major source of anaplerosis, providing most of the carbon to pools of CAC intermediates and facilitating production of macromolecular precursors in the mitochondria[22]. Glutamine-dependent reductive carboxylation enables glutamine to retain this status as a growth-promoting nutrient even when mitochondrial metabolism is impaired. Therefore, during malignant transformation the acquisition of somatic mutations that impair mitochondrial function and stimulate aerobic glycolysis (the Warburg effect) does not diminish the importance of glutamine metabolism or of the mitochondria themselves for tumour cell proliferation.

METHODS SUMMARY

WT 143B, *CYTB* 143B, UOK262, CCL16-B2 and CCL16-NDI1 cells were cultured as described[9,15,23]. MEFs were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS, Hyclone), penicillin/streptomycin and 6 mM L-glutamine. Cell growth was monitored in sub-confluent cultures by trypsinizing and counting with a haemacytometer. Glucose, lactate and glutamine were measured as described[24]. Isotopic labelling was performed in DMEM with 10% dialysed FBS supplemented with either 15 mM D[U-^{13}C]glucose or 2 mM L[U-^{13}C]glutamine. Aqueous metabolites were analysed using an Agilent 6970 gas chromatograph and an Agilent 5973 mass selective detector. To determine relative metabolite abundance across samples, the area of the total ion current peak for the relevant metabolite was compared to that of an internal standard (2-oxobutyrate) and normalized for protein content. Analysis of ^{13}C enrichment and mass isotopomer distribution was performed as published[11,25]. For radioisotope lipid synthesis assays, cells were cultured for 18 h in complete medium supplemented with $^{3}H_2O$ or ^{14}C tracers (Perkin Elmer). Lipids were extracted and analysed as described[3]. For lipid synthesis experiments using NMR, cells were plated in 10-cm dishes and allowed to proliferate exponentially in medium containing D[U-^{13}C]glucose and unlabelled glutamine, or unlabelled glucose and L[U-^{13}C]glutamine until two confluent 15-cm dishes were obtained. Lipids were then extracted and analysed by ^{13}C NMR as described[3]. Gene silencing was performed using commercial small interfering RNAs (Thermo) directed against *IDH1*, *IDH2* or *IDH3*. Cells were transfected with DharmaFECT transfection reagent (Thermo) and protein abundance was examined after 72 h using western blots with commercial antibodies (Santa Cruz Biotechnology). Cells were incubated with L[U-^{13}C]glutamine for two hours before extraction of metabolites for mass isotopomer distribution analysis. Stable silencing used lentiviral shRNAs from the Mission shRNA pLKO.1-puro library (Sigma).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 June; accepted 13 October 2011.
Published online 20 November 2011.

1. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324,** 1029–1033 (2009).
2. Gao, P. *et al.* c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* **458,** 762–765 (2009).
3. DeBerardinis, R. J. *et al.* Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc. Natl Acad. Sci. USA* **104,** 19345–19350 (2007).
4. Linehan, W. M., Srinivasan, R. & Schmidt, L. S. The genetic basis of kidney cancer: a metabolic disease. *Nature Rev. Urol.* **7,** 277–285 (2010).
5. Baysal, B. E. *et al.* Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science* **287,** 848–851 (2000).
6. Tomlinson, I. P. *et al.* Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nature Genet.* **30,** 406–410 (2002).
7. Hao, H. X. *et al.* SDH5, a gene required for flavination of succinate dehydrogenase, is mutated in paraganglioma. *Science* **325,** 1139–1142 (2009).
8. Rana, M., de Coo, I., Diaz, F., Smeets, H. & Moraes, C. T. An out-of-frame cytochrome b gene deletion from a patient with parkinsonism is associated with impaired complex III assembly and an increase in free radical production. *Ann. Neurol.* **48,** 774–781 (2000).
9. Weinberg, F. *et al.* Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl Acad. Sci. USA* **107,** 8788–8793 (2010).
10. Reitzer, L. J., Wice, B. M. & Kennell, D. Evidence that glutamine, not sugar, is the major energy source for cultured HeLa cells. *J. Biol. Chem.* **254,** 2669–2676 (1979).
11. Cheng, T. *et al.* Pyruvate carboxylase is required for glutamine-independent growth of tumor cells. *Proc. Natl Acad. Sci. USA* **108,** 8674–8679 (2011).
12. Des Rosiers, C., Fernandez, C. A., David, F. & Brunengraber, H. Reversibility of the mitochondrial isocitrate dehydrogenase reaction in the perfused rat liver. Evidence from isotopomer analysis of citric acid cycle intermediates. *J. Biol. Chem.* **269,** 27179–27182 (1994).
13. Ward, P. S. *et al.* The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17,** 225–234 (2010).
14. Frezza, C., Pollard, P. J. & Gottlieb, E. Inborn and acquired metabolic defects in cancer. *J. Mol. Med.* **89,** 213–220 (2011).
15. Yang, Y. *et al.* UOK 262 cell line, fumarate hydratase deficient (FH-/FH-) hereditary leiomyomatosis renal cell carcinoma: *in vitro* and *in vivo* model of an aberrant energy metabolic pathway in human cancer. *Cancer Genet. Cytogenet.* **196,** 45–55 (2010).
16. Rossignol, R. *et al.* Energy substrate modulates mitochondrial structure and oxidative capacity in cancer cells. *Cancer Res.* **64,** 985–993 (2004).
17. El-Mir, M. Y. *et al.* Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. *J. Biol. Chem.* **275,** 223–228 (2000).
18. Yoo, H., Antoniewicz, M. R., Stephanopoulos, G. & Kelleher, J. K. Quantifying reductive carboxylation flux of glutamine to lipid in a brown adipocyte cell line. *J. Biol. Chem.* **283,** 20621–20627 (2008).
19. Comte, B., Vincent, G., Bouchard, B., Benderdour, M. & Des Rosiers, C. Reverse flux through cardiac NADP(+)-isocitrate dehydrogenase under normoxia and ischemia. *Am. J. Physiol. Heart Circ. Physiol.* **283,** H1505–H1514 (2002).
20. Lemons, J. M. *et al.* Quiescent fibroblasts exhibit high metabolic activity. *PLoS Biol.* **8,** e1000514 (2010).
21. Olszewski, K. L. *et al.* Branched tricarboxylic acid metabolism in *Plasmodium falciparum*. *Nature* **466,** 774–778 (2010).
22. DeBerardinis, R. J. & Cheng, T. Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. *Oncogene* **29,** 313–324 (2009).
23. Brunelle, J. K. *et al.* Oxygen sensing requires mitochondrial ROS but not oxidative phosphorylation. *Cell Metab.* **1,** 409–414 (2005).
24. Yang, C. *et al.* Glioblastoma cells require glutamate dehydrogenase to survive impairments of glucose metabolism or Akt signaling. *Cancer Res.* **69,** 7986–7993 (2009).
25. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of ^{13}C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom.* **31,** 255–262 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements K. Uyeda and members of the DeBerardinis and Chandel laboratories provided critical assessment of the data, and J. Sudderth and C. Yang provided experimental assistance. We thank C. Moraes and I.F.M. de Coo for the WT 143B and *CYTB* 143B cell lines, I.E. Scheffler for CCL16-B2 cells and T. Yagi for CCL16-NDI1 cells. This work was supported by grants to R.J.D. from the NIH (K08DK072565 and R01CA157996), the Cancer Prevention and Research Institute of Texas (CPRIT, HIRP100437) and the Robert A. Welch Foundation (I1733); to N.S.C. from the NIH (R01CA123067), the LUNGevity Foundation and a Consortium of Independent Lung Health Organizations convened by Respiratory Health Association of Metropolitan Chicago; and to E.S.J. by an NIH grant (DK078933). The work was also supported by the Intramural Research Program of the NIH, National Cancer Institute Center for Cancer Research and by NIH grant RR02584. NIH training grants supported A.R.M. (5T32GM083831), W.W.W. (T32CA009560) and L.B.S. (T32GM008061). T.C. was supported by a CPRIT training grant.

Author Contributions A.R.M., W.W.W., N.S.C. and R.J.D. designed the research. A.R.M., W.W.W., L.B.S., E.S.J., T.C. and P.-H.C. performed the experiments. A.R.M., W.W.W., L.B.S., E.S.J., P.-H.C., T.C., N.S.C. and R.J.D. analysed the data. Y.Y. and W.M.L. provided the FH-deficient (UOK262) cells. A.R.M., N.S.C. and R.J.D. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.J.D. (ralph.deberardinis@utsouthwestern.edu) or N.S.C. (nav@northwestern.edu).

METHODS

Cell culture, growth and viability. WT 143B, *CYTB* 143B, CCL16-B2, CCL16-NDI1 and UOK262 cells were cultured as described[9,15,23]. Mouse embryonic fibroblasts (MEFs) were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS, Hyclone), penicillin/streptomycin and 6 mM L-glutamine. Doubling times were calculated by repeatedly allowing sub-confluent cultures to reach confluence, then trypsinizing, counting and re-plating at low confluence.

Nutrient utilization and secretion. Concentrations of glucose, lactate and glutamine in tissue culture medium were measured using a NOVA BioProfile Basic4 chemical analyser as described[24]. Rates of consumption/secretion were calculated by determining the change in metabolite concentration over 6 h, converting this change into nmoles consumed or secreted, and normalizing to protein content.

Mass isotopomer distribution and mass spectrometry. Nutrients labelled with ^{13}C were purchased from Cambridge Isotope Laboratories. All ^{13}C studies were performed in medium containing 10% dialysed FBS and prepared so that 100% of either the glucose or glutamine pool was labelled with ^{13}C, and the other pool was unlabelled. DMEM lacking both glucose and glutamine was prepared from powder (Sigma), then supplemented with either 15 mM D[U-^{13}C]glucose and 2 mM unlabelled glutamine, or 15 mM unlabelled glucose and 2 mM L[U-^{13}C]glutamine. For experiments with RPMI (Supplementary Fig. 4), we supplemented glutamine-free RPMI (11 mM unlabelled glucose, Invitrogen) with 2 mM L[U-^{13}C]glutamine. Cells were grown in 60- or 100-mm dishes until 80% confluent, then rinsed with PBS and cultured with ^{13}C-containing medium for the indicated time (usually 6 h). The cells were then extracted by freeze-thawing three times in 0.5 ml of a cold 1:1 mixture of methanol:water. Macromolecules and debris were removed by centrifugation, an internal standard was added (50 nmol of 2-oxobutyrate), and the supernatants with aqueous metabolites were evaporated completely and derivatized for 30 min at 42 °C in 100 µl of a trimethylsilyl donor (Tri-Sil, Thermo). Metabolites were analysed using an Agilent 6970 gas chromatograph networked to an Agilent 5973 mass selective detector. Retention times and mass fragmentation signatures of all metabolites were validated using pure standards. To determine relative metabolite abundance across samples, the area of the total ion current peak for the metabolite of interest was compared to that of the internal standard and normalized for protein content. The mass isotopomer distribution analysis measured the fraction of each metabolite pool that contained every possible number of ^{13}C atoms; that is, a metabolite could contain 0, 1, 2, … n ^{13}C atoms, where n = the number of carbons in the metabolite. For each metabolite, an informative fragment ion containing all carbons in the parent molecule was analysed using MSDChem software (Agilent), integrating the abundance of all mass isotopomers from $m+0$ to $m+n$, where m = the mass of the fragment ion without any ^{13}C. A list of fragment ions is in ref. 11. The abundance of each mass isotopomer was then corrected mathematically to account for natural abundance isotopes[25] and finally converted into a percentage of the total pool.

Lipid synthesis assays. All radioisotopes were from Perkin-Elmer. Culture with ^3H_2O was used to measure total lipid synthesis by preparing high-glucose DMEM from powder (Sigma) and supplementing with 2 mM L-glutamine and 10% FBS. The ^3H_2O was added during preparation of this medium to a final content of 21% of the total volume. The ^{14}C assays also used high-glucose DMEM with 2 mM L-glutamine, 10% FBS and tracer quantities of [1,2-^{14}C]acetate (5 µCi, 4.4 µM), D[U-^{14}C]glucose (10 µCi, 1.7 µM) or L[U-^{14}C]glutamine (10 µCi, 1.9 µM). In all radioisotope assays, cells were grown to 80–90% confluence then cultured for 18 h in medium containing the label. Lipids were then extracted and analysed by scintillation counting as previously described[3]. For ^{13}C NMR experiments, cells were plated at approximately 50% confluence in 10-cm dishes and allowed to proliferate exponentially in complete DMEM containing either 15 mM D[U-^{13}C]glucose and 2 mM unlabelled glutamine, or 15 mM unlabelled glucose and 2 mM L[U-^{13}C]glutamine until two confluent 15-cm dishes were obtained. Glucose and glutamine concentrations were monitored in the medium to ensure that neither nutrient was depleted during expansion of the culture. Lipids were then extracted and analysed by ^{13}C NMR as described[3] using a Varian ANOVA 14.1 T spectrometer equipped with a 3-mm broadband probe with the observe coil tuned to ^{13}C (150 MHz). Relevant peak areas were determined using ACDLabs SpecManager (Advanced Chemistry Development).

RNA interference. Transient gene silencing experiments were performed using commercial siRNA pools (siGENOME, Thermo) directed against *IDH1*, *IDH2*, *IDH3* (catalogue numbers MU-008294-00, MU-004013-00 and MU-008753-00, respectively). For the simultaneous silencing of *IDH1* and *IDH2*, individual siRNAs rather than pools were used (*IDH1*: 5′-GGACUUGGCUGCUUGC AUU-3′; *IDH2*: 5′-CAAGAACUAUGACGGAGAU-3′). Cells were transfected with DharmaFECT transfection reagent (Thermo), and protein abundance was examined after 72 h using western blots with commercial antibodies against IDH1, IDH2 or IDH3 (Santa Cruz Biotechnology). Cells were incubated with L[U-^{13}C]glutamine for two hours before extraction of metabolites for mass isotopomer distribution analysis. Stable silencing used lentiviral-mediated shRNAs from the Mission shRNA pLKO.1-puro library (Sigma). Cells were infected with each lentivirus using the manufacturer-supplied protocol, and pools of puromycin-resistant cells were used for subsequent experiments. Pools with the best silencing of IDH1 or IDH2 were used in the growth assay.

Subcellular localization studies. WT 143B and *CYTB* 143B cells were transfected with siRNA as described above. After 72 h, the transfected cells were incubated in medium containing 1 µM Mitotracker Deep Red FM (Invitrogen) for 30 min. The cells were then fixed with 3.7% formaldehyde in PBS for 15 min, permeabilized with 0.5% (v/v) Triton X-100 for 5 min, and blocked with 1% bovine serum albumin for 5 min. Immunofluorescence used commercial antibodies against IDH1 (Santa Cruz Biotechnology) or IDH2 (Abcam) followed by anti-goat or anti-mouse secondary antibodies conjugated to Alexa Fluor-488. Cells were stained with primary and secondary antibodies for one hour at room temperature and mounted on Vectashield mounting medium containing DAPI (Vector Laboratories). Images were acquired using an LD Plan-Neofluar 40×/1.3 DIC objective on a Zeiss Axioplan 2E imaging deconvolution microscope. Mitochondrial fractions were isolated by differential centrifugation as previously described[26].

26. Frezza, C., Cipolat, S. & Scorrano, L. Organelle isolation: functional mitochondria from mouse liver, muscle and cultured fibroblasts. *Nature Protocols* **2,** 287–295 (2007).

LETTER

doi:10.1038/nature10730

Differential oestrogen receptor binding is associated with clinical outcome in breast cancer

Caryn S. Ross-Innes[1], Rory Stark[1], Andrew E. Teschendorff[2], Kelly A. Holmes[1], H. Raza Ali[1,8], Mark J. Dunning[1], Gordon D. Brown[1], Ondrej Gojis[3,4,5], Ian O. Ellis[6], Andrew R. Green[6], Simak Ali[3], Suet-Feung Chin[1], Carlo Palmieri[3], Carlos Caldas[1,7,8,9] & Jason S. Carroll[1,7]

Oestrogen receptor-α (ER) is the defining and driving transcription factor in the majority of breast cancers and its target genes dictate cell growth and endocrine response, yet genomic understanding of ER function has been restricted to model systems[1–3]. Here we map genome-wide ER-binding events, by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), in primary breast cancers from patients with different clinical outcomes and in distant ER-positive metastases. We find that drug-resistant cancers still recruit ER to the chromatin, but that ER binding is a dynamic process, with the acquisition of unique ER-binding regions in tumours from patients that are likely to relapse. The acquired ER regulatory regions associated with poor clinical outcome observed in primary tumours reveal gene signatures that predict clinical outcome in ER-positive disease exclusively. We find that the differential ER-binding programme observed in tumours from patients with poor outcome is not due to the selection of a rare subpopulation of cells, but is due to the FOXA1-mediated reprogramming of ER binding on a rapid timescale. The parallel redistribution of ER and FOXA1 binding events in drug-resistant cellular contexts is supported by histological co-expression of ER and FOXA1 in metastatic samples. By establishing transcription-factor mapping in primary tumour material, we show that there is plasticity in ER-binding capacity, with distinct combinations of *cis*-regulatory elements linked with the different clinical outcomes.

Recent technological advances have allowed mapping of ER-binding events, with the goal of discovering the *cis*-regulatory elements and factors involved in mediating ER binding and transcription. Several genome-wide maps of ER in breast cancer cell-line models exist[1–3], all showing that most ER-binding events occur at distal *cis*-regulatory elements. Forkhead motifs are enriched within the regions bound by ER binding and multiple studies have identified the forkhead protein FOXA1 as an important pioneer factor for ER–chromatin interactions[4–6]. However, the ER mapping studies have been restricted to breast cancer cell lines, mostly the MCF-7 cell line. We sought to interrogate ER-binding events, for the first time, in primary frozen breast cancer samples, to determine if ER binding is dynamic and if specific *cis*-regulatory elements can distinguish tumours from patients with distinct clinical outcomes.

ER ChIP-seq was performed in eight ER^+, progesterone receptor $(PR)^+$, $HER2^-$ primary breast tumours, representative of tumours from patients with better prognosis[7], a conclusion supported by the available long-term clinical follow-up (Supplementary Fig. 1). Also included were seven primary breast tumours from patients with a poor outcome (ER^+ PR^- $HER2^-$ or ER^+ PR^+ HER^+), because PR^- or $HER2^+$ tumours are more likely to be aggressive[8,9]. As expected, the poor outcome patients who had long-term clinical follow-up died of breast cancer (Supplementary Fig. 1). Furthermore, three ER^+ distant metastatic samples from women with breast cancer were included. The metastatic locations and sample preparation can be found in Supplementary Fig. 1. As a control, we included two breast cancer samples that were ER^- (ER-α negative), but expressed high transcript levels of ER-β.

ER ChIP-seq was conducted and ER-binding peaks were called using two different algorithms, MACS[10] and SWEMBL (http://www.ebi.ac.uk/~swilder/SWEMBL/), to minimize peak caller bias. The number of sequencing reads and ER-binding events for each tumour is shown in Supplementary Fig. 2. ER binding could be mapped in all tumours, but total peak intensity and the number of identified binding events differed. Three tumours were split into two sections and ER ChIP-seq was conducted on the separate sections. We found very good concordance when comparing different sections of the same tumour ($R^2 = 0.954$) suggesting that tumour heterogeneity did not substantially influence the ER-binding signal obtained from a sample (Supplementary Fig. 3).

We initially assessed whether a conserved set of breast cancer ER-binding events could be identified. We found a core set of 484 ER-binding events that were identified in at least 75% of all the tumours, but not in either of the ER^- tumours (Fig. 1a). Peak calling details can be found in Supplementary Fig. 4. An example of a core ER-binding event is shown in Fig. 1b. This demonstrates that ER binding to chromatin still occurs even in tumours that are unlikely to respond to anti-oestrogen therapies, implying that drug resistance is not due to loss of ER binding to DNA. The average ER-binding signal intensity was highest in the metastatic samples and lowest in patients with good outcome tumours, a phenomenon observed both within the 484 core ER-binding regions (Fig. 1c) and globally (Supplementary Fig. 4). These data indicate that there is an acquisition of binding signal intensity in tumours that progress towards a poorer prognosis and ultimately metastasize. The only DNA motif found to be enriched in the core ER-binding events was an oestrogen responsive element (ERE) (Fig. 1d). The genes near (within 20 kilobases (kb): an optimal window between ER-binding events and target genes[11]) the 484 core ER-binding events exhibited elevated expression in the ER^+ tumours used for ChIP-seq, as compared to all other genes (data not shown) and were higher in ER^+ tumours relative to ER^- tumours in nine independent data sets (Supplementary Fig. 5). The genes are provided in Supplementary Fig. 6 and include classic ER target genes such as *TFF1*, *GREB1* and *RARA*. A gene predictor was generated based on genes near the core ER-binding events. Patients were stratified and the tumours with the highest 'risk index' had a poor clinical outcome when compared to the tumours with the lowest 'risk index' (Fig. 1e shows the results based on one study[12] and additional data sets are shown in Supplementary Fig. 6; only ER^+ patients were considered). These conserved *cis*-regulatory elements

[1]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [2]UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK. [3]Imperial College London, Hammersmith Campus, London W12 0NN, UK. [4]Department of Gynaecology and Obstetrics, Third Faculty of Medicine, Charles University, Ruska 87, Prague 10, 100 00, Czech Republic. [5]Department of Pathology, Third Faculty of Medicine, Charles University, Ruska 87, Prague 10, 100 00, Czech Republic. [6]Department of Histopathology, Nottingham University Hospitals NHS Trust, Nottingham City Hospital, Nottingham NG5 1PB, UK. [7]Department of Oncology, University of Cambridge CB2 0XZ, UK. [8]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [9]Cambridge Experimental Cancer Medicine Centre (ECMC), Cambridge CB2 0RE, UK.

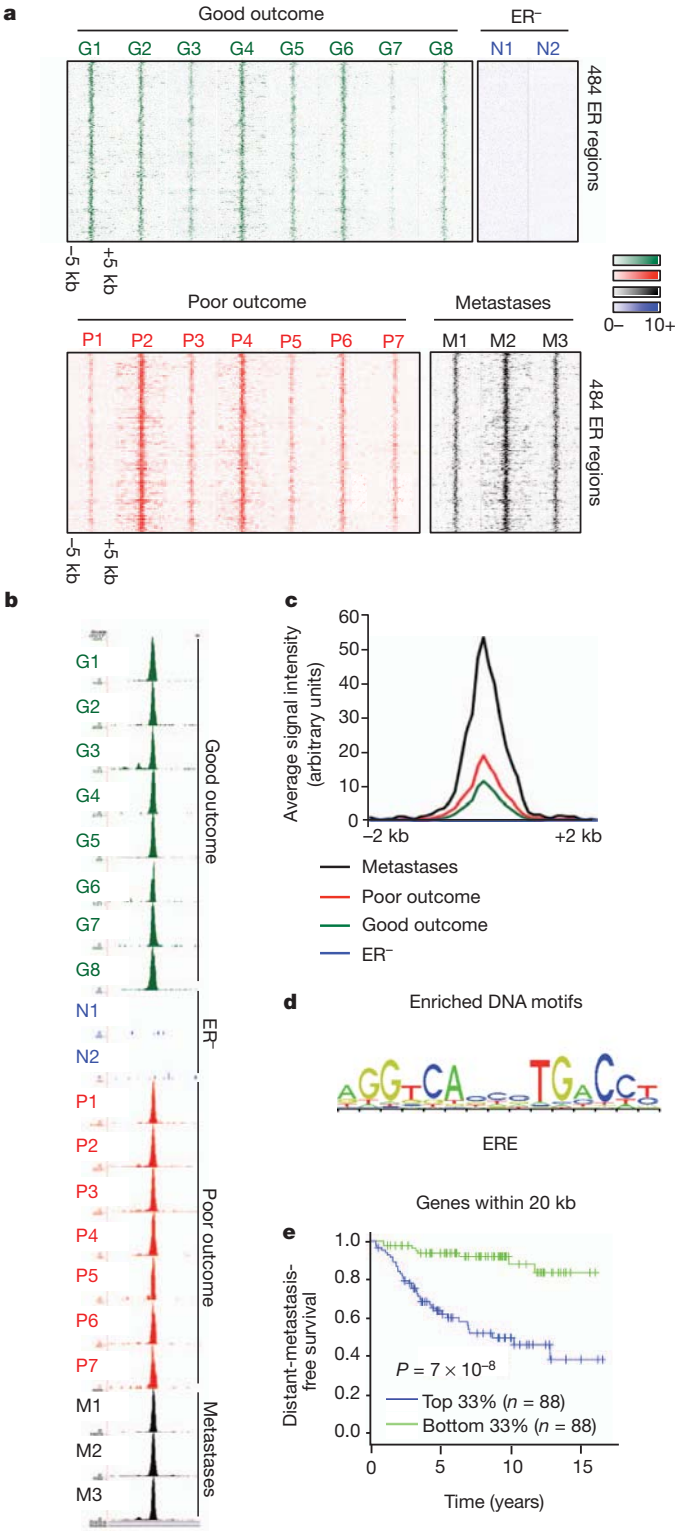


Figure 1 | A subset of ER-binding events is conserved in primary breast tumours and distant metastases. **a**, Heat map showing binding peak intensity of 484 core ER-binding events that are common to primary breast tumours and distant metastases. The window represents ±5 kb regions from the centre of the binding events. **b**, Example of an ER-binding event at the *RARA* locus, which is present in all primary ER^+ breast tumours and metastases, but not the ER^- tumours. **c**, Normalized average signal intensity of all core ER-binding events. Also included are the ER^- tumours. **d**, Motif analysis revealed the enrichment of EREs. **e**, Genes within 20 kb of the core ER-binding events were used to generate a gene predictor that was tested in independent data sets for predictive value. Tumours were stratified according to expression of this gene signature and the top one third and bottom one third of tumours were compared. Results from one data set are shown[12] and additional data sets are provided in Supplementary Fig. 6.

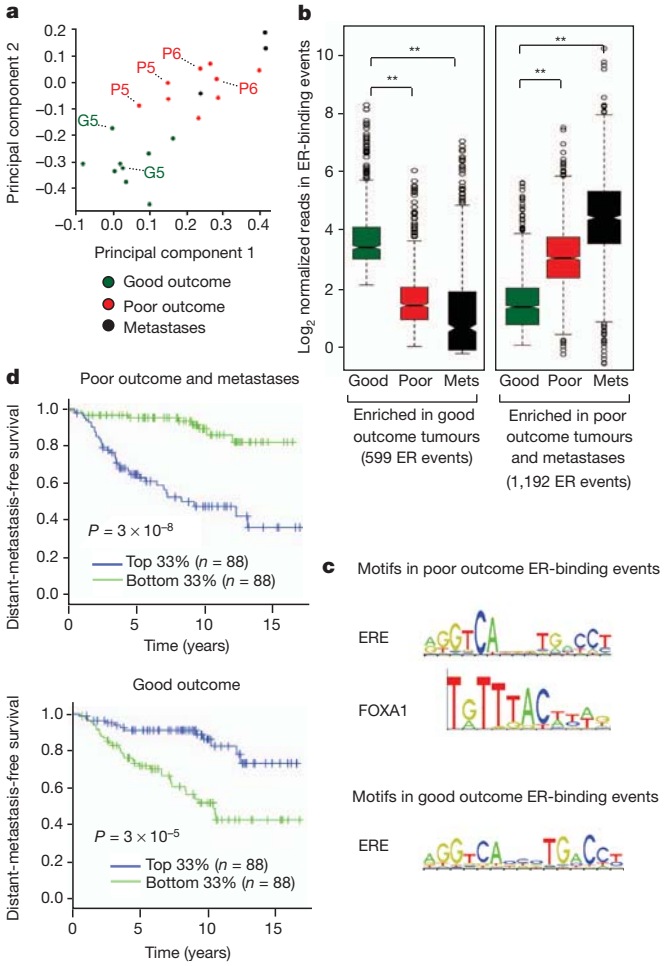


Figure 2 | ER-binding profiles can discriminate between tumours from patients with different clinical outcomes. **a**, Principal component analysis (44% of total variance) of the 1,791 ER-binding events that can discriminate between the patients with good outcome tumours and those with poor/met tumours. Included are the replicates from three tumours, which are highlighted. **b**, Box plot representing distribution of normalized read counts in differential ER-binding events that are statistically enriched in either the patients with good outcome tumours (599 ER-binding events) or the patients with poor outcome tumours and metastases (Mets) (1,192 ER-binding events). The samples were pre-normalized. $^{**}P < 1 \times 10^{-10}$. *P* values were calculated using the Wilcoxon rank sum test. **c**, Enriched motifs in the poor/met and the good outcome ER-binding events. **d**, Genes within 20 kb of the differentially bound ER-binding events were used to generate a gene predictor. Tumours were stratified according to expression of this gene signature and the top one third and bottom one third of tumours were compared. Results from one data set are shown[12] and additional data sets are provided in Supplementary Fig. 12.

and their putative target genes may be the elements that contribute to tumorigenesis and are maintained regardless of the clinical outcome of the breast cancer patient. In contrast to the primary breast cancers and metastases, we mapped ER binding in three normal human mammary glands and two normal human liver samples and found limited numbers of ER-binding events, with almost no concordance in ER binding between individuals (data not shown).

We sought to determine if differential ER-binding events could discriminate the patients with good outcome (ER^+ PR^+ $HER2^-$ tumours), from patients with poor outcome or metastases (we described the combined set as poor/met tumours). After normalization of the data to account for global differences in ER binding, differential binding analysis (DBA) was used to identify ER-binding events that were statistically enriched in one category or the other. This resulted in a set of ER-binding events that could discriminate between the two groups when using principal component analysis

(Fig. 2a). In total, DBA revealed 1,192 genomic regions that had significantly more ER binding in the poor/met group compared to the good outcome patients (Fig. 2b), and 599 ER binding regions with more ER binding in the good outcome patients when compared to the poor/met patients (Fig. 2b). The clustering of the tumours on the basis of the 1,791 differential ER-binding events can be visualized in Supplementary Fig. 7. These findings suggest that there are specific and re-occurring *cis*-regulatory elements that are occupied by ER in breast cancers, but that these are different in tumours that respond to treatment versus those that relapse and metastasize. Analysis of enriched DNA motifs identified the presence of ERE and FOXA1 motifs in the differential poor outcome ER-binding events and ERE motifs in the good outcome ER-binding events (Fig. 2c). Correlation of the poor outcome ER-binding events with known processes revealed an association with endocrine resistance and luminal B status (Supplementary Fig. 8).

To investigate if the genes near the differential ER-binding events were potentially functional in breast cancer, we analysed genes within a 20 kb window around the 1,192 poor/met and 599 good outcome ER-binding events. Using a training set, we generated a gene expression predictor for each of the good and poor outcome gene lists. The probability calculation and comparisons between the good and poor outcome genes is shown in Supplementary Figs 9 and 10. Within the poor outcome gene list was the oncogene *ERBB2* (all genes are shown in Supplementary Fig. 11). As expected, genes in the poor outcome predictor were preferentially upregulated in poor outcome patients, whereas those in the good outcome predictor were preferentially downregulated (Supplementary Fig. 10). We next tested the predictors in an independent large cohort of breast cancer patients[12], only considering ER^+ tumours. Using distant-metastasis-free survival as an endpoint, both gene sets predicted outcome ($P = 3 \times 10^{-5}$ for good and 3×10^{-8} for poor outcome genes) in this data set[12] and with the expected opposite directionality (Fig. 2d). The gene predictors were associated with survival in additional data sets (Supplementary Fig. 12) and were largely independent of histopathological factors (Supplementary Fig. 13). We tested 1,000 randomizations from the entire list of genes and determined that the probability that a random set of genes would yield an equally robust predictor of clinical outcome was $P = 0.004$. Furthermore, the good and poor gene predictors had no predictive power in four cohorts of ER^- patients (Supplementary Fig. 14). This suggests that the increased ER binding at distinct *cis*-regulatory elements is functionally and biologically relevant, resulting in altered gene expression profiles that contribute to differences in drug response and overall survival.

To validate the findings made in the tumours, we explored the possibility that ER-binding events were acquired in cell-line models of endocrine resistance. ER binding was mapped by ChIP-seq in three commonly used tamoxifen-responsive, ER^+ breast cancer cell lines (MCF-7, T-47D and ZR75-1), and two tamoxifen-resistant, ER^+ breast cancer cell lines, namely a tamoxifen-resistant MCF-7 derivative (TAM-R)[13] and BT-474 cells that are ER^+ and contain the *ERBB2* amplification (ER^+ $HER2^+$). Similar to BT474 cells, TAM-R cells have increased ERBB2 protein levels[13] and both represent cellular systems in which increased growth factor signalling results in endocrine resistance. For all five cell lines, ER ChIP-seq was performed in at least duplicate, in asynchronous cells to recapitulate the situation observed in primary tumours (Supplementary Fig. 2).

Almost 7,000 (6,920) ER-binding events were identified in all replicates of all cell lines (Fig. 3a and Supplementary Fig. 15). Almost all (98.9%) of the core ER-binding events that occurred in most primary tumours (Fig. 1a) overlapped with the cell-line core ER-binding events. DBA identified 8,188 ER-binding events with significantly stronger binding affinity in the tamoxifen-resistant cell lines, and 5,713 ER-binding events that were stronger in the tamoxifen-responsive cell lines (Fig. 3a). Examples of differentially bound regions are shown in Fig. 3b. Using the differential ER-binding events, the cell-line classification can be visualized by principal component analysis

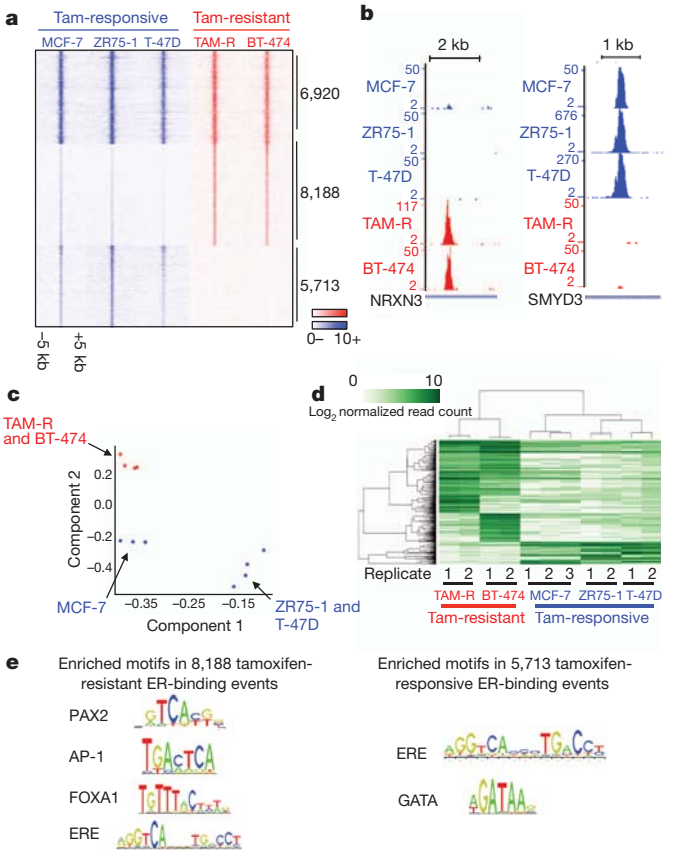


Figure 3 | Identification of a tamoxifen-resistant ER-binding profile. **a**, Heat map representing ER-binding events found in all tamoxifen (Tam)-responsive and tamoxifen-resistant cell lines, or those enriched in either responsive or resistant cell lines. The window represents ±5 kb regions from the centre of the binding events. Numbers indicate the total number of ER-binding events in each category. **b**, Examples of ER-binding events that are unique to either tamoxifen-responsive or tamoxifen-resistant cell lines. **c**, Principal component analysis (65% of total variance) of differential ER-binding events observed in responsive (blue) or resistant (red) breast cancer cells. **d**, Hierarchical clustering of the top 1,500 differentially bound ER-binding events (FDR <0.006). **e**, Enriched motifs within the ER-binding events that discriminate between tamoxifen-sensitive and resistant cell lines.

(Fig. 3c) and hierarchical clustering (Fig. 3d). Enriched motif analysis revealed ERE and FOXA1 motifs in regions showing increased ER binding in tamoxifen-resistant cell lines (Fig. 3e), which are the same motifs observed in the poor outcome ER-binding events in primary tumours (Fig. 2c). GATA motifs were enriched in ER-binding events depleted during acquisition of drug resistance (Fig. 3e), possibly due to competition between FOXA1 and GATA3, another prominent breast cancer transcription factor.

We hypothesized that the ER-binding events induced in the tamoxifen-resistant breast cancer cell lines would be the same regions that were enriched in the poor/met clinical samples. However, 79.8% of the 1,192 ER binding events enriched in the poor/met samples (from Fig. 2b) overlap with ER-binding events in wild-type MCF-7 cells, suggesting that the cell-line models are closer to the tumours and metastases from poor outcome patients. In support of this, the 599 good outcome ER-binding events observed in primary tumours (Fig. 2b) overlap poorly with the ER binding observed in MCF-7 cells (30.2% versus 79.8% for the poor outcome). Interestingly, MCF-7 cells (plus T-47D and ZR75-1 cells) were derived from the pleural effusion of metastatic breast cancer patients, but were established before tamoxifen use in the clinic. We propose that MCF-7, ZR75-1 and T-47D cell lines possess an intermediate ER-binding profile with the acquisition of additional ER-binding regions required for resistance to anti-oestrogen treatment.

The differences in ER binding between responsive and resistant contexts may be due to selection and expansion of a resistant subpopulation, or may be due to reprogramming of ER binding following specific stimuli. Growth factor pathways have long been implicated in modulating endocrine response[14,15] and have been shown to influence ER-binding potential and gene expression profiles[14]. We identified various stimuli (EGF, IL-6, TNF-α and IGF-I) shown to induce increased cellular invasion and drug resistance and treated asynchronous MCF-7 cells with control or the cocktail of mitogens for 90 min (Fig. 4a). Duplicate ER ChIP-seq replicates were performed (Supplementary Fig. 2).

Differential binding analysis identified 6,089 ER-binding regions that were differentially enriched (≥4 fold change difference, false discovery rate (FDR) <0.1) following a 90 min treatment with the cocktail. These mitogen-induced ER-binding differences could be visualized using principal component analysis (Fig. 4b). Because FOXA1 is a pioneer factor required for ER–chromatin interactions[4] and FOXA1 motifs were enriched in both the mitogen-induced ER-binding events (Supplementary Fig. 16) and the tumours from the poor outcome patients (Fig. 2c), we assessed whether the rapid, reprogrammed ER binding occurred at regions pre-determined by FOXA1. We repeated the mitogen treatment, but mapped FOXA1 binding by ChIP-seq and found that ~25% of the reprogrammed ER-binding events (1,515) occur at regions that are already bound by FOXA1, before mitogen treatment (Fig. 4c). A substantial proportion (37.6%) of the other reprogrammed ER-binding events occur at regions where FOXA1 binding is also induced by mitogens. As such, ~53% of mitogen-induced ER-binding events occur at regions pre-bound by FOXA1 or at regions that also acquire FOXA1 binding (Fig. 4c), a level of concordance that mirrors the ~50% overlap observed in wild-type cells[6].

To determine if FOXA1 expression was present in ER$^+$ distant metastases, we obtained 24 metastatic samples (bone, brain and liver) from ER$^+$ breast cancer patients and performed immunohistochemistry for ER and FOXA1 (Fig. 4d). We found that ~87% of the metastases retained ER expression and that FOXA1 expression occurred in ~95% of the metastases (Fig. 4e). Importantly, the concordance between ER and FOXA1 was high ($R^2 = 0.585$), regardless of the site of metastasis. Therefore, the co-expression of ER and FOXA1 in distant metastases supports our conclusions that FOXA1 mediates ER reprogramming.

By mapping ER binding in clinical samples, we provide a first glimpse of the primary regulatory regions that contribute to differences within ER$^+$ breast cancers, rather than secondary events such as gene expression profiles. Our findings suggest that there is plasticity in ER binding, with distinct ER-binding profiles associated with clinical outcome. These differential ER-binding profiles seem to be mediated by FOXA1. A remaining question is what dictates differential FOXA1 and subsequently ER binding. Possibilities include changes in the genomic landscape, alterations in co-factor levels or changes in FOXA1 structure and function, potentially by post-translational modifications. By establishing transcription factor mapping in primary samples, we find that differential ER-binding patterns govern gene expression programs and are associated with clinical outcome in ER$^+$ cancer.

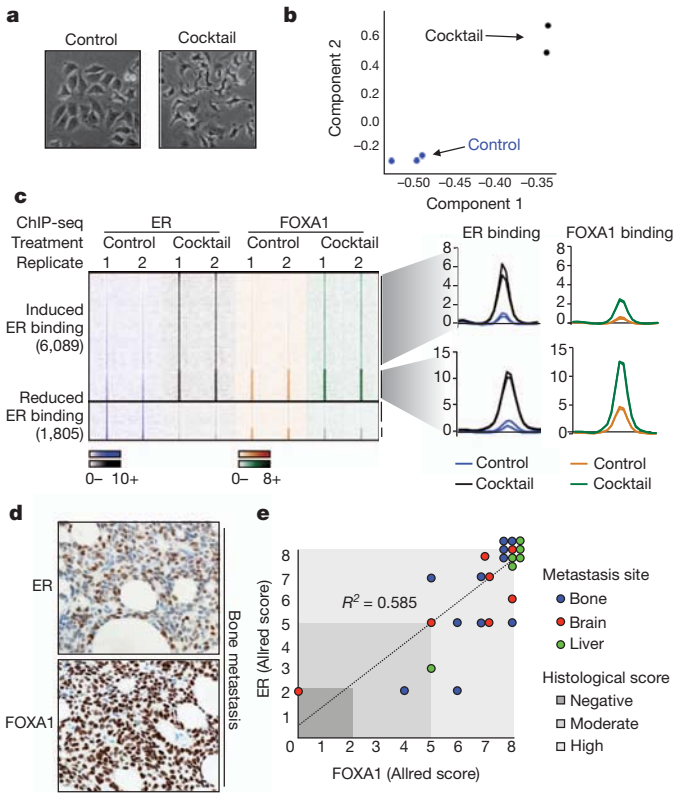


Figure 4 | ER and FOXA1 binding is dynamic and their expression correlates in metastases. **a**, Morphological changes in MCF-7 cells treated with a cocktail of mitogens (EGF, IGF-1, IL-6 and TNF-α) for 90 min. Magnification, ×40. **b**, Principal component analysis (93% of total variance) of differentially bound ER-binding events can discriminate between the control and mitogenic cocktail-treated samples in the first (and second) component. **c**, Heat map showing induction and loss of ER and FOXA1 binding after 90 min of treatment with the mitogenic cocktail. Also shown is the normalized average signal intensity of all ER- and FOXA1-binding events within the 'gained' ER-binding events. **d**, Metastatic samples from different sites were stained by immunohistochemistry for ER and FOXA1. Example of ER and FOXA1 expression in a breast cancer bone metastasis. Magnification, ×40. **e**, Graphical representation of ER and FOXA1 co-expression in metastases. Location of metastases is provided, as is the Allred score.

METHODS SUMMARY

MCF-7, ZR75-1, T-47D and BT-474 human cell lines were obtained from the American type culture collection (ATCC) and grown in the relevant media. TAM-R cells[13] were a gift from I. Hutcheson and R. Nicholson (Cardiff). The ER$^+$ breast cancer tumours were obtained from the Nottingham Tenovus primary breast cancer series, Addenbrooke's Hospital and Imperial College Healthcare NHS Trust, with appropriate ethical approval from the repositories. The malignant pericardial effusion and the two distant metastases were obtained from Imperial College Healthcare NHS Trust. For ChIP in the tumours and metastases, the frozen sample was cut into smaller pieces before ChIP, which was then performed as previously described[16]. For the malignant pericardial effusion, epithelial cells were first enriched using Dynabeads conjugated with Epcam[17]. For ChIPs from cell-line material, proliferating cells were cross-linked and processed for ChIP as previously described[16]. The antibodies used were anti-ER (sc-543) from Santa Cruz Biotechnologies and anti-FOXA1 (ab5089) from Abcam. Sequences generated by the Illumina Genome Analyzer were processed by the Illumina analysis pipeline version 1.6.1, and aligned to the Human Reference Genome (assembly hg18, NCBI build 36.1, March 2008) using BWA version 0.5.5 (ref. 18). Differential binding analysis was performed using the DiffBind package[19]. For immunohistochemical analyses, ER staining was conducted using the 6F11/2 mouse monoclonal antibody (Novocastra, Leica Microsystems) and FOXA1 staining was conducted using a rabbit polyclonal antibody (ab23738) from Abcam. An Allred scoring system was used to assess staining accounting for both staining intensity and the proportion of cells stained.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 May; accepted 23 November 2011.
Published online 4 January 2012.

1. Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nature Genet.* **38,** 1289–1297 (2006).
2. Lin, C. Y. *et al.* Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genet.* **3,** e87 (2007).
3. Welboren, W. J. *et al.* ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.* **28,** 1418–1428 (2009).

4. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122,** 33–43 (2005).
5. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132,** 958–970 (2008).
6. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genet.* **43,** 27–33 (2011).
7. EBCTCG. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365,** 1687–1717 (2005).
8. Kun, Y. *et al.* Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. *Hum. Mol. Genet.* **12,** 3245–3258 (2003).
9. Arpino, G. *et al.* Estrogen receptor-positive, progesterone receptor-negative breast cancer: association with growth factor receptor expression and tamoxifen resistance. *J. Natl Cancer Inst.* **97,** 1254–1261 (2005).
10. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
11. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
12. Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.* **25,** 1239–1246 (2007).
13. Knowlden, J. M. *et al.* Elevated levels of epidermal growth factor receptor/c-erbB2 heterodimers mediate an autocrine growth regulatory pathway in tamoxifen-resistant MCF-7 cells. *Endocrinology* **144,** 1032–1044 (2003).
14. Lupien, M. *et al.* Growth factor stimulation induces a distinct ERα cistrome underlying breast cancer endocrine resistance. *Genes Dev.* **24,** 2219–2227 (2010).
15. Nagashima, T. *et al.* Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. *J. Biol. Chem.* **282,** 4045–4056 (2007).
16. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48,** 240–248 (2009).
17. Gomm, J. J. *et al.* Isolation of pure populations of epithelial and myoepithelial cells from the normal human mammary gland using immunomagnetic separation with Dynabeads. *Anal. Biochem.* **226,** 91–99 (1995).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
19. Stark, R. & Brown, G. D. DiffBind: differential binding analysis of ChIP-seq peak data. *Bioconductor* http://bioconductor.org/packages/release/bioc/html/DiffBind.html.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank D. Schmidt for assistance with figures, J. Hadfield for Illumina sequencing, S. MacArthur, O. Rueda, S. Vowler, R. Russell and M. Wilson for technical and bioinformatics help. We thank J. Stingl and his laboratory for help with the normal mammary gland work. We would like to acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. The authors would like to thank Imperial College Healthcare NHS Trust, Human Biomaterials Resource Centre (Tissue Bank). Tumour samples from Cambridge were obtained with support from NIHR Biomedical Research Centre and the Experimental Cancer Medicine Centre. C.S.R.-I. is supported by a Commonwealth Scholarship. O.G. is part funded by a grant awarded by the Ministry of Education of the Czech Republic (Project "Oncology" MSM 0021620808) and is also a recipient of a Translational Research Fellowship from the European Society of Medical Oncology. C.P. is funded by Cancer Research UK. J.S.C. is supported by an ERC starting grant and an EMBO Young investigator award.

Author Contributions C.S.R.-I., R.S., C.C. and J.S.C. designed all experiments. Experimental work was conducted by C.S.R.-I. with help from K.A.H. Computational analysis was conducted by R.S. and A.E.T., with help from M.J.D. and G.D.B. All clinical samples, clinical information and help with sample processing was provided by C.C., C.P., S.-F.C., S.A., A.R.G., I.O.E. and O.G. Histological analysis was conducted by H.R.A. The manuscript was written by C.S.R.-I., R.S., C.C. and J.S.C. with assistance from other authors.

Author Information Alignment and peak data are deposited in the Gene Expression Omnibus under accession number GSE32222. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.S.C. (jason.carroll@cancer.org.uk) or C.C. (carlos.caldas@cancer.org.uk).

METHODS

Cell culture. MCF-7, ZR75-1, T-47D and BT-474 human cell lines were obtained from the American type culture collection (ATCC). MCF-7 cells were grown in DMEM containing 10% heat-inactivated FBS, 2 mM L-glutamine, 50 U ml^{-1} penicillin and 50 μg ml^{-1} streptomycin. ZR75-1, T-47D and BT-474 cells were grown in RPMI containing 10% heat-inactivated FBS, 2 mM L-glutamine, 50 U ml^{-1} penicillin and 50 μg ml^{-1} streptomycin. To validate the cell lines, samples were genotyped by the Health Protection Agency (http://www.hpa.org.uk) (MCF-7, ZR75-1, T-47D) or by in-house genotyping (BT-474). TAM-R cells[13] were obtained from I. Hutcheson and R. Nicholson (Cardiff) and were maintained in phenol-red-free DMEM containing 5% charcoal dextran-treated FBS, 2 mM L-glutamine, 50 U ml^{-1} penicillin and 50 μg ml^{-1} streptomycin and 10 nM 4-hydroxytamoxifen.

Primary tumour material. The ER$^+$ breast cancer tumours were obtained from the Nottingham Tenovus primary breast cancer series, Addenbrooke's Hospital and Imperial College Healthcare NHS Trust, with appropriate ethical approval from the repositories. The malignant pericardial effusion and the two distant metastases were obtained from Imperial College Healthcare NHS Trust. See Supplementary Fig. 2 for clinical details.

Chromatin immunoprecipitations. The antibodies used were anti-ER (sc-543) from Santa Cruz Biotechnologies and anti-FOXA1 (ab5089) from Abcam. For ChIP in the tumours and metastases, the frozen sample was cut into smaller pieces and thawed in 1% (final concentration) formaldehyde for 20 min at room temperature (20 °C). The reaction was quenched by adding 0.1 volume of 2 M glycine for 10 min. The sample was disaggregated by Dounce homogenization and processed according to standard ChIP procedures[16]. The DNA was subsequently amplified as previously described[16]. For the malignant pericardial effusion, epithelial cells were first enriched using Dynabeads conjugated with Epcam[17]. For ChIPs from cell-line material, proliferating cells were cross-linked and processed for ChIP as previously described[16]. For the TAM-R cells, ER ChIP-seq was performed on cells grown in DMEM containing 10% FBS and 10 nM tamoxifen for 24 h. For the cocktail experiments, cells were treated with 100 ng ml^{-1} IGF-1, 100 ng ml^{-1} EGF, 1 ng ml^{-1} TNF-α and 10 ng ml^{-1} IL-6 for 90 min.

High-throughput sequencing and enrichment analysis. Sequences generated by the Illumina Genome Analyzer were processed by the Illumina analysis pipeline version 1.6.1, and aligned to the Human Reference Genome (assembly hg18, NCBI build 36.1, March 2008) using BWA version 0.5.5 (ref. 18). Reads were filtered by removing those with a BWA alignment quality score less than 15. Enriched regions of the genome were identified by comparing the ChIP samples to input samples using the MACS peak caller version 1.3.7.1 (ref. 10). Additional peak calls were determined for the tumour samples using the SWEMBL peak caller version 3.2 (S. Wilder *et al.*, manuscript in preparation) with default parameters except with $-R = 0.005$. For tumours without a corresponding input, all the available tumour input reads were combined and down sampled to derive a control track for peak calling purposes.

Differential binding analysis. After identifying a common set of peaks (see Supplementary Figs 4A and 15), significantly differentially bound sites were identified by first counting the number of reads in the ChIP samples overlapping each identified peak (subtracting the number of overlapping reads from the corresponding input sample). Differential analysis was performed using the edgeR package[20]. Normalization factors were computed using the TMM technique, after which tagwise dispersions were calculated and subjected to an exact test[20]. Resulting *P* values were subjected to Benjamini–Hochberg multiple testing correction to derive FDRs; only sites differentially bound with a FDR < 0.1 (cell line samples) or a *P* value <0.01 were considered for further analysis. The differential binding analysis and related plots were performed using the DiffBind package[19].

Heat-map generation. To generate the heat maps of the raw ChIP-seq data, ER-binding peaks in the different samples were used as targets to centre each window. Each window was divided into 100 bins of 100 bp in size. An enrichment value was assigned to each bin, counting the number of sequencing reads in that bin normalized to 10,000,000 sequencing reads, after subtracting the number of normalized input reads per bin.

Motif analysis. For analysis of enriched DNA motifs, Weeder was used, with the default settings[21].

Distant-metastasis-free-survival analysis. First, we computed for each gene in the given ER-binding list (ER-core-binding events, good outcome ER binding, poor/met ER binding) a *P* value and statistic from a univariate Cox regression, individually in nine independent data sets[12,22–29]. This was done to ascertain if the observed directionalities of association between gene expression and outcome were consistent with the predictions from the differential ER-binding events. Histograms of *P* values over all cohorts clearly confirmed the association of the ER-binding events with distant-metastasis-free survival (DMFS; Supplementary Fig. 9). Moreover, genes more bound in poor prognosis and metastatic tumours were generally more highly expressed in poor prognosis tumours (Supplementary Fig. 10).

We adopted a two-step strategy to first learn from the pool of genes in a given ER-binding list, a subset with patterns of gene expression that are associated with DMFS in a training set, and second to then validate this predictor in completely independent data. We used the ER$^+$ breast cancer samples from two independent data sets as training sets[22,23], re-normalized the gene expression profiles to mean zero and unit variance and applied a singular value decomposition to the resulting reduced expression matrices. For each of the top 10 eigenvectors (principal components) we correlated their profiles across samples to DMFS using a Cox regression, and next identified DMFS-associated eigenvectors, which were also highly correlated (or anticorrelated) between the two training sets. Thus, our algorithm is similar to the supervised principal components algorithm (SPCA) previously described[30], but differs in that we extend this procedure to two training sets. We consistently identified DMFS-associated eigenvectors that were also highly correlated between the two training sets. A prognostic predictor was constructed by averaging the DMFS-associated eigenvectors from each training set. The predictor was subsequently tested in seven independent ER$^+$ breast cancer cohorts[12,24–29]. Specifically, for each sample we computed the Pearson correlation of the predictor with the sample expression profile over the genes that could be mapped. The resulting correlation can be viewed as the risk index. In each cohort this risk index was then evaluated in a Cox regression to test ability to predict DMFS. The data shown in Fig. 1e and Fig. 2d is derived from the Loi data set[12], but the additional data sets are provided in Supplementary Fig. 6 and 12.

The Sweave document describing all analyses can be found at http://c3315341.r41.cf0.rackcdn.com/Data.tar.gz.

Immunohistochemistry. Immunohistochemistry of full-face sections was performed using a BondMax Autoimmunostainer (Leica Microsystems). ER staining was conducted using the 6F11/2 mouse monoclonal antibody (Novocastra, Leica Microsystems) at 1 in 70 and antigen retrieval was performed by heating with pH 6 citrate buffer for 30 min. FOXA1 staining was conducted using a rabbit polyclonal antibody (ab23738) from Abcam, at 1 in 800 with antigen retrieval performed by heating with pH 6 citrate buffer for 20 min. Bound primary antibody was detected using a BOND polymer detection kit (Leica) and developed with 3-3′-diaminobenzidine (DAB). The degree of staining was assessed by a pathologist (H.R.A.) using a standard light microscope. An Allred scoring system was used to assess staining accounting for both staining intensity (0 = none, 1 = weak, 2 = moderate, 3 = strong) and the proportion of cells stained (0 = 0%, 1 < 1%, 2 = 1–10%, 3 = 11–33%, 4 = 34–66%, 5 > 66%), providing a composite score (intensity + proportion = 0–8).

20. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
21. Pavesi, G. *et al.* MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.* **34,** W566–W570 (2006).
22. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365,** 671–679 (2005).
23. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347,** 1999–2009 (2002).
24. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98,** 262–272 (2006).
25. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10,** 529–541 (2006).
26. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68,** 5405–5413 (2008).
27. Buffa, F. M. *et al.* microRNA associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res.* **71,** 5635 (2011).
28. Naderi, A. *et al.* A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* **26,** 1507–1516 (2007).
29. Hoadley, K. A. *et al.* EGFR associated expression profiles vary with breast tumor subtype. *BMC Genomics* **8,** 258 (2007).
30. Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* **2,** e108 (2004).

LETTER

doi:10.1038/nature10624

Natural killer cells act as rheostats modulating antiviral T cells

Stephen N. Waggoner[1], Markus Cornberg[2], Liisa K. Selin[1] & Raymond M. Welsh[1]

Antiviral T cells are thought to regulate whether hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infections result in viral control, asymptomatic persistence or severe disease, although the reasons for these different outcomes remain unclear. Recent genetic evidence, however, has indicated a correlation between certain natural killer (NK)-cell receptors and progression of both HIV and HCV infection[1–3], implying that NK cells have a role in these T-cell-associated diseases. Although direct NK-cell-mediated lysis of virus-infected cells may contribute to antiviral defence during some virus infections—especially murine cytomegalovirus (MCMV) infections in mice and perhaps HIV in humans[4,5]—NK cells have also been suspected of having immunoregulatory functions. For instance, NK cells may indirectly regulate T-cell responses by lysing MCMV-infected antigen-presenting cells[6,7]. In contrast to MCMV, lymphocytic choriomeningitis virus (LCMV) infection in mice seems to be resistant to any direct antiviral effects of NK cells[5,8]. Here we examine the roles of NK cells in regulating T-cell-dependent viral persistence and immunopathology in mice infected with LCMV, an established model for HIV and HCV infections in humans. We describe a three-way interaction, whereby activated NK cells cytolytically eliminate activated CD4 T cells that affect CD8 T-cell function and exhaustion. At high virus doses, NK cells prevented fatal pathology while enabling T-cell exhaustion and viral persistence, but at medium doses NK cells paradoxically facilitated lethal T-cell-mediated pathology. Thus, NK cells can act as rheostats, regulating CD4 T-cell-mediated support for the antiviral CD8 T cells that control viral pathogenesis and persistence.

Intravenous (i.v.) inoculation of C57BL/6 mice with a low (5×10^4 plaque-forming units (p.f.u.), medium (2×10^5 p.f.u.), or high (2×10^6 p.f.u.) dose of LCMV, strain clone 13, resulted in different degrees of pathology, as indicated by weight loss (Fig. 1a) and by histological analysis of lung sections at day 15 post-infection (p.i.) (Fig. 1b). The high dose caused a precipitous drop in body weight during the first week of infection (Fig. 1a, right) but, thereafter, clonal exhaustion and deletion of LCMV-specific T cells resulted in a persistent infection[9,10] associated with minimal lung pathology (Fig. 1b, right) and 100% (77 of 77) survival (Fig. 1c, top). Selective depletion of NK cells using 25 μg of anti-NK1.1 monoclonal antibodies (Supplementary Fig. 1) 1 day before high-dose infection resulted in 58% (35 of 65) mortality between days 9 and 13 of infection (Fig. 1c, top) associated with severe pulmonary oedema (data not shown) and reduced viral titres by day 7 p.i. (Fig. 1d, right). Under these high-dose conditions, therefore, the presence of NK cells promoted persistence and prevented mortality.

In contrast to the beneficial role of NK cells during high-dose infection, NK-cell depletion prevented the severe weight loss (Fig. 1a, middle) and tissue pathology (Fig. 1b, middle) associated with the medium dose of LCMV. Twenty-three per cent (7 of 31) of control-treated mice succumbed to the medium dose during the second week of infection, and the lungs of surviving mice exhibited bronchus-associated lymphoid tissue, pulmonary oedema and interstitial mononuclear infiltration. Lung pathology was absent in NK-cell-depleted mice, which uniformly survived medium-dose challenge (Fig. 1c, bottom). Moreover, although high levels of replicating virus persisted in surviving control mice at day 15 p.i., NK-cell depletion resulted in complete viral clearance (Fig. 1d, middle). In this case the presence of NK cells was detrimental for the host, as they promoted immune pathology and death.

Irrespective of the presence of NK cells, inoculation with a low dose of virus was uniformly non-lethal in 18 of 18 (100%) control and 18 of 18 (100%) of NK-cell-depleted mice by >50 days p.i., with minimal weight loss (Fig. 1a, left) and minimal lung pathology (Fig. 1b, left). Virus was completely cleared in both groups of mice by day 15 of low-dose infection (data not shown), but NK-cell depletion resulted in more rapid elimination of LCMV in liver by day 7 p.i. (Fig. 1d, left).

The weight loss, lung pathology and mortality observed in medium-dose-infected wild-type mice (Fig. 1a, b) did not occur after infection of αβ T-cell-receptor-deficient ($Tcrb^{-/-}$)mice, and NK-cell depletion of $Tcrb^{-/-}$ mice did not alter weight loss or viral burden (Supplementary Fig. 2). Thus, NK cells regulate viral clearance and immunopathology during LCMV infection through a T-cell-dependent mechanism.

As early as day 6 after medium-dose infection, the proportion and number of interferon-γ (IFN-γ)$^+$ LCMV-specific CD8 T cells was increased two- to sixfold in mice depleted of NK cells (Fig. 2a and Supplementary Fig. 3), and antiviral T cells from these mice showed an enhanced ability to co-produce tumour necrosis factor (TNF) (Supplementary Fig. 3). The number of LCMV epitope $NP_{396–404}$ tetramer-binding CD8 T cells in the spleen on day 5 p.i. was increased 4- to 20-fold in NK-cell-depleted mice relative to non-depleted control mice after infection with all doses of virus (Fig. 2b). The number of virus-specific IFN-γ$^+$ CD4 T cells was also amplified 7- to 20-fold by NK-cell depletion compared to control mice on different days after medium-dose infection (Fig. 2c). Moreover, co-production of TNF and interleukin-2 (IL-2) by antiviral CD4 T cells was augmented by NK-cell depletion (Fig. 2d and Supplementary Fig. 3). The increased magnitude of the LCMV-specific T-cell response in the absence of NK cells during medium-dose infection correlated with rapid viral clearance (Fig. 2e). Depletion of NK cells using a carefully titrated dose of anti-asialo GM1 antibody, which eliminates NK cells but not CD8 T cells[11], also enhanced antiviral CD4 and CD8 T-cell responses during medium-dose infection (Supplementary Fig. 4).

The enhanced antiviral T-cell responses suggested that NK-cell depletion may augment proliferation of LCMV-specific T cells. Transfer of carboxyfluorescein diacetate succinimidyl ester (CFSE)-labelled Thy1.1$^+$ T cells revealed a larger population of CFSE^{low} donor CD4 (Fig. 2f) and CD8 (data not shown) T cells in multiple host tissues at day 6 p.i. of high-dose infection in the absence of NK cells. There was also greater specific lysis of viral-peptide-coated target cells as detected by a conventional *in vivo* cytotoxicity assay at day 4 of infection (Supplementary Fig. 5). Moreover, LCMV-specific Ly5.1$^+$ TCR transgenic (P14) CD8 T cells (transfer 10^4) were recovered from tissues of NK-cell-depleted recipient (Ly5.2$^+$) mice at two- to ninefold greater

[1]Department of Pathology and Program in Immunology and Virology, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA. [2]Department of Gastroenterology and Hepatology, Medizinische Hochschule, Hannover 30625, Germany.

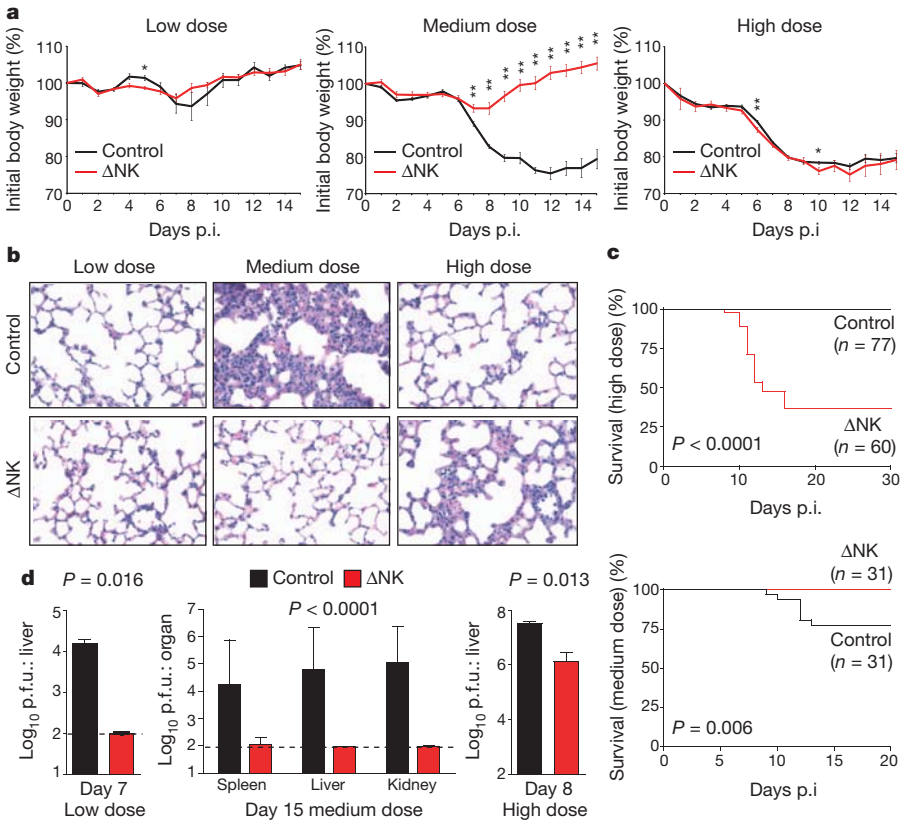


Figure 1 | NK cells influence T-cell-dependent pathology and viral persistence during LCMV infection. **a**–**d**, C57BL/6 mice treated with IgG2a (Control) or anti-NK1.1 (ΔNK) were infected with low (5×10^4 p.f.u.), medium (2×10^5 p.f.u.), or high (2×10^6 p.f.u.) doses of LCMV. **a**, Weight loss (mean ± s.e.m.) during infection ($n = 3$–43 per group per day), $^*P < 0.05$, $^{**}P < 0.01$. **b**, Haematoxylin & eosin (H&E) staining of lung (×400) at day 15 p.i. **c**, Survival after high- or medium-dose infection. **d**, Viral titres (mean ± s.e.m.) after low (day 7, $n = 3$ per group), medium (day 15, $n = 9$–15 per group), or high (day 8, $n = 3$ per group) dose infection. Dotted line represents limit of detection.

numbers than control-treated mice 6 days after low-dose infection (Supplementary Fig. 5). Together these results indicate that NK1.1$^+$ cells repress the size of the antiviral T-cell response during LCMV infection.

The activities of CD4 T cells are important for maintaining CD8 T-cell function during LCMV infection[12–14]. To assess whether CD4 T cells were involved in the NK-cell suppression of LCMV-specific CD8 T cells, mice were treated with antibodies to concurrently deplete both NK and CD4 T cells. Whereas depletion of NK cells before medium-dose LCMV infection resulted in a >200-fold reduction in splenic viral titres at day 7 p.i. relative to control and CD4-depleted (ΔCD4) mice (Fig. 3a), depletion of both NK and CD4 T cells (ΔNKΔCD4) had no effect on viral titres. The increase in the number of antiviral CD8 T cells producing more than one cytokine caused by NK-cell depletion was also prevented by co-depletion of CD4 T cells (Fig. 3b). In contrast, co-depletion of NK and CD8 T cells did not prevent an increase in IFN-γ^+ LCMV epitope $GP_{61–80}$-specific CD4 T cells (control: 3.8 ± 0.5% versus ΔNK: 9.8 ± 0.7% versus ΔNK/ΔCD8: 10.7 ± 1.6%, $n = 3$, $P < 0.05$ versus control) at day 12 of medium-dose infection.

Paradoxically, at the high virus dose, co-depletion of NK and CD4 T cells prevented the severe pulmonary oedema (Fig. 3c) and increased mortality (Fig. 3d) associated with depletion of NK cells alone. In this experiment, mice were harvested on day 12 p.i., when three surviving NK-cell-depleted mice were moribund and required euthanasia, whereas all double-depleted mice showed relatively normal vigour. The livers of NK/CD4 double-depleted mice contained 25-fold more p.f.u. than livers from mice depleted of NK cells alone (NK: 5.7 ± 0.2 p.f.u. versus ΔNK/ΔCD4: 7.1 ± 0.1 p.f.u., $n = 5$, $P < 0.0001$). Enhancement of LCMV-specific CD8 T cells in the absence of NK cells was also abrogated by concurrent depletion of CD4 T cells (Supplementary Fig. 6). Together these data indicate that CD4 T cells are needed for NK-cell modulation of antiviral CD8 T-cell responses associated with viral clearance, persistence and immunopathology.

We used a modified *in vivo* cytotoxicity assay by injecting splenocytes from medium-dose-infected NK-cell-depleted mice (Ly5.1$^+$, day 4 p.i.) into medium-dose-infected NK-cell-depleted (ΔNK) or isotype IgG2a-treated (control) recipient mice (Ly5.2$^+$, day 3 p.i.). After 5 h, similar proportions of total donor T (control: 0.16 ± 0.03% versus ΔNK: 0.15 ± 0.02%, $n = 21$, $P = 0.80$) and B cells (control: 1.8 ± 0.2% versus ΔNK: 1.7 ± 0.2%, $n = 21$, $P = 0.88$) were recovered from infected recipients, regardless of NK-cell depletion. Likewise, recovery of activated (CD44^{hi} CD43(1B11)$^+$) donor CD8 T cells was similar from spleens of control and ΔNK mice, with minimal loss relative to uninfected control mice (Fig. 4a). In contrast, there was a substantial loss of activated donor CD4 T cells in infected relative to uninfected recipients, and this loss was prevented by depletion of NK cells (Fig. 4a). The magnitude of NK-cell-dependent loss of activated donor CD4 T cells was similar in low-, medium- and high-dose-infected recipients (Fig. 4b). More activated CD4 T cells, both donor and host derived, in infected (control) mice stained positively for the apoptosis indicator annexin V in comparison to naive donor CD4 T cells or to activated donor CD4 T cells in medium-dose-infected ΔNK recipient mice (Fig. 4c). In contrast to activated donor CD4 T cells, the recoveries of naive (CD44^{low}) phenotype CD4 and CD8 donor T cells were not altered by NK-cell depletion (data not shown). These data indicate that NK cells in wild-type mice selectively and rapidly target activated CD4 T cells for elimination during LCMV infection.

We next examined the involvement of NK-cell cytolytic mediators FasL, TNF and perforin (Prf1) in this process. The loss of activated wild-type donor CD4 T cells in infected wild-type recipient mice (Fig. 4a, d) was seen when activated *Fas*^{lpr} (Fas mutant) mouse donor cells were transferred into wild-type recipient mice or when wild-type donor cells were transferred into *Tnf*$^{-/-}$ recipient mice (Fig. 4d). In contrast, there was relatively little loss of activated wild-type donor CD4 T cells in *Prf1*$^{-/-}$ hosts (Fig. 4d), whose retention of activated donor CD4 cells was not significantly different ($P > 0.1$) from that in NK-cell-depleted wild-type or *Prf1*$^{-/-}$ hosts. Thus, NK-cell elimination of activated CD4 T cells is mediated through a perforin-dependent pathway that does not require Fas or TNF.

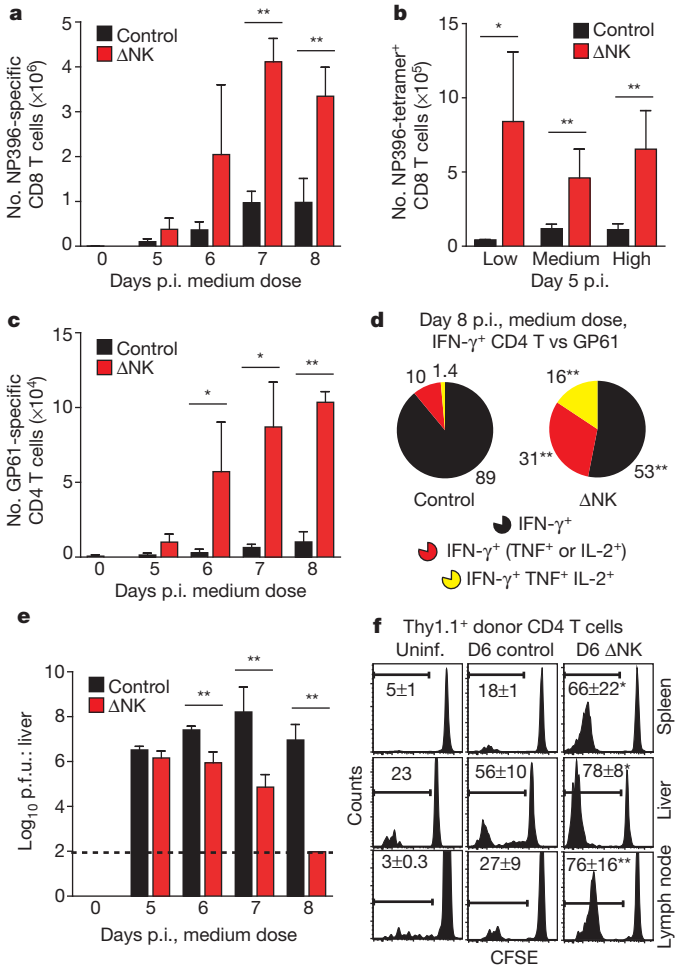


Figure 2 | LCMV-specific T-cell responses enhanced in NK-cell-depleted mice. **a**–**c**, Number (mean ± s.e.m.) of LCMV-specific T cells measured by IFN-γ expression (**a**, **c**) or tetramer-binding (**b**) at various days p.i. (**a**, **c**) (medium dose, $n = 3$ per group per day) or at day 5 p.i. (**b**) (various doses, $n = 3$–11 per group). **d**, Co-production of TNF and IL-2 by gated IFN-γ$^+$ CD4 T cells (day 8 p.i., medium dose) after *in vitro* stimulation with GP61 peptide. **e**, Viral titres in liver ($n = 3$–11 per group per day, medium dose). **f**, CFSE dilution by donor (Thy1.1$^+$) CD4 T cells in uninfected (Uninf.) and infected (medium dose, day 6 p.i.) control and ΔNK Thy1.2$^+$ host mice. Control versus ΔNK mice, $^*P < 0.05$, $^{**}P < 0.01$.

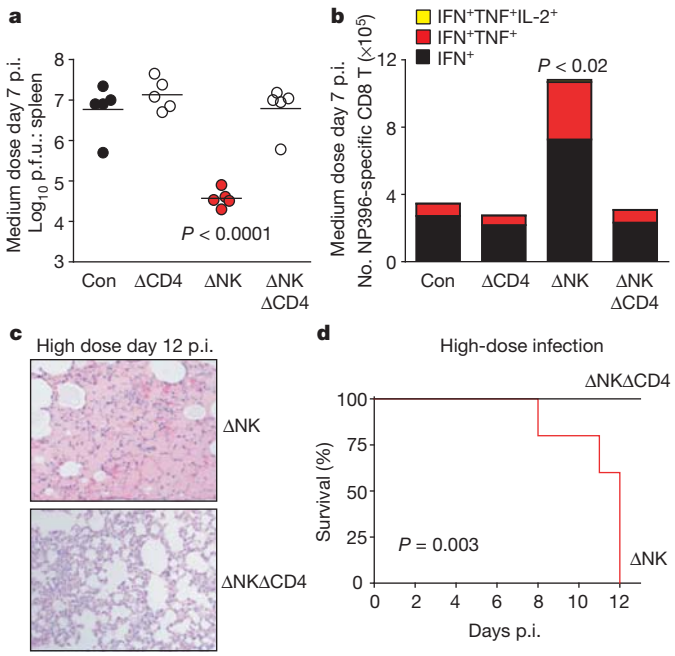


Figure 3 | Role of CD4 T cells in NK suppression of antiviral CD8 T cells, viral control and immunopathology. **a**–**e**, Prior to medium- (**a**, **b**) or high-dose (**c**–**e**) infection, mice were injected with isotype (Con, control), anti-NK1.1 (ΔNK), anti-CD4 (ΔCD4), or both (ΔNKΔCD4). **a**, **b**, Viral titres in spleen (**a**) and number of cytokine-producing $NP_{396–404}$-specific CD8 T cells ($n = 5$ per group, ± s.e.m.) in spleen at day 7 of medium-dose infection (**b**). **c**, **d**, H&E staining of lung (×100) (**c**) and survival at day 12 of high-dose infection ($n = 3$–5 per group) (**d**).

Previous work has implicated NKG2D (also known as Klrk1) in targeting of activated T cells by murine NK cells *in vitro*[15–17], but we observed no differences in activated wild-type donor CD4 T-cell survival in wild-type versus *Nkg2d*$^{-/-}$ recipients (Fig. 4d) or in wild-type mice treated with a blocking monoclonal antibody to NKG2D[18] (data not shown). Of note is that we did not observe the expression of ligands for activating NK-cell receptors including NKG2D, NKp46 (also known as NCR1), DNAM-1 (also known as CD226) and TRAIL (also known as TNFSF10) on these early activated CD4 T cells, and NK-cell-mediated elimination of activated donor CD4 T cells also occurred in antibody-deficient (μMT$^{-/-}$) mice (data not shown), precluding a role for antibody-dependent mechanisms. Activated CD4 T cells did, however, express much higher levels of adhesion molecules than naive cells, and these molecules have previously been shown to trigger NK-cell cytotoxicity via LFA-1[19,20]. Somewhat surprising was the observation that the activated CD4 T cells were far more susceptible than activated CD8 T cells to direct killing by the NK cells, even though both expressed high levels of adhesion molecules. We previously had shown that the presence of the negatively signalling receptor CD244 (2B4) on NK cells prevented NK-cell-mediated lysis of activated CD8 T cells[21]. We found here that although expression of the CD244 ligand CD48 was upregulated on T cells after medium-dose LCMV infection, expression levels of CD48 were much higher on activated CD8 than on activated CD4 cells (mean fluorescence intensity (MFI) of CD48: activated CD4, 3,423 ± 147; activated CD8, 6,180 ± 166; $n = 9$, $P < 0.0001$) (Supplementary Fig. 7).

To assess whether NK-cell-mediated lysis of activated CD4 T cells is a general principle of virus infections, we examined the loss of LCMV-activated CD4 T cells after transfer into mice inoculated with an unrelated *Arenavirus*, Pichinde virus (PV), the *Coronavirus* mouse hepatitis virus (MHV), or the interferon inducer and NK-cell activator polyinosinic:polycytidylic acid (polyI:C). All three stimuli induced measureable loss of activated donor CD4 T cells that was dependent upon the presence of NK cells (Fig. 4e). In reciprocal experiments, CD4 T cells activated during infection with PV, MHV, vaccinia virus (VV) or MCMV were lost upon transfer into mice infected with medium-dose LCMV when NK cells were present (Fig. 4f).

An analysis, by *in vivo* cytotoxicity assays after transfer into NK-cell-sufficient mice, of the window of time during which NK-cell regulation of T cells occurred in the LCMV medium-dose model showed reduced frequencies of activated donor CD4 T cells relative to uninfected recipients 1 (43%), 2 (32%), 3 (26%), 4 (4%) and 5 (8%) days after infection (Supplementary Fig. 8). The frequencies of activated donor CD4 T cells were increased by NK-cell depletion in recipient mice only at day 2 and day 3 p.i. These results indicate that NK cells target activated CD4 T cells mainly on the second and third day of infection, when the cytolytic activity of NK cells is at its peak[22].

These results show that NK cells can have a crucial role in controlling virus-associated morbidity, mortality and persistence in the absence of direct NK-cell-mediated control of virus replication, and they do so by altering the numbers and polyfunctionality of virus-specific T cells. Their effect on activated CD4 T cells was presumably due to direct cytotoxicity, as demonstrated by rapid perforin-dependent elimination of activated CD4 T cells by NK cells in short-term *in vivo* cytotoxicity assays and the observation of enhanced annexin reactivity

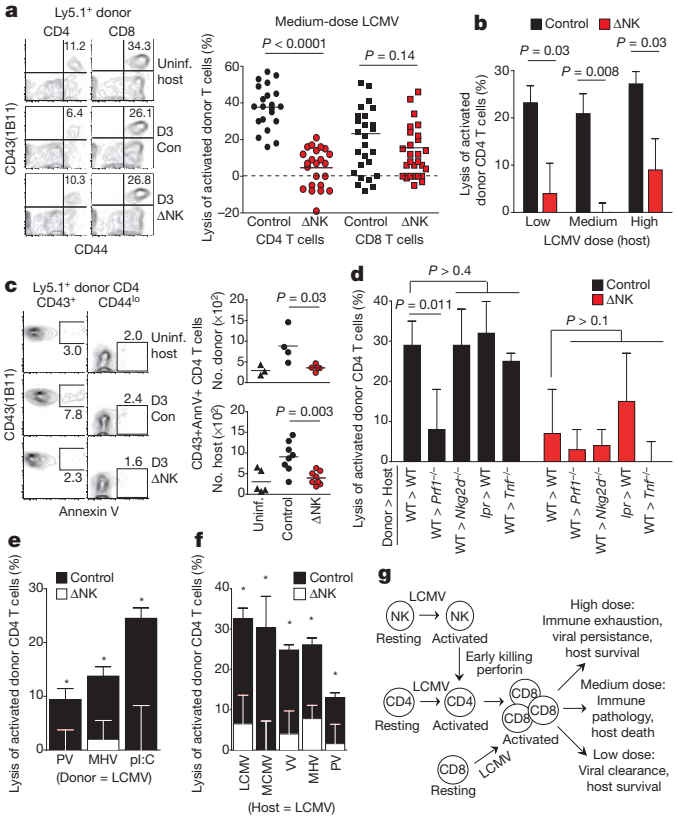


Figure 4 | NK cells rapidly eliminate activated CD4 T cells. **a**–**f**, *In vivo* cytotoxicity assays were performed as described in Methods using donor cells from NK-cell-depleted wild-type (Ly5.1$^+$) (**a**–**f**) or *Fas*^{lpr} (*lpr*) (**d**) mice 4 days p.i. with medium-dose LCMV (**a**–**e**) or other viruses (**f**). Recovery of donor T cells was examined 5 h after transfer into control or NK-depleted (ΔNK) Ly5.2$^+$ mice (**a**, $n = 21$–28 per group; **b**–**f**, $n = 3$–5 per group). Host mice were uninfected or inoculated with medium-dose LCMV (day 3 p.i.) (**a**, **c**, **d**, **f**), various doses of LCMV (day 3 p.i.) (**b**), or PV (day 2), MHV (day 3) and polyI:C (pI:C) (day 1) (**e**). **a**, **b**, **d**–**f**, Lysis (mean ± s.e.m.) of donor T cells in infected relative to uninfected recipient mice; control versus ΔNK mice, $^*P < 0.05$. WT, wild type. **c**, Annexin V reactivity of donor (Ly5.1$^+$) and host (Ly5.1$^-$) CD4 T cells. **g**, Proposed model connecting NK-cell killing of CD4 T cells to CD8 T cells and infection outcome in the presence of NK cells.

of CD4 T cells in the presence of NK cells. The effect of activated NK cells on CD8 T cells and virus clearance seemed to be indirect, and depended on the presence of CD4 T cells, which are known to produce factors that preserve CD8 T-cell viability and functionality[12,13,23–25]. This suggests a three-way interaction, whereby NK cells suppress the CD4 T-cell response, thereby preventing augmentation of the CD8 T-cell response, which, in turn, directly regulates viral clearance and immunopathology in this system (Fig. 4f).

Our observation of direct NK-cell-mediated lysis of T cells during virus infection is distinct from published accounts of NK-cell regulation of antiviral T cells during MCMV infection, in which NK-cell-mediated lysis of virus-infected cells contributes to control of viral burden and persistence of MCMV-infected dendritic cells that in turn regulate activity of antiviral T cells[6,7,26]. Moreover, we found no NK-cell-dependent changes in the number and antigen-presenting function of splenic dendritic cells during LCMV infection (Supplementary Fig. 9), consistent with our finding that NK cells directly regulated the T cells. A possible concern was that *in vivo* cellular depletion with antibodies against NK1.1 and CD4 have the potential to affect frequencies of NKT, γδ T and regulatory T cells, but the frequencies of these lymphocytes in the spleen of LCMV-infected mice were not altered after 3 to 4 days of infection by the low concentration of anti-NK1.1 used in these studies (Supplementary Figs 1 and 10). Moreover, depletion of NK cells in γδ T-cell-deficient (*Tcrd*$^{-/-}$) and NK T-cell-deficient (*Cd1d*$^{-/-}$) mice during medium-dose infection enhanced LCMV-specific T-cell responses and reduced viral loads (Supplementary Fig. 9), similar to that in wild-type mice. Therefore, these lymphocyte lineages seem to be dispensable for NK-cell immunoregulatory function during LCMV infection.

By adjusting the dose of the LCMV inocula, it is possible to generate diverse patterns of CD8 T-cell-regulated pathogenesis, similar to the variety of pathogenic patterns a human HCV infection can take, including rapid viral clearance, severe T-cell-dependent immunopathology and long-term persistence. We show here that at a high dose of LCMV, NK cells act beneficially by suppressing T-cell responses, thereby preventing severe pathology and mortality while enabling the development of a persistent infection from which mice eventually recover and clear the virus[27]. At the medium-dose inoculum, NK-cell suppression of T cells is detrimental to the host, as virus clearance is impaired due to the limited number and functionality of T cells. However, a medium dose of virus is not sufficient for complete clonal exhaustion of T cells, ultimately resulting in severe T-cell-dependent immunopathology that can lead to death of the host.

These results indicate that NK cells can serve as rheostats, or master regulators, of antiviral T-cell responses. Consistent with the fact that many virus infections induce cytokines that potently activate NK cells[28], we found that NK-cell lysis of activated CD4 T cells was triggered by several viruses as well as after inoculation with polyI:C, which induces interferon and activates NK cells. Although a previous study found that NK-cell depletion did not alter the magnitude of antiviral T-cell responses during infection with the Armstrong strain of LCMV[29], we have observed enhanced antiviral T-cell responses and improved viral control at early time points after infection of NK-cell-depleted mice with both LCMV Armstrong and Pichinde virus (S.N.W., unpublished observations). Thus, the timing and the type of evaluation may be important to detect detrimental effects of NK cells on T cells during more benign viral infections.

METHODS SUMMARY

Infection model. One day before infection, male C57BL/6 mice were selectively depleted of NK cells through a single intraperitoneal (i.p.) injection of 25 μg anti-NK1.1 monoclonal antibody (PK136) or a control mouse IgG2a (both from Bio-X-Cell), as previously described[21] (Supplementary Fig. 1). In some cases, mice were also depleted of CD4 T cells by i.p. injection of 100 μg anti-CD4 (GK1.5) at days −1 and +3 of infection. Mice were then infected i.v. with 5×10^4 (low dose), 2×10^5 (medium dose) or 2×10^6 (high dose) p.f.u. of the clone 13 variant of LCMV. Virus was titrated by plaque assay on Vero cells. In some experiments, mice were inoculated i.p. with 1.5×10^7 p.f.u. of PV, 8×10^5 p.f.u. of MHV strain A59, 1×10^6 p.f.u. of VV strain Western Reserve, 1×10^6 p.f.u. of Smith strain MCMV, or 200 μg of polyI:C (Invivogen).

Immune assays. The number of LCMV-specific T cells was measured by H-2D^b-NP$_{396–404}$ tetramer staining or by intracellular cytokine staining after 5 h *ex vivo* stimulation with 1 μM viral peptide in the presence of brefeldin A. T-cell cytolytic activity was measured *in vivo* as described previously[21].

***In vivo* NK-cell assay.** An unconventional *in vivo* cytotoxicity assay was previously established to determine NK-cell killing of lymphocyte populations[21]. Donor mice were depleted of NK cells and then infected i.v. or i.p. with different viruses. At day 4 p.i., single-cell splenocyte suspensions were prepared from these mice, labelled with 2 μM CFSE, and transferred (2×10^7) into various strains of NK-cell-depleted or control recipient mice that were either uninfected or had been infected with virus 1–5 days previously. Spleens of recipient mice were harvested 5 h after transfer and assessed for survival of donor T cells.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 April; accepted 12 October 2011.
Published online 20 November 2011.

1. Martin, M. P. *et al.* Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nature Genet.* **31,** 429–434 (2002).
2. Jennes, W. *et al.* Cutting edge: resistance to HIV-1 infection among African female sex workers is associated with inhibitory KIR in the absence of their HLA ligands. *J. Immunol.* **177,** 6588–6592 (2006).

3. Khakoo, S. I. *et al.* HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science* **305,** 872–874 (2004).
4. Alter, G. *et al.* HIV-1 adaptation to NK-cell-mediated immune pressure. *Nature* **476,** 96–100 (2011).
5. Bukowski, J. F., Woda, B. A., Habu, S., Okumura, K. & Welsh, R. M. Natural killer cell depletion enhances virus synthesis and virus-induced hepatitis *in vivo*. *J. Immunol.* **131,** 1531–1538 (1983).
6. Robbins, S. H. *et al.* Natural killer cells promote early CD8 T cell responses against cytomegalovirus. *PLoS Pathog.* **3,** e123 (2007).
7. Andrews, D. M. *et al.* Innate immunity defines the capacity of antiviral T cells to limit persistent infection. *J. Exp. Med.* **207,** 1333–1343 (2010).
8. Welsh, R. M., Brubaker, J. O., Vargas-Cortes, M. & O'Donnell, C. L. Natural killer (NK) cell response to virus infections in mice with severe combined immunodeficiency. The stimulation of NK cells and the NK cell-dependent control of virus infections occur independently of T and B cell function. *J. Exp. Med.* **173,** 1053–1063 (1991).
9. Ahmed, R., Salmi, A., Butler, L. D., Chiller, J. M. & Oldstone, M. B. Selection of genetic variants of lymphocytic choriomeningitis virus in spleens of persistently infected mice. Role in suppression of cytotoxic T lymphocyte response and viral persistence. *J. Exp. Med.* **160,** 521–540 (1984).
10. Zajac, A. J. *et al.* Viral immune evasion due to persistence of activated T cells without effector function. *J. Exp. Med.* **188,** 2205–2213 (1998).
11. Yang, H., Yogeeswaran, G., Bukowski, J. F. & Welsh, R. M. Expression of asialo GM1 and other antigens and glycolipids on natural killer cells and spleen leukocytes in virus-infected mice. *Nat. Immun. Cell Growth Regul.* **4,** 21–39 (1985).
12. Battegay, M. *et al.* Enhanced establishment of a virus carrier state in adult $CD4^+$ T-cell-deficient mice. *J. Virol.* **68,** 4700–4704 (1994).
13. Matloubian, M., Concepcion, R. J. & Ahmed, R. $CD4^+$ T cells are required to sustain $CD8^+$ cytotoxic T-cell responses during chronic viral infection. *J. Virol.* **68,** 8056–8063 (1994).
14. Fuller, M. J., Khanolkar, A., Tebo, A. E. & Zajac, A. J. Maintenance, loss, and resurgence of T cell responses during acute, protracted, and chronic viral infections. *J. Immunol.* **172,** 4204–4214 (2004).
15. Noval Rivas, M. *et al.* NK cell regulation of CD4 T cell-mediated graft-versus-host disease. *J. Immunol.* **184,** 6790–6798 (2010).
16. Rabinovich, B. A. *et al.* Activated, but not resting, T cells can be recognized and killed by syngeneic NK cells. *J. Immunol.* **170,** 3572–3576 (2003).
17. Soderquest, K. *et al.* Cutting edge: $CD8^+$ T cell priming in the absence of NK cells leads to enhanced memory responses. *J. Immunol.* **186,** 3304–3308 (2011).
18. Ogasawara, K. *et al.* NKG2D blockade prevents autoimmune diabetes in NOD mice. *Immunity* **20,** 757–767 (2004).
19. Naganuma, H. *et al.* Increased susceptibility of IFN-γ-treated neuroblastoma cells to lysis by lymphokine-activated killer cells: participation of ICAM-1 induction on target cells. *Int. J. Cancer* **47,** 527–532 (1991).
20. Barber, D. F., Faure, M. & Long, E. O. LFA-1 contributes an early signal for NK cell cytotoxicity. *J. Immunol.* **173,** 3653–3659 (2004).
21. Waggoner, S. N., Taniguchi, R. T., Mathew, P. A., Kumar, V. & Welsh, R. M. Absence of mouse 2B4 promotes NK cell-mediated killing of activated $CD8^+$ T cells, leading to prolonged viral persistence and altered pathogenesis. *J. Clin. Invest.* **120,** 1925–1938 (2010).
22. Welsh, R. M. Jr. Cytotoxic cells induced during lymphocytic choriomeningitis virus infection of mice. I. Characterization of natural killer cell induction. *J. Exp. Med.* **148,** 163–181 (1978).
23. Yi, J. S., Du, M. & Zajac, A. J. A vital role for interleukin-21 in the control of a chronic viral infection. *Science* **324,** 1572–1576 (2009).
24. Frohlich, A. *et al.* IL-21R on T cells is critical for sustained functionality and control of chronic viral infection. *Science* **324,** 1576–1580 (2009).
25. Elsaesser, H., Sauer, K. & Brooks, D. G. IL-21 is required to control chronic viral infection. *Science* **324,** 1569–1572 (2009).
26. Bukowski, J. F., Woda, B. A. & Welsh, R. M. Pathogenesis of murine cytomegalovirus infection in natural killer cell-depleted mice. *J. Virol.* **52,** 119–128 (1984).
27. Oldstone, M. B. Biology and pathogenesis of lymphocytic choriomeningitis virus infection. *Curr. Top. Microbiol. Immunol.* **263,** 83–117 (2002).
28. Welsh, R. M. Regulation of virus infections by natural killer cells. *Nat. Immun. Cell Growth Regul.* **5,** 169–199 (1986).
29. Su, H. C. *et al.* NK cell functions restrain T cell responses during viral infections. *Eur. J. Immunol.* **31,** 3048–3055 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Hearn, C. Baer, J. Suschak and P. Afriyie for technical support; K. Daniels and M. Seedhom for insightful discussions; R. Taniguchi and V. Kumar for sharing unpublished observations; and H. Ducharme for mouse husbandry. We thank L. Berg for NKT tetramer, L. Lanier for anti-NKG2D blocking antibody (CX5), B. Polić for *Nkg2d*$^{-/-}$ mice and M. Exley for *Cd1d*$^{-/-}$ mice. This work was supported by National Institutes of Health (NIH) training grant AI07349 (S.N.W.) and research grants AI-17672, AI-081675, CA34461 (R.M.W.), AI46578 (L.K.S.), a German Research Foundation fellowship CO310-2/1 (M.C.) and an institutional Diabetes Endocrinology Research Center (DERC) grant DK52530. The views expressed are those of the authors and do not necessarily express the views of the NIH.

Author Contributions S.N.W. designed the study, performed experiments, analysed data and wrote the manuscript; R.M.W. designed the study, analysed data and wrote the manuscript; M.C. and L.K.S. were involved in study design, discussed results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.M.W. (Raymond.Welsh@umassmed.edu).

METHODS

Mice. C57BL/6, Thy1.1$^+$, *Tcrd*$^{-/-}$, *Tcrb*$^{-/-}$, *Fas*^{lpr}, *Prf1*$^{-/-}$ and μMT$^{-/-}$ mice were purchased from The Jackson Laboratories. Ly5.1$^+$ mice were from Taconic Farms. *Nkg2d*$^{-/-}$ and *Cd1d*$^{-/-}$ mice on a C57BL/6 background were obtained from B. Polić[30] and M. Exley[31], respectively. Congenic (Ly5.1$^+$) TCR transgenic P14 (ref. 32) mice and *Tnf*$^{-/-}$ mice on a C57BL/6 background were bred at the University of Massachusetts Medical School (UMMS). Male mice at 6–16 weeks of age were routinely used in experiments. Mice were maintained under specific pathogen-free conditions, and experiments were performed in compliance with institutional guidelines as approved by the Institutional Animal Care and Use Committee of UMMS.

Virus infections and *in vivo* cell depletions. The clone 13 variant of LCMV was propagated in baby hamster kidney BHK21 cells[9] and titrated by plaque assay on Vero cells. Mice were infected i.v. with 5×10^4 (low dose), 2×10^5 (medium dose), or 2×10^6 (high dose) p.f.u. of LCMV. Selective depletion of NK cells was achieved through a single i.p. injection of 25 μg anti-NK1.1 monoclonal antibodies (PK136) or a control mouse IgG2a produced by Bio-X-Cell, as previously described[21] (Supplementary Fig. 1). Alternatively, mice received a carefully titrated dose of 10 μl of anti-asialo GM1 antibody (Wako Biochemicals) diluted in 200 μl PBS i.p. 1 day before virus infection. Anti-NKG2D monoclonal antibody (CX5) was a gift of L. Lanier, and 200 μg was injected i.p. at the time of infection. Mice were depleted of T cells by i.p. injection of either 100 μg anti-CD4 (GK1.5) or 50 μg anti-CD8 (2.43) produced by Bio-X-Cell at day −1 and day +3 of infection. In some experiments, mice were inoculated i.p. with 1.5×10^7 p.f.u. of PV, 8×10^5 p.f.u. of MHV strain A59, 1×10^6 p.f.u. of VV strain Western Reserve, 1×10^6 p.f.u. of Smith strain MCMV, or 200 μg of polyI:C (Invivogen).

Tetramers and peptides. T-cell epitopes encoded by LCMV include $NP_{396–404}$ (FQPQNGQFI), $GP_{33–41}$ (KAVYNFATC) and $GP_{61–80}$ (GLKGPDIYKGVYQF KSVEFD)[33–35]. Peptides were purchased from 21st Century Biochemicals and purified by reverse phase-HPLC to 90% purity. H-2D^b-$NP_{396–404}$ tetramers were prepared as described[36]. CD1d-PBS57-allophycocyanin tetramers provided by NIAID Tetramer Facility were a gift from L. Berg.

Antibodies and FACS analysis. Fluorescently labelled antibodies and reagents were purchased from BD Biosciences, eBioscience, BioLegend and R&D Biosystems. Flow cytometric analyses of cells were performed on a LSR II cytometer (BD Biosciences) equipped with FACSDiva software and data were analysed using FlowJo software (Tree Star).

CFSE labelling and adoptive transfer. Spleens from donor mice were mechanically disrupted, and erythrocytes were lysed using a 0.84% NH_4Cl solution in order to generate single-cell leukocyte suspensions. Cells were labelled for 15 min at 37 °C with the 2 μM fluorescent dye CFSE (CFDA-SE, Molecular probes), washed, and transferred i.v. (3×10^7 cells) to recipient mice.

***In vivo* cytotoxicity assays.** T-cell cytolytic activity was measured *in vivo* as described previously[21]. Briefly, single-cell suspensions were prepared from spleens of uninfected mice, and separate fractions of cells were then loaded with LCMV peptides (1 μM) for 45 min at 37 °C before labelling with CFSE (2.5, 1 or 0.4 μM, Molecular Probes) for 15 min at 37 °C. After washing, these populations were combined at equal ratios and transferred i.v. into either naive or infected recipients. Survival of each transferred population in the spleens of recipient mice was assessed 16 h after transfer. Specific lysis was calculated as follows: 100 − ((% LCMV target population in infected experimental/% unlabelled population in infected experimental) ÷ (% LCMV target population in naive control/% unlabelled population in naive control)) × 100).

An unconventional *in vivo* cytotoxicity assay was previously established to determine NK-cell killing of lymphocyte populations *in vivo*[21]. Wild-type or *Fas*^{lpr} donor mice were depleted of NK cells and then infected with VV, MCMV, MHV, PV, or a medium dose of LCMV clone 13. At day 4 p.i., single cell splenocyte suspensions were prepared from these mice, labelled with CFSE, and then transferred (2×10^7) into experimental recipient mice on day 3 of medium dose LCMV infection, unless otherwise noted. Recipients included WT, *Prf1*$^{-/-}$, *Tnf*$^{-/-}$, or *Nkg2d*$^{-/-}$ mice that were administered anti-NK1.1 or isotype control antibodies one day before inoculation with PV, MHC, polyI:C, or various doses of LCMV clone 13. Some recipient mice were uninfected and served as controls. Spleens of recipient mice were harvested 5 h after transfer and assessed for survival of donor T cells.

***In vitro* antigen presentation assay.** Stimulator cells were prepared by isolation of single-cell suspensions from the spleens of uninfected as well as isotype-treated or anti-NK1.1-treated mice infected 3 days previously with a medium dose of LCMV i.v. Following irradiation, stimulator cells (5×10^4) were plated at a 1:10 ratio with CFSE-labelled Ly5.1$^+$ LCMV-specific P14 CD8 T cells (5×10^5) in T-cell stimulation medium (RPMI supplemented with 100 U ml^{-1} penicillin G, 100 μg ml^{-1} streptomycin sulphate, 2 mM L-glutamine, 10 mM HEPES, 1 mM sodium pyruvate, 0.1 mM non-essential amino acids, 0.05 mM 2-mercaptoethanol and 10% heat-inactivated (56 °C, 30 min) FBS), which was refreshed every 2 days. P14 cells were enumerated and analysed for dilution of CFSE as a measure of proliferation every 24 h after initiation of co-culture.

Lymphocyte preparation and intracellular cytokine assay. Single-cell leukocyte suspensions from spleens, inguinal lymph nodes, lung and liver were prepared as described previously[21] and were plated at 2×10^6 cells per well in 96-well plates. Cells were stimulated for 5 h at 37 °C with either 1 μM viral peptide or 2.5 μg ml^{-1} anti-CD3 monoclonal antibody in the presence of brefeldin A and 0.2 U ml^{-1} rhIL-2. Stimulated cells were then pre-incubated with a 1:200 dilution of Fc Block (2.4G2) in FACS buffer (HBBS, 2% FCS, 0.1% NaN_3) and stained for 20 min at 4 °C with various combinations of fluorescently tagged monoclonal antibodies. After washing, cells were permeabilized using BD Cytofix/Cytoperm solution and then stained in BD Permwash using monoclonal antibodies specific for various cytokines. AnnexinV staining was performed in azide-free FACS buffer directly *ex vivo* according to manufacturer's instructions (BD Biosciences).

Statistical analysis. Results are routinely displayed as mean ± s.e.m., with statistical differences between experimental groups determined using a two-tailed unpaired Student's *t*-test, where a P value of <0.05 was deemed significant. Statistical differences in survival were determined by log rank (Mantel–Cox) analysis. Graphs were produced and statistical analyses were performed using GraphPad Prism.

30. Zafirova, B. *et al.* Altered NK cell development and enhanced NK cell-mediated resistance to mouse cytomegalovirus in NKG2D-deficient mice. *Immunity* **31,** 270–282 (2009).
31. Exley, M. A. *et al.* Innate immune response to encephalomyocarditis virus infection mediated by CD1d. *Immunology* **110,** 519–526 (2003).
32. Pircher, H. *et al.* T cell tolerance to Mlsa encoded antigens in T cell receptor V beta 8.1 chain transgenic mice. *EMBO J.* **8,** 719–727 (1989).
33. van der Most, R. G. *et al.* Identification of D^b- and K^b-restricted subdominant cytotoxic T-cell responses in lymphocytic choriomeningitis virus-infected mice. *Virology* **240,** 158–167 (1998).
34. Oxenius, A. *et al.* Presentation of endogenous viral proteins in association with major histocompatibility complex class II: on the role of intracellular compartmentalization, invariant chain and the TAP transporter system. *Eur. J. Immunol.* **25,** 3402–3411 (1995).
35. Klavinskis, L. S., Whitton, J. L., Joly, E. & Oldstone, M. B. Vaccination and protection from a lethal viral infection: identification, incorporation, and use of a cytotoxic T lymphocyte glycoprotein epitope. *Virology* **178,** 393–400 (1990).
36. Mylin, L. M. *et al.* Quantitation of CD8$^+$ T-lymphocyte responses to multiple epitopes from simian virus 40 (SV40) large T antigen in C57BL/6 mice immunized with SV40, SV40 T-antigen-transformed cells, or vaccinia virus recombinants expressing full-length T antigen or epitope minigenes. *J. Virol.* **74,** 6922–6934 (2000).

JOB SECURITY UK junior faculty members face poor employment outlook **p.401**

UNITED STATES Age of researchers at first grant is increasing, says study **p.401**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

S. STAFFORD

OUTREACH

Field hospitality

Hosting guests on research trips can give scientists a chance to showcase their work — but it can also cause distractions.

BY LUCAS LAURSEN

Early in his career, Paul Olsen sat in front of a television, expecting to see his own image. He had hosted a television crew on a research expedition to Manicouagan Crater in Canada, where he and his team were investigating the Triassic–Jurassic boundary in the geological record. Olsen, a palaeontologist at Lamont–Doherty Earth Observatory of Columbia University in Palisades, New York, had spent hours explaining and re-explaining for the camera how scientists used the site to reconstruct ancient ecologies. As the opening credits rolled, Olsen wondered how he would come across on the small screen.

But by the end credits, he was still wondering. All that had appeared on screen was a brief flash of Olsen's disembodied hand, as the documentary focused on other researchers from the expedition. In the decades since then, Olsen has hosted at least a dozen teams of journalists, from National Geographic, the BBC and the Canadian Broadcasting Corporation among others, each time devoting days to assisting them and solving logistical challenges on their behalf. He has learned that a journalist's gratitude does not always translate into a starring role — and nor should it. Although some of his colleagues express disappointment when their hours of storytelling don't make the final cut of a documentary, Olsen acknowledges that journalists must be "fairly ruthless", because "they're there to tell a story". Scientist hosts might find that they're not part of that tale.

Hosting guests such as journalists, policy-makers or teachers during research fieldwork is a gamble. In exchange for a shot at publicizing their work or educating a wide audience, scientists must give up an unknowable amount of time and accept the risk — and the added cost — of taking extra people to a remote location. If it goes well, a visiting teacher might help to spread the word, or a policy-maker might be able to make more informed decisions; but scientists may also end up with recalcitrant or shivering visitors who hole up in their rooms instead of getting close to the fieldwork. To make the most of bringing guests to the field, hosts should plan ahead, communicate their goals, try to head off conflicts and expect to exercise patience.

BE PREPARED

Hosting may give researchers an opportunity to sell their science. "It theoretically helps your career," says Kevin Krajick, a media-relations manager and science writer at Lamont–Doherty, who helps to connect journalists with scientists. "It's a bit of marketing — organizations such as the National Science Foundation love to see media — and it can help educate the general public."

The impact may be more important for general fields of study than for any one researcher's work. "I don't think there are too many benefits directly to the scientist," says Olsen. Yet he continues to host guests owing to a sense of responsibility to members of the public, whose taxes fund most of his work. Jill Baron, an ecologist at the US Geological Survey in Fort Collins, Colorado, says that bringing policy-makers closer to the wilderness helps to persuade them that such places are worth protecting. Visits can also establish useful relationships. "When they need information they are likely to call you," says Baron, who has hosted policy-makers ranging from local air-quality commissioners to US senators.

Communicating science in the field can help to further a cause and even, in rare cases, catalyse research-vindicating policies. In 2007, after decades of visits by Colorado officials to Baron's field sites, a consortium including the US Environmental Protection Agency, Colorado's air-quality commission and the US National Park Service drafted a plan to reduce nitrogen deposition in the Rocky Mountain National Park. Baron says that without such visits, "it's hard for people to actually visualize" environmental threats. She once spent hours explaining to a mayoral staffer from Denver how agricultural fertilizer used on Colorado's plains infiltrated the Rocky Mountains. After walking all day through the Rockies, the staffer stopped and asked her: "'You mean what we do down in Denver makes a difference up here?'" Baron recalls. "You can talk until you're blue ▶

Left, palaeontologist Paul Olsen (kneeling) hosting a camera crew in Morocco. Right, Kevin Krajick, a science writer, (left) on a field trip in Bangladesh.

but until they see it they won't understand," she says.

When David McGee, a climatologist at the Massachusetts Institute of Technology in Cambridge, hosted freelance writer Douglas Fox on a field trip, he expected to teach Fox about the science involved in his work. But McGee was surprised by what he learned from the experience himself. "It improved my ability to talk about what I do," he says. "You get better at breaking out of the jargon and lingo and assumptions about what is important" to outsiders (see *Nature* **468,** 465–467; 2010). McGee took Fox, who is based in San Francisco, California, on a trip to the dry bed of Lake Bonneville, which once covered much of what is now northern Utah. McGee says he was excited that someone was interested in his work and helping to publicize it. The expedition led to a feature story in *High Country News*, a magazine for the western United States (go.nature.com/qiiyqj); McGee now shares the article with prospective graduate students and postdocs.

As useful as these outreach lessons were, they did cause frustrating and time-consuming complications. One of McGee's Utah field sites was on US military property. McGee had arranged permits long before the expedition — for scientists. But when his military contacts discovered that Fox was joining the trip, they decided that the team would need a military supervisor. They also decided that they were willing to send the supervisor for only one day. So Fox missed out on some of the expedition's scientific research and McGee got stuck with lengthy, last-minute negotiations. "Once you've invited this person along and they've got their plane tickets, you end up with some responsibility to help as an intermediary," he says. "It all worked out in the end, but it was a headache."

McGee also learned that journalists are attracted to the little conflicts that scientists are less eager to broadcast. With Fox around all the time, every prickly exchange among the researchers, no matter how routine, was potential story material. "Sometimes with collaborators you're short about someone's difference of opinion," says McGee. "But that's not the thing you want to see in print." As a result, some of his colleagues were not comfortable with Fox's presence, and McGee now knows that he needs to check with other members of his team before taking a journalist along.

LESSON PLANNING

A proactive approach before the field visit can help to mitigate misunderstandings, says Krajick. "You want to be prepared," he warns scientists before they take visitors on trips. "Be ready to explain your stuff and have a talk over the phone before going on the expedition to make sure you're on the same page." That allows scientists to lay ground rules such as when they will be available for questions; and it gives journalists a chance to explain when in the editorial process they might want to consult the scientist to check facts.

Olsen discusses mutual expectations with all guests ahead of time. He also prepares handouts for teachers and journalists, summarizing the main points of interest of the field site and the objectives of the expedition. Baron and her students and technicians separate groups of policy-makers into different hiking parties, on the basis of their physical abilities. "Don't take staffers or policy-makers or senators, especially senators, anywhere they're going to be cold or thirsty or get altitude sickness," she advises.

It helps for scientists to have some control over who goes into the field with them. "One of the things we have learned in the past is that the researchers need to be able to pick their teacher," says Janet Warburton, an education coordinator at PolarTREC, a programme based in Fairbanks, Alaska, that gives school teachers a hands-on research experience in the polar regions. PolarTREC filters applicants for motivation and the ability to translate their experience into lesson plans; researchers themselves must ensure that the teacher will be able to cope with the challenges of life and work on their field trip. Lack of filtering can cause wasted opportunities: in one unfortunate case, says Warburton, a teacher not up to the physical challenge of the daily walk to a field site remained in the barracks instead of participating in the fieldwork.

Mark Goldner, a teacher at Heath School in Brookline, Massachusetts, learned about research expectations first hand. Goldner was shortlisted for an expedition to the Svalbard archipelago in the Norwegian Arctic last July. Before he could join the trip, he had to pass a thorough interview with the principal investigator — Julie Brigham-Grette, a palaeoclimatologist at the University of Massachusetts in Amherst. "Julie came right out and said, 'Here's some things we expect of you — what do you think?'" Goldner remembers.

"Be ready to explain your stuff and make sure you're on the same page."

As a trip leader, Brigham-Grette always tells potential participants in her field trips about the physical demands of excursions, such as long walks in the cold, and asks them about their ability to handle responsibilities that include blogging about the trip and helping with expedition equipment. She had one teacher who wasn't able to operate his computer or blog effectively on a trip — and because Brigham-Grette considers communication an important part of a visiting teacher's role, she now asks about those skills. "When I'm in the field, I can't help them navigate that," she says.

Meeting in advance can mitigate stress and

T. MARTIN/ARCUS

Tim Martin, a teacher, tags along with palaeoclimatologist Julie Brigham-Grette in Siberia.

distractions once the trip is under way — in the field, it is sometimes hard for scientists to explain the big picture of the research effort because they are so caught up in the mechanical details of data collection. "When teachers are going out in the field, they're seeing only one piece of the process," says Warburton. "The team is usually very stressed collecting data and may not have the energy to explain the big picture in the moment."

For teacher trips, pre-planning should entail some thought about how the experience will benefit the classroom. Before his own expedition to Siberia with Brigham-Grette, Tim Martin, a teacher at Greensboro Day School in North Carolina, discussed with her how his web design and photography skills could help to translate her science into useful lessons for his students. He also improved his understanding of her work.

Since their expedition together, Brigham-Grette and Goldner have stayed in touch so that she can share her group's analyses of the data that they collected. "The whole goal of PolarTREC shouldn't be about me, or my trip to the Arctic, it's really about my students and the outreach that I can do," says Goldner. "That ongoing collaboration is really important."

ON THE RECORD

Allowing guests, especially journalists, to participate in expeditions may not be the best choice for every expedition. Anything from precarious logistics to bad weather or sensitive politics might cause scientists to postpone, says Olsen. Krajick adds, "If you're not prepared to go on the record all the time, I don't think you should take someone along." Most fieldwork takes place in too intimate a setting to expect much control over gossip or frayed nerves.

Sharing the experience doesn't have to mean relinquishing all control. Olsen has experimented with one way of dealing with the time demanded by television crews: he has decided to not bother trying to conduct real science for the cameras. Mock observations are enough for television journalists, and not having to worry about botching precise measurements makes answering questions easier.

Indeed, hosting often requires flexibility. "You can't plan which days are for discovery and which are photo days," says McGee. He acknowledges that Fox's visit was rather long at four days, but McGee liked the result — a nuanced depiction of his work.

Field guests can be an asset, even beyond outreach: scientists may be able to put visitors to work and make use of their skills. When Brigham-Grette first offered him a place on her trip, Martin was "ready to mop floors". But the team found more useful things for him to do: he had once worked building houses in needy communities, and had construction skills. On one occasion, Brigham-Grette assigned Martin to work on a drill. He didn't falter when the sub-freezing temperatures chilled the drill's fluids and shut it down. Instead, he assumed the role of foreman. "We brought it inside for warmth," he recalls. "But the exhaust was poisonous, so I built an enclosure to channel the exhaust outside."

Goldner happened to know how to drive boats, so Brigham-Grette asked him to ferry scientists and equipment around. And on a remote field site in Morocco, Olsen once asked a television crew with a large budget to transport some of his team members between sites. When it comes to field-trip guests, he says,"their needs and your needs can overlap." ■

Lucas Laursen *is a freelance journalist based in Madrid, Spain.*

JOB SECURITY

British prospects grim

Recruitment and job security for early-career faculty members in the United Kingdom will become more uncertain as government funding for higher education declines, says a trade union. The University and College Union (UCU) in London released a study of higher-education revenues on 5 January. The report says that the proportion of university income accounted for by government funding, estimated to be 31.6% for 2010–11, will fall to 15% by 2014–15. That will probably lead to a reduction in long-term, renewable contracts for junior faculty members, says Stephen Court, the UCU's senior research officer. "The shifting pattern of income from the government will make university employment more precarious," he says.

BIOMEDICAL SCIENCE

Jackson Lab to expand

The Jackson Laboratory has completed plans to open a branch in Connecticut. On 5 January, state governor Dannel Malloy finalized an agreement with Jackson, a biomedical research centre based in Bar Harbor, Maine. The lab will initially lease space from the University of Connecticut Health Center in Farmington. Next year, construction will begin on a permanent facility in Farmington that is set to employ 300 scientists and support staff by 2017. Recruitment is already under way in computational biology and systems genomics for the temporary lab, which will have 27 employees, including scientists, by the end of 2012, says Michael Hyde, Jackson's vice-president for advancement.

UNITED STATES

Scientists miss their peak

US biomedical scientists rarely earn their first major grants during their optimum innovative years, concludes a study (K. R. W. Matthews *et al. PLoS ONE* **6,** e29738; 2011). In 2008, the average age of a scientist getting a first grant from the US National Institutes of Health (NIH) was 42, the authors found. But researchers who won Nobel prizes in medicine or chemistry between 1980 and 2010 did their pioneering work at an average of 41 years; 78% did so before 51, the average age of NIH investigators now. Part of the problem is that the NIH is risk-averse and unwilling to fund nascent work, argues Kirstin Matthews, lead author of the study and a science and technology policy fellow at Rice University in Houston, Texas.

T. MARTIN/ARCUS

Tim Martin, a teacher, tags along with palaeoclimatologist Julie Brigham-Grette in Siberia.

distractions once the trip is under way — in the field, it is sometimes hard for scientists to explain the big picture of the research effort because they are so caught up in the mechanical details of data collection. "When teachers are going out in the field, they're seeing only one piece of the process," says Warburton. "The team is usually very stressed collecting data and may not have the energy to explain the big picture in the moment."

For teacher trips, pre-planning should entail some thought about how the experience will benefit the classroom. Before his own expedition to Siberia with Brigham-Grette, Tim Martin, a teacher at Greensboro Day School in North Carolina, discussed with her how his web design and photography skills could help to translate her science into useful lessons for his students. He also improved his understanding of her work.

Since their expedition together, Brigham-Grette and Goldner have stayed in touch so that she can share her group's analyses of the data that they collected. "The whole goal of PolarTREC shouldn't be about me, or my trip to the Arctic, it's really about my students and the outreach that I can do," says Goldner. "That ongoing collaboration is really important."

ON THE RECORD

Allowing guests, especially journalists, to participate in expeditions may not be the best choice for every expedition. Anything from precarious logistics to bad weather or sensitive politics might cause scientists to postpone, says Olsen. Krajick adds, "If you're not prepared to go on the record all the time, I don't think you should take someone along." Most fieldwork takes place in too intimate a setting to expect much control over gossip or frayed nerves.

Sharing the experience doesn't have to mean relinquishing all control. Olsen has experimented with one way of dealing with the time demanded by television crews: he has decided to not bother trying to conduct real science for the cameras. Mock observations are enough for television journalists, and not having to worry about botching precise measurements makes answering questions easier.

Indeed, hosting often requires flexibility. "You can't plan which days are for discovery and which are photo days," says McGee. He acknowledges that Fox's visit was rather long at four days, but McGee liked the result — a nuanced depiction of his work.

Field guests can be an asset, even beyond outreach: scientists may be able to put visitors to work and make use of their skills. When Brigham-Grette first offered him a place on her trip, Martin was "ready to mop floors". But the team found more useful things for him to do: he had once worked building houses in needy communities, and had construction skills. On one occasion, Brigham-Grette assigned Martin to work on a drill. He didn't falter when the sub-freezing temperatures chilled the drill's fluids and shut it down. Instead, he assumed the role of foreman. "We brought it inside for warmth," he recalls. "But the exhaust was poisonous, so I built an enclosure to channel the exhaust outside."

Goldner happened to know how to drive boats, so Brigham-Grette asked him to ferry scientists and equipment around. And on a remote field site in Morocco, Olsen once asked a television crew with a large budget to transport some of his team members between sites. When it comes to field-trip guests, he says,"their needs and your needs can overlap." ■

Lucas Laursen *is a freelance journalist based in Madrid, Spain.*

JOB SECURITY

British prospects grim

Recruitment and job security for early-career faculty members in the United Kingdom will become more uncertain as government funding for higher education declines, says a trade union. The University and College Union (UCU) in London released a study of higher-education revenues on 5 January. The report says that the proportion of university income accounted for by government funding, estimated to be 31.6% for 2010–11, will fall to 15% by 2014–15. That will probably lead to a reduction in long-term, renewable contracts for junior faculty members, says Stephen Court, the UCU's senior research officer. "The shifting pattern of income from the government will make university employment more precarious," he says.

BIOMEDICAL SCIENCE

Jackson Lab to expand

The Jackson Laboratory has completed plans to open a branch in Connecticut. On 5 January, state governor Dannel Malloy finalized an agreement with Jackson, a biomedical research centre based in Bar Harbor, Maine. The lab will initially lease space from the University of Connecticut Health Center in Farmington. Next year, construction will begin on a permanent facility in Farmington that is set to employ 300 scientists and support staff by 2017. Recruitment is already under way in computational biology and systems genomics for the temporary lab, which will have 27 employees, including scientists, by the end of 2012, says Michael Hyde, Jackson's vice-president for advancement.

UNITED STATES

Scientists miss their peak

US biomedical scientists rarely earn their first major grants during their optimum innovative years, concludes a study (K. R. W. Matthews *et al. PLoS ONE* **6,** e29738; 2011). In 2008, the average age of a scientist getting a first grant from the US National Institutes of Health (NIH) was 42, the authors found. But researchers who won Nobel prizes in medicine or chemistry between 1980 and 2010 did their pioneering work at an average of 41 years; 78% did so before 51, the average age of NIH investigators now. Part of the problem is that the NIH is risk-averse and unwilling to fund nascent work, argues Kirstin Matthews, lead author of the study and a science and technology policy fellow at Rice University in Houston, Texas.

A GAME OF SELF-DECEIT

The resurrection shuffle

BY CLAYTON LOCKE

I saw him duck under the cover of the awning, collapsing and shaking his umbrella, and step into the lazy smoke of the *izakaya*. When he raised his gaze to sweep the diners and drinkers at their tables, it took me a moment to work out what was wrong. Then I got it: I was looking at myself, not as I usually did, left–right inverted in the mirror, but as I saw myself in photographs. It's funny that my brain should pick out such a subtle detail. He saw me sitting at one of the outer tables, next to the plastic sheeting turned opaque with hammering rain. Then, as he moved away from the door, another figure entered. He was with my wife.

His greeting bow was the slightest nod. "*Konbanwa*."

"Why did you bring her, you fool?" Without thinking I had addressed him using *temee*, an aggressive personal pronoun. I didn't even dare look at my wife, for fear of choking emotion.

"Because, Koji-*kun*, I want everything we do here to be witnessed. I'm not underestimating you."

I composed my face into the ever-necessary mask of neutrality. He had taken me off-guard, and I silently reprimanded myself. I harboured a sense of superiority over him, although I knew it to be illogical. This man was exactly me. Well, almost. I had the memories of the mountains and the time of cold, the experience that has shifted my entire way of thinking. A way of thinking I intended to preserve.

"This is illegal. You have to stay more than 20 kilometres away from me," he said.

I couldn't help but laugh at the notion that a bystander would pick us as clones; a salaryman in a crisply tailored suit, and a man in ill-fitting rags. He was right though, yet still he came. That's good, it meant he was afraid. I flicked a glance at my wife. For a heartbeat our eyes met and the blood sang in my ears.

"I've got a proposal for you," I said. I nodded to the empty chairs at my table and after a moment of hesitation they sat. I reached down for the plastic bag beside my chair and placed it on the table between us. Inside were documents and photos.

"So it's blackmail? I credited myself with more imagination than this. Where did you get these?" He paused a moment: when he spoke again his voice was level. "What's your problem? It's legal, you've been given a new identity, you're still *you*. You're wallowing in this self-serving misery, don't look to me to pull you out of it!"

"You don't know the half of it," I said. I reflected on all that had brought me here, how I should never have gone mountaineering without a backup hoverbelt. It was only when the primary one didn't catch and I slipped and fell that I'd realized how stupid I'd been. I'd come to a stop in a deep ravine, my knees shattered, coughing blood. A blow to my head had cracked the personal smartphone chip embedded in my skull. Suddenly I was cut off from the world, no Internet access, no emergency calls, and a silence like I'd not known in many years. I'd been missing for less than two days, but who could survive those subzero temperatures? Lost in the mist and rain, hearing the electric hiss of hover-copters as they glided overhead, and shouting myself hoarse. I'd finally been found, but by the time the gears of bureaucracy had turned, nearly two weeks had passed, and to the world I was dead.

JACEY

How easily they'd discarded me. Life was cheap now, but death was cheaper. Already a clone body had been grown and my latest memory backup imprinted on the fresh brain. It wasn't possible to re-enter my life, for my new self had all the rights and identity: everything from bank accounts and job qualifications to friends and my wife.

I had changed in those weeks. Although I cursed myself for my stupidity, I held the experience in the mountains as the most important thing that had ever happened to me. I wanted to keep it and shape my life with it as part of me. I'd spent long nights seething with jealousy at the imposter taking over my life. Then I had a plan, and the very next day I spent the last yen of my pension on a full up-to-the-minute memory backup at a professional storage agency.

I blinked back into the present. The new me snatched across the table and tried to grab the plastic bag but I caught his wrist and in a moment we were both on our feet and struggling. In the confusion I pulled the gun from my rear pocket. This is where my preparations paid off, my risky burglary of what had once been my own home to retrieve this very weapon.

The gun was between us now (it was an evenly matched fight, I observed wryly) and I sensed a crowd gathering. I twisted backwards, keeping my hold on him, and we fell to the ground, and the gun was pointed at my own temple, my finger around the trigger. I didn't think I'd have the courage, but now in the moment it was all too easy.

"Murderers don't come back," I said. I saw his eyes widen as he recognized his pistol, the realization he'd been out-played. He knew what I knew, the death penalty was sure and a criminal would not be resurrected. It would be the *real* me, who had lived through the mountain cold, who would live again. He tried to withdraw his hands from the pistol but it was too late. The muzzle tight against the side of my head, I pulled the trigger. ■

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

Clayton Locke *is an Australian physicist currently researching frequency standards and metrology in Tokyo.*

LETTER

doi:10.1038/nature10720

Pathway complexity in supramolecular polymerization

Peter A. Korevaar[1,2], Subi J. George[2], Albert J. Markvoort[1,3], Maarten M. J. Smulders[1,2], Peter A. J. Hilbers[1,3], Albert P. H. J. Schenning[2], Tom F. A. De Greef[1,2,3] & E. W. Meijer[1,2]

Self-assembly provides an attractive route to functional organic materials, with properties and hence performance depending sensitively on the organization of the molecular building blocks[1–5]. Molecular organization is a direct consequence of the pathways involved in the supramolecular assembly process, which is more amenable to detailed study when using one-dimensional systems. In the case of protein fibrils, formation and growth have been attributed to complex aggregation pathways[6–8] that go beyond traditional concepts of homogeneous[9–11] and secondary[12–14] nucleation events. The self-assembly of synthetic supramolecular polymers has also been studied and even modulated[15–18], but our quantitative understanding of the processes involved remains limited. Here we report time-resolved observations of the formation of supramolecular polymers from π-conjugated oligomers. Our kinetic experiments show the presence of a kinetically favoured metastable assembly that forms quickly but then transforms into the thermodynamically favoured form. Quantitative insight into the kinetic experiments was obtained from kinetic model calculations, which revealed two parallel and competing pathways leading to assemblies with opposite helicity. These insights prompt us to use a chiral tartaric acid as an auxiliary to change the thermodynamic preference of the assembly process[19]. We find that we can force aggregation completely down the kinetically favoured pathway so that, on removal of the auxiliary, we obtain only metastable assemblies.

The π-conjugated oligomer *S*-chiral oligo(*p*-phenylenevinylene) (SOPV, Fig 1a) serves as a functional material in a variety of organic electronic devices, with performance critically dependent on the material's morphology[20]. This morphology is determined by the assembly mechanisms that transform the molecular dissolved monomers into the materials used in devices. The hydrogen-bonded dimers of SOPV self-assemble in apolar solvents via π–π interactions into one-dimensional helical stacks, and earlier circular dichroism (CD) and ultraviolet–visual spectroscopy studies under thermodynamic control revealed exclusive formation of left-handed *M*-type helical aggregates (*M*-SOPV) through a nucleated growth mechanism[21]. In practice, however, the assembly process is often under kinetic control—prompting us to examine in detail the supramolecular polymerization process under non-equilibrium conditions.

Upon rapid quenching of SOPV from the molecularly dissolved state to 273 K, we observed the formation of a mixture of *M*-SOPV and aggregates with opposite helicity, as evidenced by the opposite sign of the bisignated Cotton effect (Fig. 1b). At 298 K, these right-handed *P*-type aggregates (*P*-SOPV) slowly converted into thermodynamically stable *M*-SOPV aggregates (Supplementary Fig. 1). The observation of metastable *P*-SOPV aggregates indicated that the supramolecular polymerization of SOPV involves two different aggregation pathways, which we termed the on-pathway (leading to *M*-SOPV) and off-pathway (leading to *P*-SOPV) (Fig. 1c).

To study the aggregation mechanism of SOPV and quantify the self-assembly pathways under kinetic control, we conducted stopped-flow experiments in which a concentrated solution of molecularly dissolved SOPV in chloroform was mixed with an excess of methylcyclohexane (MCH) to initiate self-assembly. The subsequent formation of helical SOPV aggregated in MCH was probed using CD spectroscopy (Supplementary Discussion 1). These kinetic experiments were conducted at 293 K and 308 K, with different concentrations of SOPV. At 293 K and the lowest SOPV concentrations, the rate of aggregate formation initially increased with time, characteristic of a lag phase (Fig. 2a, c). The time-dependent CD signal was always negative under these conditions, suggesting the direct formation of thermodynamically stable on-pathway *M*-SOPV aggregates. At higher concentrations ($\geq$9 μM) a positive CD signal appeared in the initial stages of the assembly process and then developed into a negative CD signal at later times, suggesting the initial formation of off-pathway *P*-SOPV aggregates that then convert into thermodynamically stable *M*-SOPV aggregates. Remarkably, the time at which 50% of the aggregation process was completed ($t-50$) is longer at 15 μM than at 10 μM (Fig. 2d). Analogous kinetic studies at 308 K also revealed the presence of a lag phase in experiments at the lowest concentrations (Fig. 2b, e) and an inverted dependence of $t-50$ on concentration, although the shortest $t-50$ time shifts to a higher SOPV concentration (20 μM; Fig. 2d). Detailed analyses unambiguously prove that both the on- and off-pathway are formed via a homogeneously nucleated growth mechanism (Supplementary Figs 5 and 6).

To rationalize the experimental aggregation kinetics, we extended models known in the field of protein fibrillization[6,9,11] by incorporating two competing, nucleated assembly pathways in which prenucleus oligomers (oligomer aggregates below the critical nucleation size) and helical aggregates change size through monomer association and dissociation (Fig. 3a, Supplementary Discussion 2). For the *M*-SOPV aggregates, non-helical prenucleus oligomers change size with rate constants for association and dissociation of a and b, respectively. Once the critical nucleus (with size n) is reached, the helical aggregates are in the elongation regime and change size through monomer association with the same rate constant a, while dissociation proceeds with rate constant c. The steady-state concentration of the aggregates in the nucleation phase is determined by the nucleation equilibrium constant $K_n = a/b$, while in the elongation phase it is determined by the elongation equilibrium constant $K_e = a/c$. The nucleated growth of off-pathway *P*-SOPV aggregates is described analogously, with nucleus size n^*, rate constants a^*, b^* and c^*, and equilibrium constants K_n^* and K_e^*. An essential model assumption is that the transition from metastable to thermodynamically stable aggregates occurs via depolymerization of *P*-SOPV aggregates and subsequent growth of *M*-SOPV aggregates. This is justified by the high helix reversal penalty (8.1 k_BT, where k_B is the Boltzmann constant and T is temperature), obtained by 'majority rules' experiments, which rules out intrastack stereomutation as an alternative transition mechanism (Supplementary Fig. 8)[22].

[1]Institute for Complex Molecular Systems, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, The Netherlands. [2]Laboratory of Macromolecular and Organic Chemistry, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, The Netherlands. [3]Biomodeling and Bioinformatics Group, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, The Netherlands.

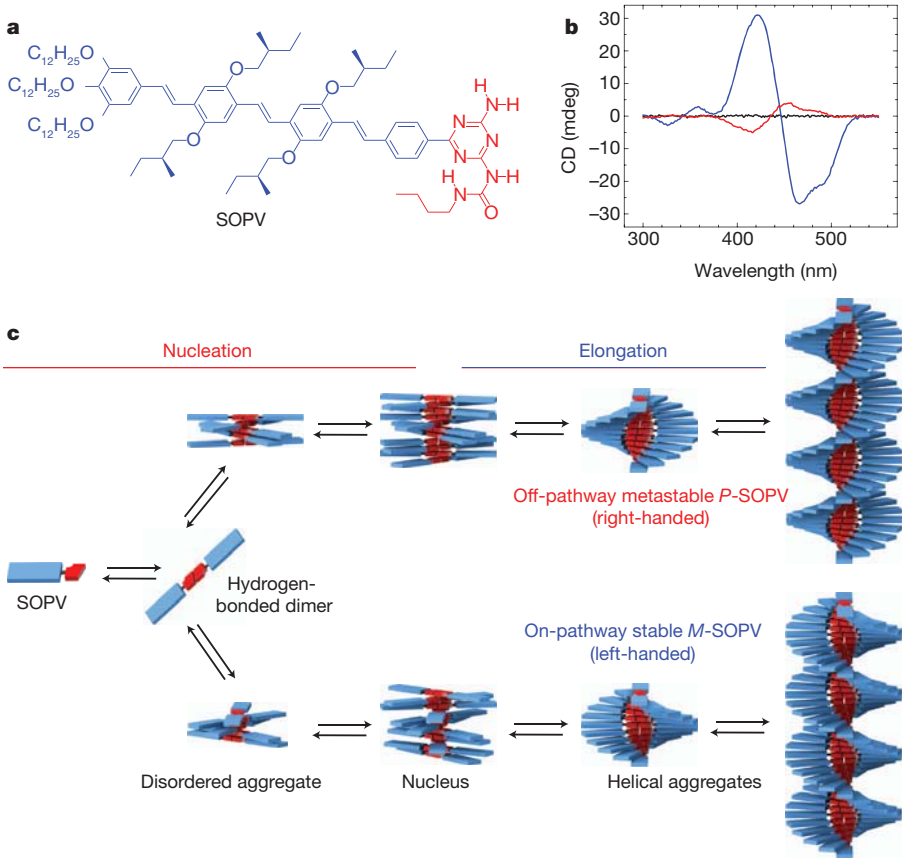


Figure 1 | Pathway complexity in supramolecular polymerization of SOPV. **a**, Molecular structure of SOPV with *S*-chiral butoxy side chains. **b**, CD spectra of SOPV in MCH in disassembled state (343 K, black line), thermodynamically stable *M*-SOPV (273 K, blue line) and mixture of *M*-SOPV and metastable *P*-SOPV (273 K, red line). **c**, Schematic representation of the aggregation pathways of SOPV, including the growth of two competing assemblies. Self-assembly of SOPV in apolar solutions is initiated by formation of a quadruply hydrogen-bonded dimer. The hydrogen-bonded dimer further self-assembles into helical stacks via a nucleation–elongation growth mechanism. Right-handed *P*-helices form quickly but are less stable than the left-handed *M*-helices, which form more slowly.

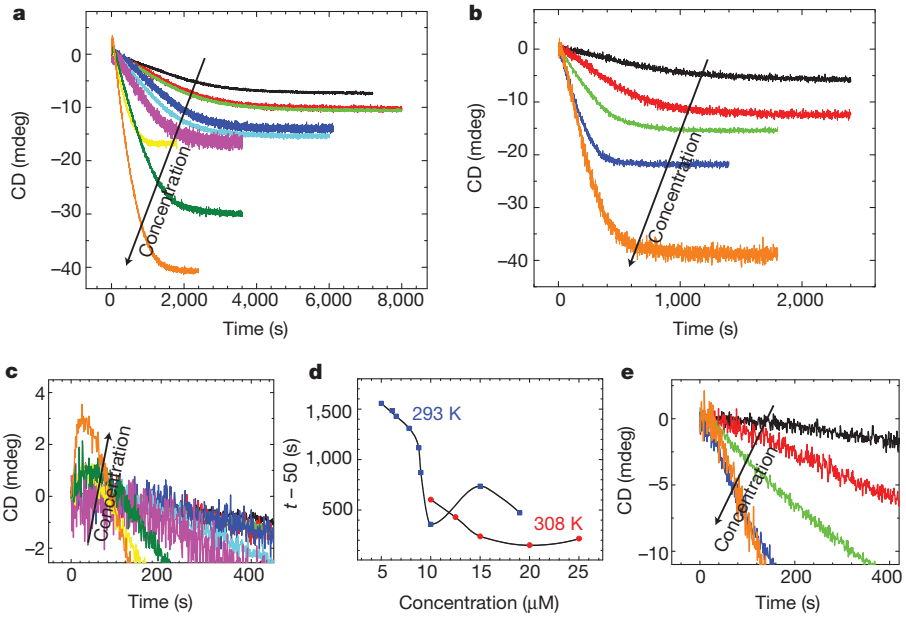


Figure 2 | Experiments under kinetic control. Stopped-flow studies on self-assembly of SOPV. **a**, Concentration-dependent kinetics at 293 K. SOPV concentrations (black arrow shows increasing concentration) are colour-coded as follows: 5 μM (black), 6.1 μM (red), 6.5 μM (light green), 7.8 μM (dark blue), 8.7 μM (light blue), 9 μM (pink), 10 μM (yellow), 15 μM (dark green) and 19 μM (orange). **b**, Concentration-dependent kinetics at 308 K. SOPV concentrations are 10 μM (black), 12.5 μM (red), 15 μM (green), 20 μM (blue) and 25 μM (orange). **c**, Zoom in for initial stages of growth process studied at 293 K in panel **a**. **d**, Time at which 50% of final CD intensity is reached ($t - 50$) versus concentration at 293 K and 308 K. The lines serve to guide the eye. **e**, Zoom in for initial stages of growth process studied at 308 K in panel **b**.

Simulations with this model showed that off-pathway *P*-SOPV aggregates only appear in the initial stages of the self-assembly process if sufficient monomers are present and *P*-SOPV formation is kinetically favoured over that of *M*-SOPV. Specifically, the total monomer concentration (c_{tot}) has to exceed the critical concentration of *P*-SOPV ($c_{tot} > K_e^{*-1}$), which depends on the elongation equilibrium constant of this species. Moreover, the initial formation of *P*-SOPV is kinetically favoured when the forward rate constant of nucleation and elongation is larger for the off-pathway aggregates ($a^* > a$) or if the pre-nucleus oligomers of *P*-SOPV are kinetically more stable ($K_n^* > K_n$, or $a^*/b^* > a/b$) (Supplementary Fig. 9). As a result, off-pathway *P*-SOPV aggregates can only self-assemble at high concentrations and low temperatures (high K_e^*), as was observed experimentally. With $a^* > a$, the kinetic model successfully described the experimental kinetic data, including the initial formation of *P*-SOPV (Fig. 3b). The calculated Gibbs free energy diagram revealed that the *P*-SOPV nucleus is thermodynamically less unstable than the *M*-SOPV nucleus (that is, $K_n^* > K_n$), whereas *M*-SOPV is more stable in the elongation phase (Fig. 3c). Kinetic studies on protein fibrillation have shown that for the initial stages of homogeneously nucleated aggregation involving a thermodynamically unstable nucleus, the conversion to the stable product is proportional to time squared, t^2 (ref. 9). In our experiments with the lowest SOPV concentrations, the changes of the CD signal in the lag phases indeed can be described as being proportional to t^2, demonstrating homogeneous nucleation in the aggregation of SOPV (Supplementary Fig. 10).

To corroborate the proposed involvement of two aggregation pathways, we further analysed the concentration dependence of the kinetic data and found that the changes in the $t - 50$ values with concentration can only be rationalized by taking off-pathway aggregation into account (Fig. 3d, e): the relatively large amount of metastable *P*-type aggregates formed at higher concentrations sequester monomers and lower their solution concentration, thereby hampering the formation of the thermodynamically stable *M*-type aggregates. Our simulations did not show a further decrease of $t - 50$ at higher concentrations with reasonable model parameters. Because our model only considers monomer addition and dissociation, oligomer reactions and fragmentation[14]—which become more important at higher concentrations—may explain this difference between simulated and observed behaviour. Simulations exploring the effect of temperature showed that the maximum in the aggregation rate shifts to higher concentrations and the aggregation rate gets faster (that is, $t - 50$ values get lower) on increasing temperature, as was observed experimentally (Supplementary Fig. 11).

The discovery of these two different assembly pathways, producing structurally different aggregates, encouraged us to attempt to direct the assembly of SOPV towards metastable products. To force the assembly of SOPV into exclusively *P*-type aggregates instead of the mixture obtained upon fast cooling (see above), we used a two-step non-covalent synthetic method (Fig. 4a, b) inspired by previous assembly control experiments[19]. The earlier work involved achiral oligo(*p*-phenylenevinylene) monomers that assemble into chiral helical architectures of equal energy, and showed that the addition of chiral auxiliaries transforms an initially racemic mixture of these structures so that one of the two possible chiral helical architectures dominates[19]. Here, we used a chiral auxiliary to direct the assembly so that, upon removal of the auxiliary, we obtain the thermodynamically unstable *P*-type aggregates.

In the first step, *S*-chiral dibenzoyl tartaric acid (DTA) was used to dictate its chirality to the SOPV monomer via two-fold hydrogen bonding, orthogonal to the SOPV dimerization (Supplementary Discussion 3). Aggregates having opposite helicity (*P*-DTA-SOPV) compared to equilibrium conditions (*M*-SOPV) are formed, as demonstrated by the opposite CD spectrum (Fig. 4c). In the second step, the complete removal of DTA from the SOPV aggregates by aqueous extraction at 273 K using ethylene diamine resulted in the formation of transiently stable *P*-SOPV aggregates at this temperature (Supplementary Fig. 12). The exclusive formation of *P*-SOPV is demonstrated by the absolute ellipticity of *P*-SOPV, which equals the ellipticity of a solution containing exclusively *M*-SOPV aggregates

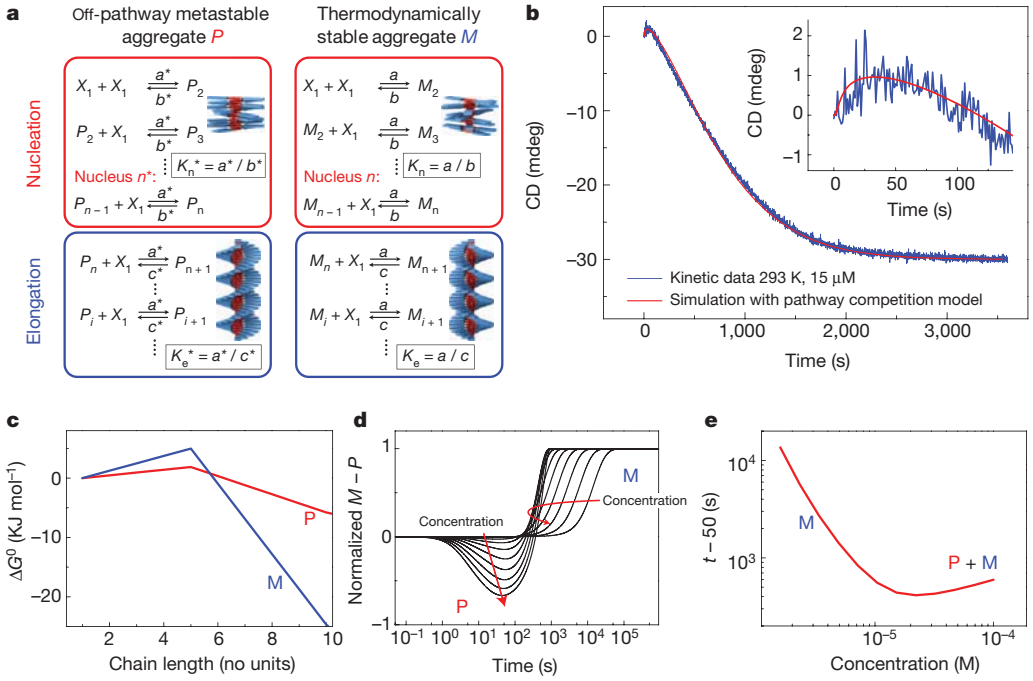


Figure 3 | Analysis aggregation pathway competition with kinetic model. **a**, Schematic representation of reversible kinetic model for nucleated supramolecular polymerization, including a metastable off-pathway and a thermodynamically stable on-pathway competing for free monomer. **b**, Kinetic trace of the self-assembly of SOPV and a simulated kinetic trace using the pathway competition model. The inset to **b** shows a zoom in for the initial stages of the growth process. **c**, Calculated Gibbs free energy diagram for formation of *P*-SOPV and *M*-SOPV. **d**, Concentration-dependent simulation of kinetics with pathway competition model. **e**, $t - 50$ versus SOPV hydrogen-bonded dimer concentration obtained from simulated concentration-dependent kinetics.

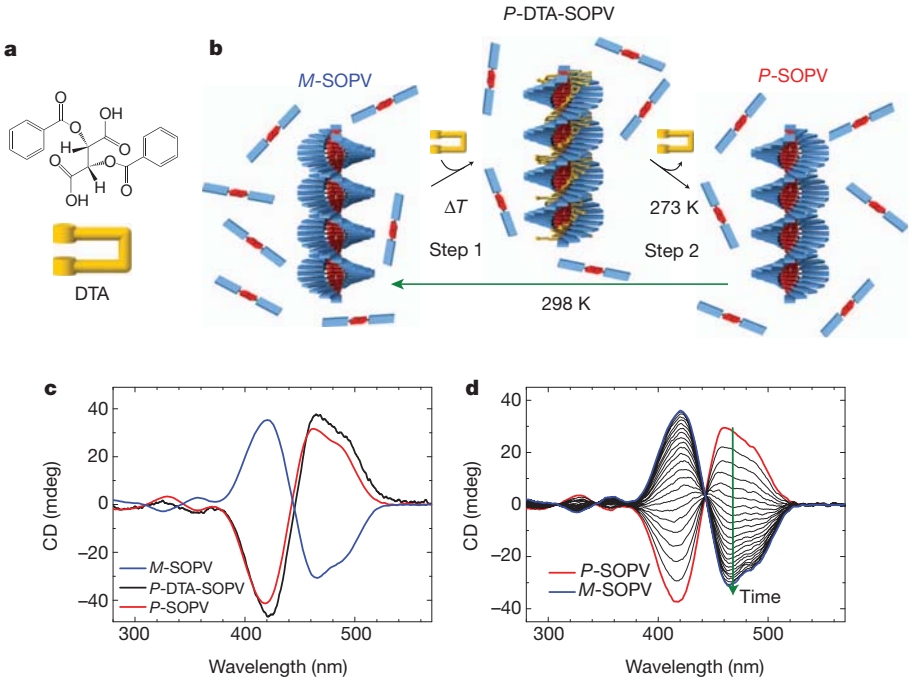


Figure 4 | Controlled formation of exclusively metastable aggregates via a two-step non-covalent synthetic methodology. **a**, Molecular structure of *S*-chiral dibenzoyl tartaric acid. **b**, Schematic representation of preparation of pure *P*-SOPV via a two-step non-covalent synthetic methodology. **c**, CD spectra of *M*-SOPV, *P*-DTA-SOPV and *P*-SOPV. **d**, CD spectra acquired during conversion from metastable *P*-SOPV towards *M*-SOPV upon annealing at 298 K.

under equilibrium conditions. The kinetic lability of *P*-SOPV is demonstrated by annealing at 298 K, resulting in a time-dependent stereomutation of the CD spectra indicative of a conversion from *P*-SOPV towards *M*-SOPV aggregates (Fig. 4d). The single isosbestic point in these time-dependent CD spectra and their apparent first-order kinetics (Supplementary Fig. 13) indeed indicated that two distinct states are involved in the conversion process: *P*-SOPV and *M*-SOPV.

Our observations established that the aggregation of SOPV involves two competing pathways leading to assemblies with opposite helicity, one of which is favoured kinetically and the other thermodynamically. We note that kinetically controlled formation of one-dimensional supramolecular aggregates has been reported[23–26], but the metastable states involved have so far only been probed in terms of their static properties (such as morphology or photophysical properties). Our kinetic studies, by probing the full aggregation process starting from free monomers, reveal that the influence of the metastable state—the off-pathway aggregates—on the overall assembly process is mediated through the equilibrium with free monomers. This effect is a common mechanism for one-dimensional supramolecular systems. We also note that almost all organic materials used in functional devices are processed starting from solutions of the molecular dissolved components. This means that by influencing the subtleties of the self-assembly process generating these materials, through tuning of the on-pathway or off-pathway mechanisms, the resulting morphologies could potentially be controlled to arrive at optimized self-assembled functional materials.

METHODS SUMMARY

Thermodynamically stable *M*-SOPV (MCH, 100 μM) is obtained upon slow cooling from the molecularly dissolved state (60 K h^{-1}) to 273 K, and equilibrium conditions are verified by time-dependent measurements, showing no changes in time. Metastable *P*-SOPV (MCH, 100 μM) is obtained upon quenching molecularly dissolved SOPV to 273 K (ice bath). Stopped-flow experiments are performed by mixing a concentrated SOPV solution in chloroform with MCH (1:50 volume ratio, Berger ball mixer, mixing time <10 ms). The subsequent formation of helical aggregates is probed by a CD spectrometer (466 nm) connected in-line with the stopped-flow instrument. The absence of (oligomeric) SOPV aggregates at the start of the experiment is verified by ^{1}H nuclear magnetic resonance (NMR) studies. The nucleated aggregation pathway competition model describes the kinetic data with parameters $n = n^* = 5$; $a = 2.9 \times 10^4\ M^{-1}\ s^{-1}$; $K_e = 1.52 \times 10^6\ M^{-1}$, $K_n/K_e = 0.0526$, $K_n^*/K_n = 1.38$, $K_e^*/K_e = 0.164$ and $a^*/a = 3.79$ (Matlab, ode15s solver). The value of K_e used is independently assessed by melting experiments. Alternatively, models using diffusion-controlled kinetics afford unrealistic values of $K_n/K_e < 10^{-4}$. The Gibbs free energy as a function of chain length i is calculated using $\Delta G^0_{(i-1)\to i} = -RT\ln(Kc_{tot})$, where R is the gas constant. To obtain *P*-DTA-SOPV, DTA and SOPV are mixed in chloroform (1:1). Subsequently chloroform is evaporated; the residual DTA-SOPV complex is dissolved in MCH (100 μM) on heating above the critical elongation temperature. Cooling from the molecularly dissolved state to 293 K results in *P*-DTA-SOPV aggregates. On adding an equal volume of aqueous ethyl diamine solution (0.15 M) at 273 K, vigorous mixing of the water and MCH phase (1 min) and full separation of both phases, the top MCH layer containing metastable *P*-SOPV is isolated.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 February; accepted 11 November 2011.
Published online 18 January 2012.

1. Yamamoto, Y. *et al.* Photoconductive coaxial nanotubes of molecularly connected electron donor and acceptor layers. *Science* **314,** 1761–1764 (2006).
2. Percec, V. *et al.* Self-organization of supramolecular helical dendrimers into complex electronic materials. *Nature* **417,** 384–387 (2002).
3. Würthner, F. *et al.* Supramolecular p–n-heterojunctions by co-self-organization of oligo(*p*-phenylene vinylene) and perylene bisimide dyes. *J. Am. Chem. Soc.* **126,** 10611–10618 (2004).
4. Capito, R. M., Azevedo, H. S., Velichko, Y. S., Mata, A. & Stupp, S. I. Self-assembly of large and small molecules into hierarchically ordered sacs and membranes. *Science* **319,** 1812–1816 (2008).
5. Dankers, P. Y. W., Harmsen, M. C., Brouwer, L. A., Van Luyn, M. J. A. & Meijer, E. W. A modular and supramolecular approach to bioactive scaffolds for tissue engineering. *Nature Mater.* **4,** 568–574 (2005).
6. Powers, E. T. & Powers, D. L. Mechanisms of protein fibril formation: nucleated polymerization with competing off-pathway aggregation. *Biophys. J.* **94,** 379–391 (2008).
7. Baskakov, I. V., Legname, G., Baldwin, M. A., Prusiner, S. B. & Cohen, F. E. Pathway complexity of prion protein assembly into amyloid. *J. Biol. Chem.* **277,** 21140–21148 (2002).
8. Mulder, A. M. *et al.* Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science* **330,** 673–677 (2010).
9. Ferrone, F. A. Analysis of protein aggregation kinetics. *Methods Enzymol.* **309,** 256–274 (1999).

10. Oosawa, F. & Kasai, M. A theory of linear and helical aggregations of macromolecules. *J. Mol. Biol.* **4,** 10–21 (1962).
11. Powers, E. T. & Powers, D. L. The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the ''supercritical concentration''. *Biophys. J.* **91,** 122–132 (2006).
12. Bishop, M. F. & Ferrone, F. A. Kinetics of nucleation-controlled polymerization. *Biophys. J.* **46,** 631–644 (1984).
13. Xue, W.-F. Homans, S. W. & Radford, S. E. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *Proc. Natl Acad. Sci. USA* **105,** 8926–8931 (2008).
14. Knowles, T. P. J, *et al.* An analytical solution to the kinetics of breakable filament assembly. *Science* **326,** 1533–1537 (2009).
15. De. Greef, T. F. A. *et al.* Supramolecular polymerization. *Chem. Rev.* **109,** 5687–5754 (2009).
16. Cui, H. Chen, Z. Zhong, S., Wooley, K. L. & Pochan, D. J. Block copolymer assembly via kinetic control. *Science* **317,** 647–650 (2007).
17. Lund, R, *et al.* Structural observation and kinetic pathway in the formation of polymeric micelles. *Phys. Rev. Lett.* **102,** 188301 (2009).
18. Pasternack, R. F, *et al.* A nonconventional approach to supramolecular formation dynamics. the kinetics of assembly of DNA-bound porphyrins. *J. Am. Chem. Soc.* **120,** 5873–5878 (1998).
19. George, S. J, *et al.* Helicity induction and amplification in an oligo(*p*-phenylenevinylene) assembly through hydrogen-bonded chiral acids. *Angew. Chem. Int. Edn Engl.* **46,** 8206–8211 (2007).
20. Hoeben, F. J. M. Jonkheijm, P. Meijer, E. W. & Schenning, A. P. H. J. About supramolecular assemblies of π-conjugated systems. *Chem. Rev.* **105,** 1491–1546 (2005).
21. Jonkheijm, P. Van der Schoot, P. Schenning, A. P. H. J. & Meijer, E. W. Probing the solvent-assisted nucleation pathway in chemical self-assembly. *Science* **313,** 80–83 (2006).
22. Van Gestel, J. Amplification of chirality in helical supramolecular polymers: the majority-rules principle. *Macromolecules* **37,** 3894–3898 (2004).
23. Pashuck, E. T. & Stupp, S. I. Direct observation of morphological transformation from twisted ribbons into helical ribbons. *J. Am. Chem. Soc.* **132,** 8819–8821 (2010).
24. Lohr, A. & Würthner, F. Evolution of homochiral helical dye assemblies: involvement of autocatalysis in the ''majority-rules'' effect. *Angew. Chem. Int. Edn Engl.* **47,** 1232–1236 (2008).
25. Wolffs, M, *et al.* The role of heterogeneous nucleation in the self-assembly of oligothiophenes. *Chem. Commun.* 4613–4615 (2008).
26. Ajayaghosh, A. Varghese, R. Praveen, V. K. & Mahesh, S. Evolution of nano- to microsized spherical assemblies of a short oligo(*p*-phenylene-ethynylene) into superstructured organogels. *Angew. Chem. Int. Edn Engl.* **45,** 3261–3264 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Ž. Tomović for providing the SOPV. We are grateful to R. M. Kellogg (Syncom, Groningen, The Netherlands) for providing the tartaric acid derivate. We also thank H. W. H. van Roekel and R. de Bruijn for discussions. Artwork was provided by K. Pieterse. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC grant agreement number 246829, and from the Netherlands Organization for Scientific Research.

Author Contributions P.A.K. and M.M.J.S. performed the stopped-flow experiments. P.A.K., A.J.M. and T.F.A.D.G. analysed the data and developed the kinetic model. S.J.G. performed the experiments with DTA and SOPV. E.W.M., T.F.A.D.G., A.P.H.J.S. and P.A.J.H. supervised the research. P.A.K., T.F.A.D.G. and E.W.M. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.F.A.D.G. (t.f.a.d.greef@tue.nl) or E.W.M. (e.w.meijer@tue.nl).

METHODS

Materials. The synthesis of SOPV has been described previously[27]. All solvents were used as received.

Instrumentation. CD, linear dichroism and corresponding ultraviolet–visual spectra were recorded using a Jasco J-815 CD spectrometer. Sensitivity, response time and scanning speed were chosen appropriately. The temperature was controlled using a Jasco Peltier temperature controller with a range of 263–383 K and adjustable temperature slope. Stopped-flow studies were performed using a Biologic SFM 400 stopped-flow setup with Berger Ball mixer, Biologic TC 100 cuvet (optical path length 1 cm), and Biologic MPS-60 controller unit. The stopped-flow cuvet was connected in-line with a Jasco J-815 CD spectrometer. To control the temperature of the cuvet and syringes, SFM 400 was connected to a Julabo F12 temperature controller (ethylene glycol bath with thermostat). ^{1}H-NMR spectra were recorded at 298 K on a Varian Unita Inova (500 MHz), in $CDCl_3$ (Cambridge Isotope Laboratories). Fluorescence measurements were performed on an Edinburgh Instruments FS920 double-monochromator luminescence spectrometer using a Peltier-cooled red-sensitive photomultiplier. Ultraviolet–visual spectra acquired on the complexation of α-cyclodextrin with *p*-nitrophenolate were recorded on a Perkin Elmer L40 photospectrometer.

Experiments under kinetic control. Thermodynamically stable *M*-SOPV (100 μM) in methylcyclohexane (MCH, spectroscopic grade, Sigma-Aldrich) was obtained upon slow cooling (60 K h^{-1}) from the molecularly dissolved state (343 K) to 273 K; equilibrium conditions were verified by time-dependent measurements, showing no further change in time. Metastable *P*-SOPV (MCH, 100 μM) was obtained upon quenching molecularly dissolved SOPV from 343 K to 273 K (ice bath). The optical path length for the reported spectra was 1 mm. The fraction of *P*-SOPV and *M*-SOPV obtained upon quenching SOPV (MCH, 13 μM) from different temperatures was evaluated from the CD spectra by simulation of a linear combination of the spectra corresponding to a solution consisting of pure *P*-SOPV (obtained via the two-step non-covalent synthetic methodology) and a solution containing *M*-SOPV exclusively. The fraction of free SOPV monomer present before quenching was derived from the CD spectra under equilibrium conditions assuming $\varphi_{monomer} = 1 - \varphi_{helical\ aggregate}$. The optical path length of the reported spectra for these quenching experiments was 1 cm.

Stopped-flow experiments. To probe the kinetics of the supramolecular polymerization of SOPV, SOPV in chloroform (spectroscopic grade, Sigma-Aldrich) was mixed with MCH (50:1; 1,508 μl:30 μl, injection rate 8–9 ml s^{-1}, mixing time <10 ms), after which CD was followed in time ($\lambda = 466$ nm; $\Delta\lambda = 1$ nm; $\Delta t = 1$ s; standard sensitivity; high tension voltage adjusted to get a direct-current voltage around 1 V). The SOPV concentration and the duration of the measurement were adjusted appropriately. The absence of (oligomeric) SOPV aggregates at the start of the experiment was verified by ^{1}H-NMR studies. Kinetic curves were averaged over multiple stopped-flow experiments as follows. At 293 K: 5 μM, 9; 6.1 μM, 2; 6.5 μM, 6; 7.8 μM, 1; 8.7 μM, 4; 9 μM, 2; 10 μM, 9; 15 μM, 3; 19 μM, 6. At 308 K: 10 μM, 6; 12.5 μM, 6; 15 μM, 11; 20 μM, 9; 25 μM, 6. After the measurement, the concentration was verified using CD (466 nm) and ultraviolet–visual (438 nm) spectra. To test the mixing in the stopped-flow setup, the complexation of α-cyclodextrin and *p*-nitrophenolate was measured. Stopped-flow experiments were performed upon 100:1 mixing of an α-cyclodextrin (Across) and a *p*-nitrophenolate (Merck) solution. Both solutions were buffered at pH = 11 (HPO_4^{2-}/PO_4^{3-} buffer). Settings: injection volumes (100:1) 2,011 μl:20 μl, injection rate 7 ml s^{-1}, mixing time <10 ms, $\lambda = 410$ nm, $\Delta\lambda = 1$ nm, $l = 1$ cm, $\Delta t = 1$ s, standard sensitivity, high tension voltage = 300 V; data averaged over 5 injections.

Aggregation pathway competition kinetics model. The kinetic model describes the competing nucleated aggregation pathways of both *P*-SOPV and *M*-SOPV via a sequence of differential equations. Each differential equation describes monomer association and dissociation to the respective aggregate with length *i*. Assuming that for $i > N$, with $N \gg n$ ($N = 100$), $[X_{i+1}] = \alpha[X_i]$, the required number of differential equations was reduced to $2(N + 2)$. The resulting system of differential equations is solved using the ode15s solver in Matlab. The aggregation pathway competition model describes the kinetic data with parameters $n = n^* = 5$; $a = 2.9 \times 10^4$ M^{-1} s^{-1}; $K_e = 1.52 \times 10^6$ M^{-1}, $K_n/K_e = 0.0526$, $K_n^*/K_n = 1.38$, $K_e^*/K_e = 0.164$ and $a^*/a = 3.79$. The value of K_e used was independently assessed by melting experiments. Alternatively, models using diffusion-controlled kinetics afforded unrealistic values of $K_n/K_e < 10^{-4}$. The Gibbs free energy diagram was calculated via $\Delta G^0_{(i-1)\rightarrow i} = -RT\ln(Kc_{tot})$, with gas constant *R*, temperature *T*, equilibrium constant $K = K_n$ for $i \leq n$ and $K = K_e$ for $i > n$ and assuming the total SOPV hydrogen-bonded dimer concentration c_{tot} as the reference state, whereas the standard state was defined[28] at 1 M. The role of temperature on the aggregation pathway competition was rationalized by performing kinetic simulations at different temperatures. The temperature-dependency of the forward rate constant was described by the well-known Arrhenius equation: $a = A\exp(-E_{act}/RT)$, where *A* is the kinetic prefactor and E_{act} the activation energy. The temperature-dependency of K_e was described using the Van't Hoff equation: $K_e = \exp(-(\Delta H_e^0 - T\Delta S_e^0)/RT)$, where ΔH_e^0 and ΔS_e^0 represent the standard enthalpy and entropy of elongation, respectively. The other rate constants are defined via temperature-independent ratios.

Two-step non-covalent synthetic methodology. To obtain *P*-DTA-SOPV, the DTA (provided by Syncom) and SOPV were mixed in chloroform (1:1). Subsequently, chloroform was evaporated by purging with nitrogen gas; the residual DTA-SOPV complex (red colour) was dissolved in MCH (100 μM) upon sonication and heating until the sample turned green, indicating molecular dissolution. Cooling from this molecularly dissolved state to 293 K resulted in *P*-DTA-SOPV aggregates. On adding an equal volume of aqueous ethyl diamine solution (0.15 M) at 273 K, vigorous mixing of the water and MCH phase (1 min) and full separation of both phases, the top MCH layer containing metastable *P*-SOPV could be isolated.

27. Jonkheijm, P. *et al.* Transfer of π-conjugated columnar stacks from solution to surfaces. *J. Am. Chem. Soc.* **125,** 15941–15949 (2003).
28. Xue, W.-F., Homans, S. W. & Radford, S. E. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *Proc. Natl Acad. Sci. USA* **105,** 8926–8931 (2008).

LETTER

doi:10.1038/nature10734

Multi-isotope imaging mass spectrometry quantifies stem cell division and metabolism

Matthew L. Steinhauser[1,2], Andrew P. Bailey[3], Samuel E. Senyo[1,2], Christelle Guillermier[2,4,5], Todd S. Perlstein[1,2], Alex P. Gould[3], Richard T. Lee[1,2,6] & Claude P. Lechene[2,4,5]

Mass spectrometry with stable isotope labels has been seminal in discovering the dynamic state of living matter[1,2], but is limited to bulk tissues or cells. We developed multi-isotope imaging mass spectrometry (MIMS) that allowed us to view and measure stable isotope incorporation with submicrometre resolution[3,4]. Here we apply MIMS to diverse organisms, including *Drosophila*, mice and humans. We test the 'immortal strand hypothesis', which predicts that during asymmetric stem cell division chromosomes containing older template DNA are segregated to the daughter destined to remain a stem cell, thus insuring lifetime genetic stability. After labelling mice with ^{15}N-thymidine from gestation until post-natal week 8, we find no ^{15}N label retention by dividing small intestinal crypt cells after a four-week chase. In adult mice administered ^{15}N-thymidine pulse-chase, we find that proliferating crypt cells dilute the ^{15}N label, consistent with random strand segregation. We demonstrate the broad utility of MIMS with proof-of-principle studies of lipid turnover in *Drosophila* and translation to the human haematopoietic system. These studies show that MIMS provides high-resolution quantification of stable isotope labels that cannot be obtained using other techniques and that is broadly applicable to biological and medical research.

MIMS combines ion microscopy with secondary ion mass spectrometry (SIMS), stable isotope reporters and intensive computation (Supplementary Fig. 1). MIMS allows imaging and measuring of stable isotope labels in cell domains smaller than one micrometre cubed. We tested the potential of MIMS to track DNA labelling quantitatively with ^{15}N-thymidine *in vitro*. In proliferating fibroblasts, we detected label incorporation within the nucleus by an increase in the ^{15}N:^{14}N ratio above natural ratio (Fig. 1a). The labelling pattern resembled chromatin with either stable isotope-tagged thymidine or thymidine analogues (Fig. 1b). We measured dose-dependent incorporation of ^{15}N-thymidine over three orders of magnitude (Fig. 1c and Supplementary Fig. 2). We

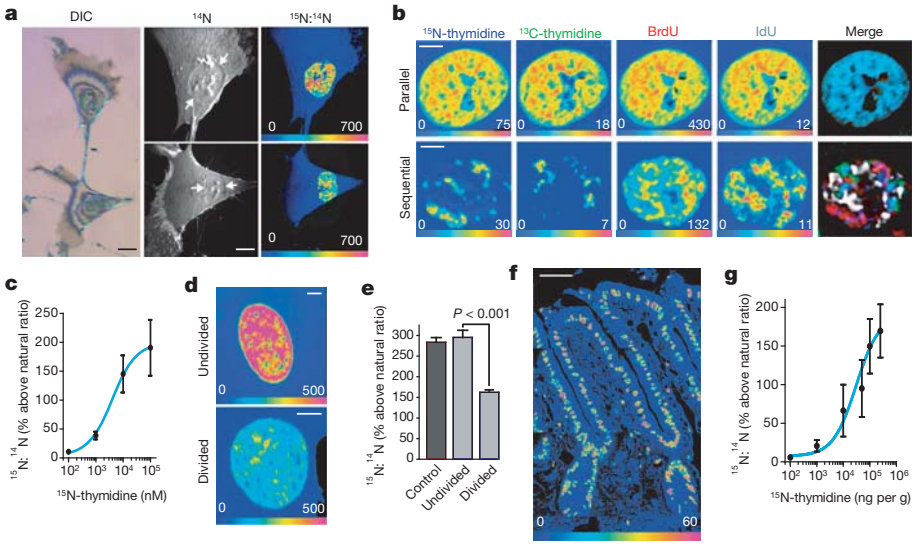


Figure 1 | MIMS quantification of stable isotope-labelled thymidine incorporation by dividing cells. **a**, Dividing fibroblast labelled with ^{15}N-thymidine. The cell surface was sputtered to reach the nuclei. Left, differential interference contrast reflection microscopy. Middle, ^{14}N image revealing subcellular details including the nucleus (white arrows) with nucleoli. Right, hue saturation intensity (HSI) image mapping the ^{15}N:^{14}N ratio. The rainbow scale ranges from blue, set to natural ratio (0.37%, expressed as 0% above natural ratio), to red, where the ratio is several fold above natural ratio (700%, eight times the natural ratio). ^{15}N labelling is concentrated in the nucleus. Scale bars, 10 μm. **b**, Fibroblast nuclei after serial or parallel DNA labelling with ^{15}N-thymidine, ^{13}C-thymidine, BrdU (^{81}Br) or iododeoxyuridine (^{127}I). Parallel labelling (top row): colocalization of label confirmed by the merged image (far right). Sequential labelling (bottom row): non-superimposable nuclear labelling. Scale bars, 5 μm. **c**, Concentration-dependent nuclear ^{15}N-labelling in ^{15}N-thymidine treated fibroblasts. Pixel-by-pixel quantification of ^{15}N-incorporation. Data are derived from the mean of intranuclear pixels. Sigmoidal dose response curve: $R^2 = 0.99$. **d**, Nuclei from ^{15}N-thymidine-labelled human foreskin fibroblasts after 24-h chase (cycle approximately 18 h). Scale bars, 5 μm. **e**, One group labelled similarly to control cells (undivided), the other with labelling that was approximately half that of control (divided). **f**, ^{15}N:^{14}N HSI image of the small intestine. ^{15}N-thymidine labelling (one week) of nuclei extends from the crypts to the tips of the villi. Mosaic: 8 tiles, 80 μm each. Scale bar, 30 μm. **g**, Sigmoidal dose response curve: ^{15}N-thymidine labelling after single subcutaneous injection ($R^2 = 0.99$).

[1]Department of Medicine, Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [2]Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Division of Physiology and Metabolism, Medical Research Council National Institute for Medical Research, Mill Hill, London NW7 1AA, UK. [4]National Resource for Imaging Mass Spectroscopy, 65 Landsdowne St., Cambridge, Massachusetts 02139, USA. [5]Department of Medicine, Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [6]Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA.

also tracked fibroblast division after a 24-h label-free chase (Fig. 1d, e and Supplementary Fig. 3). Cells segregated into two populations, one indistinguishable from control cells suggesting no division, the other with halving of label, consistent with one division during the chase.

We found similar results by tracking cell division *in vivo* in the small intestine (Fig. 1f, g and Supplementary Figs 4–6). We measured dose-dependent ^{15}N-thymidine incorporation within nuclei of actively dividing crypt cells (Fig. 1g and Supplementary Fig. 4), down to a dose of 0.1 μg per g (Supplementary Fig. 2). The cytoplasm was slightly above natural ratio, probably due to low-level soluble ^{15}N-thymidine or mitochondrial incorporation (Supplementary Fig. 2). We measured halving of label with each division during label-free chase (Supplementary Fig. 6).

We then tested the 'immortal strand hypothesis', a concept that emerged from autoradiographic studies[5] and that predicted long-term label retaining cells in the small intestinal crypt[6,7]. It proposes that asymmetrically dividing stem cells also asymmetrically segregate DNA, such that older template strands are retained by daughter cells that will remain stem cells and newer strands are passed to daughters committed to differentiation (Supplementary Fig. 7)[5,6]. Modern studies continue to argue both for[8–12] or against[13–16] the hypothesis, leading to the suggestion that definitive resolution of the debate will require a new experimental approach[17].

Although prior evidence suggests a concentration of label-retaining cells in the +4 anatomic position[7,8], we searched for DNA label retention irrespective of anatomic position or molecular identity. We labelled mice with ^{15}N-thymidine for the first 8 weeks of life when intestinal stem cells are proposed to form[8]. After a 4-week chase, mice received bromodeoxyuridine (BrdU) for 24 h before euthanasia to identify proliferating cells (Fig. 2a and Supplementary Fig. 8, Experiment 1), specifically crypt base columnar (CBC) cells and transit amplifying cells (Supplementary Fig. 9), which cycle at a rate of one and two times per 24 h, respectively[18] (Supplementary Fig. 10). All crypt cell nuclei were highly labelled upon completion of ^{15}N-thymidine; after a 4-week chase, however, we found no label retention by non-Paneth crypt cells (Fig. 2b–f; $n = 3$ mice, 136 crypts analysed). ^{15}N-labelling in $^{15}N^+$/BrdU$^-$ Paneth and mesenchymal cells was equivalent to that measured at pulse completion (Fig. 2b, c) indicating quiescence during the chase (values above ^{15}N:^{14}N natural ratio: Paneth pulse = 107.8 ± 5.0% s.e.m. $n = 51$ versus Paneth pulse-chase = 96.3 ± 2.8% s.e.m. $n = 218$; mesenchymal pulse = 92.0 ± 5.0% s.e.m. $n = 89$ versus mesenchymal pulse-chase = 90.5 ± 2.2% s.e.m. $n = 543$). The number of randomly selected crypt sections was sufficient to detect a frequency as low as one label-retaining stem cell per crypt irrespective of anatomic location within the crypt. Because each anatomic level contains approximately 16 circumferentially arrayed cells[8], a two-dimensional analysis captures approximately one eighth of the cells at each anatomic position (one on each side of the crypt; Supplementary Fig. 9a). Therefore, assuming only one label-retaining stem cell per crypt we should have found 17 label-retaining cells in the 136 sampled crypts (one eighth of 136) but we found 0 (binomial test $P < 0.0001$). The significance of this result held after lowering the expected frequency of label-retaining cells by 25% to account for the development of new crypts, a process thought to continue into adulthood[19]. In three additional experiments, using shorter labelling periods and including *in utero* development, we also found no label-retaining cells in the crypt other than Paneth cells (Supplementary Fig. 8, Experiments 2–4).

To address the possibility that long-term thymidine exposure or frequent high-dose injections introduced pitfalls, we limited labelling to the previously reported peak time period of label-retaining stem cell formation[8] and reduced the dose by 50-fold compared to prior experiments (Supplementary Fig. 8, Experiment 5). Approximately six label-retaining cells per intestinal circumference were observed previously[8] with a similar protocol; with about 90 crypts per circumference, this translates to 1 label-retaining cell per 15 sectioned crypts. We analysed 330 crypts and observed 19 $^{15}N^+$/BrdU$^-$ Paneth cells (one per 17 crypt

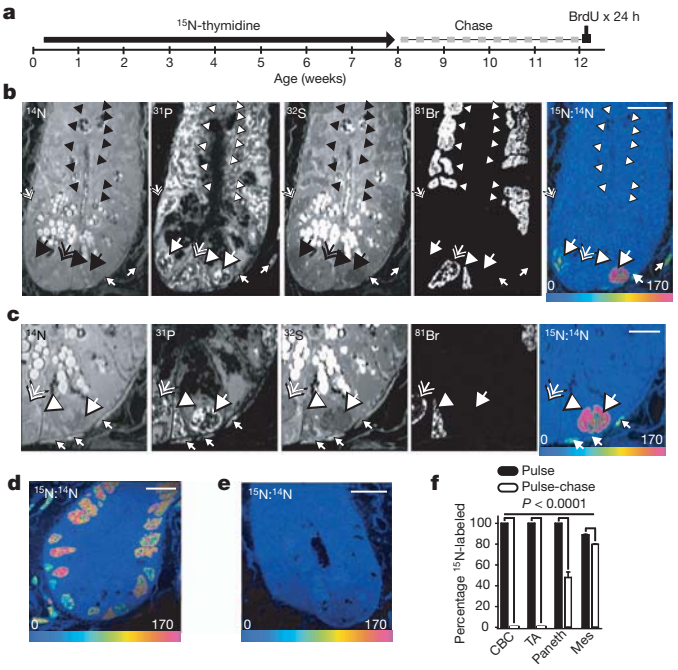


Figure 2 | No label-retaining stem cells in the small intestinal crypt. **a**, ^{15}N-thymidine administered from post-natal day 4 to week 8. After 4-week chase, BrdU was administered (500 μg intraperitoneally every 6 h) for 24 h before euthanasia (see Supplementary Fig. 8). **b**, ^{14}N: crypt structure and intense signal in intracytoplasmic Paneth granules at the crypt base. ^{31}P: intense intranuclear signal. ^{32}S: intense signal within cytoplasmic Paneth cell granules. ^{81}Br: direct measure of BrdU incorporation. ^{15}N:^{14}N: HSI image of ^{15}N-thymidine labelling within $^{15}N^+$/BrdU$^-$ Paneth cells (large arrow) and mesenchymal cells (small arrows). No other cells reveal ^{15}N retention. Large arrow head: recently formed (BrdU$^+$/$^{15}N^-$) Paneth cell. Small hatched arrow (middle left side of the crypt): unlabelled Paneth cell, ($^{15}N^-$/BrdU$^-$). Scale bar, 15 μm. **c**, Continued analysis of the same crypt after narrowing the acquisition field. High ^{15}N signal in a BrdU$^-$ Paneth cell (large arrow) and mesenchymal cells closely associated with the crypt (small arrows). BrdU$^+$ CBC (hatched arrow) and Paneth cell (arrow head) are $^{15}N^-$. Scale bar, 5 μm. **d**, ^{15}N:^{14}N HSI image of small intestinal crypt at the end of ^{15}N-thymidine pulse. All nuclei are labelled. Nuclei with lesser degrees of labelling likely represent cells born during a period of lower circulating ^{15}N-thymidine as expected given the different labelling protocols (Supplementary Fig. 8). Scale bar, 20 μm. **e**, Unlabelled mouse image. The entire crypt contains the natural abundance of ^{15}N. Scale bar = 10 μm. **f**, Mean % $^{15}N^+$ cells at the completion of pulse and after pulse-chase (± standard error of CBC, transit amplifying (TA), Paneth and mesenchymal (Mes) cells ($n = 3$ mice per group).

sections), but no non-Paneth label-retaining cells (expected, 21; observed, 0; binomial test $P < 0.0001$) or doubly labelled ($^{15}N^+$/BrdU$^+$) cells irrespective of cellular identity. Thus, after labelling during all potential periods of stem cell formation, spanning *in utero* through post-natal development, we found no label-retaining stem cells in the small intestinal crypt other than Paneth cells, which are largely quiescent. We conclude therefore that the phenomenon of label-retaining cells as described by others was most likely due either to reagent toxicity or to insufficient resolution with the consequent misidentification of labelled post-mitotic Paneth cell nuclei as label retaining +4 stem cells. MIMS minimizes these potential artefacts due to its ability to detect non-radioactive isotopes and the high imaging resolution both laterally and vertically. Although our data cannot directly exclude the migration of label-retaining stem cells into the crypt under non-homeostatic conditions such as radiation injury[11], prior examinations of crypt clones argue against contributions from either a resting stem cell population or from stem cells originating outside the crypt[20].

We also studied label release in adult mice after stem cell formation. These experiments are particularly relevant given the recent suggestion

that crypt stem cells exhibit monoclonal drift over 8 months[21,22], a model that is incompatible with lifetime survival of label-retaining stem cells but that does not exclude the possibility of biased template strand segregation. After administering ^{15}N-thymidine for 2 weeks to label all proliferating crypt cells, ^{15}N-thymidine was stopped and BrdU was administered for 24 h to label cells dividing during the ^{15}N-thymidine-free chase (Fig. 3a). A consequence of the immortal strand hypothesis would be that a stem cell would produce a daughter in which ^{15}N label was completely lost with the first division during chase (Supplementary Fig. 7). We observed no such BrdU^{+}/^{15}N^{-} crypt cells, even after extending the ^{15}N-thymidine-free chase to 48 h (Fig. 3b and Supplementary Fig. 11). Proliferating (BrdU^{+}) crypt cells (n = 3mice; 563 cells analysed) diluted ^{15}N-thymidine label (mean 71% above natural ratio) compared to undivided (BrdU^{-}) cells (129% above natural ratio) or cells analysed at the end of ^{15}N-thymidine pulse (128% above natural ratio), as expected of cells undergoing random chromosomal segregation (Fig. 3c). We also analysed crypt cell nuclei in late mitosis, finding ^{15}N label and BrdU in both sets of segregating chromosomes (Fig. 3d and Supplementary Fig. 12, n = 232). Together, these data result from an analysis of 625 crypts, in which we did not find a single BrdU^{+} nucleus or chromosomal complement that was not ^{15}N-labelled, indicating that dividing cells in the crypt uniformly dilute labelled DNA in a pattern consistent with random DNA strand segregation.

The broad applicability of MIMS is demonstrated by proof-of-principle studies in *Drosophila* and humans. *Drosophila* provides a model system in which to study the genetics of lipid metabolism, *in vivo*. Due to its small size, however, it is difficult to obtain sufficient material from specific tissue/cell types for measurement with conventional mass spectrometry. A previous study showed numerous lipid droplets in the enterocytes of a segment of the *Drosophila* larval intestine, the anterior midgut[23]. We therefore used MIMS to measure the incorporation and turnover of diet-derived ^{13}C-palmitate within these cells. After larval exposure to dietary ^{13}C-palmitate, label is incorporated into enterocyte lipid droplets with a subcellular pattern similar to that seen with a neutral lipid dye (compare Fig. 4a with supplementary Fig. 13). Dietary ^{13}C-palmitate pulses of varying periods indicate that the rate of ^{13}C label incorporation into lipid droplets is greater in anterior

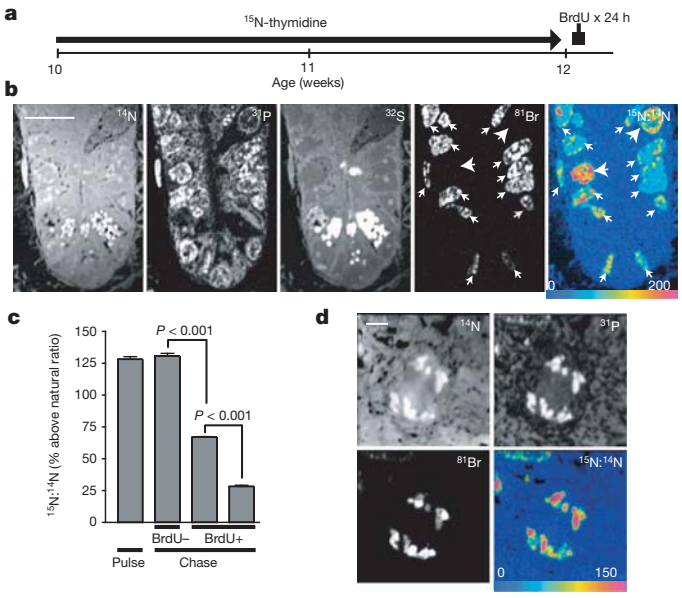


Figure 3 | Label-dilution in adult mice indicates random segregation of DNA strands. **a**, ^{15}N-thymidine administered for 2 weeks to adult mice (osmotic mini-pump, 20 μg h^{-1}), then BrdU for 24 h before euthanasia (See Supplementary Fig. 8). **b**, Dividing cells (BrdU^{+}) dilute ^{15}N-thymidine label (small arrows) relative to undivided cells (BrdU^{-}) (large arrows). Note two CBC cells with elongated nuclei at the crypt base. Scale bar, 10 μm. **c**, Divided crypt cells (BrdU^{+}), residing in CBC or +4–10 positions, demonstrated ^{15}N-dilution consistent with one or two rounds of division during the chase (n = 3 mice per group). **d**, Mitotic crypt cell. Segregating chromosomes are visible in ^{14}N and ^{31}P images. ^{15}N-label and BrdU were measured in both segregating chromosomal complements consistent with symmetric chromosomal segregation. Scale bar, 2 μm.

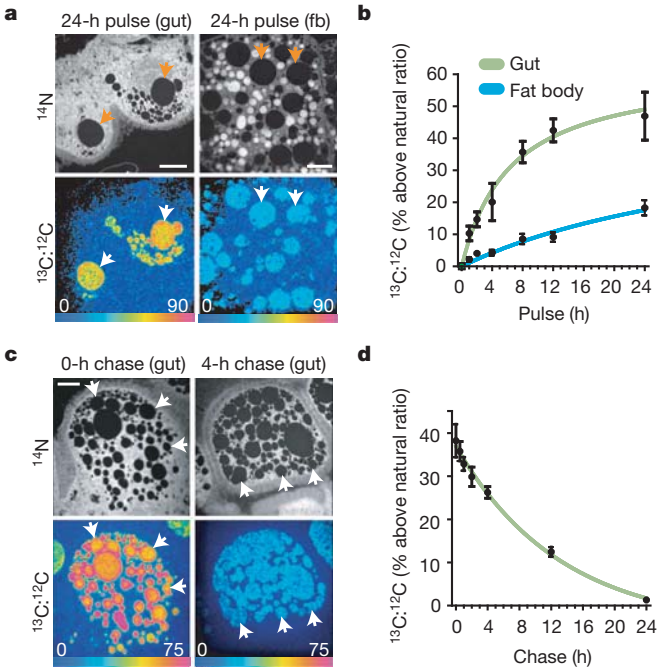


Figure 4 | Extending MIMS to *Drosophila* and human metabolism. **a**, *Drosophila* anterior midgut enterocytes (gut, left panels) accumulate more ^{13}C from 24-h exposure to dietary ^{13}C-palmitate than adipocytes (fat body cells (fb), right panels). Lipid droplets (arrows) have low nitrogen content and so appear dark in ^{14}N images (top panels). The ^{13}C:^{12}C HSI images reveal ^{13}C incorporation into lipid droplets. Scale bars, 10 μm. **b**, Time course of ^{13}C accumulation in *Drosophila* lipid droplets of anterior midgut enterocytes (Gut) and adipocytes (Fat body) over 24 h of larval exposure to dietary ^{13}C-palmitate. Initial rates of ^{13}C-incorporation in the lipid droplets were 9.7% h^{-1} (Gut) and 1.16% h^{-1} (Fat body) with the first time point after pulse taken at 1 h. **c**, Rapid lipid turnover in *Drosophila* anterior midgut enterocytes (gut) after larval dietary ^{13}C-palmitate pulse-chase. Upper panels show ^{14}N images and lower panels show corresponding ^{13}C:^{12}C HSI images, indicating strong ^{13}C incorporation at the end of the pulse (bottom left) and rapid elimination during a 4-h chase (bottom right). Scale bar = 5 μm. **d**, ^{13}C tracer dilution in the lipid droplets of anterior midgut enterocytes during a label-free chase fits an exponential decay curve (R^2 = 0.90, half-life = 9.1 h). **e**, ^{15}N-thymidine administered for 48 h: no labelled white blood cell were found. After a 4-week chase we found few labelled lymphocytes. Top: unlabelled polymorphonuclear leukocyte (PMN) with multi-lobed nucleus. Bottom: ^{15}N-labelled lymphocyte after chase. The round nucleus seen in the ^{31}P image occupies the majority of the cell, as is typical for lymphocytes. Scale bars, 3 μm.

midgut enterocytes than in the adipocytes of the fat body, the major lipid storage depot in *Drosophila* (Fig. 4a, b). Pulse-chase experiments with dietary ^{13}C-palmitate also demonstrate a reproducible exponential decay ($R^2 = 0.9$) of the ^{13}C signal within enterocyte lipid droplets and provide an estimate for the half-life of approximately 9 h (Fig. 4c, d). These results demonstrate the utility of MIMS for precise measurement of *in vivo* lipid turnover at high resolution within individual *Drosophila* lipid droplets.

In the first MIMS human experiment, we administered ^{15}N-thymidine to a healthy volunteer by intravenous infusion for 48 h. We performed MIMS analysis on peripheral white blood cell smears. We found no labelled white blood cells at the end of the infusion (3,600 analysed). After a 4-week chase we found a few labelled white blood cells (Fig. 4c, 4 of approximately 2,000 cells analysed) consistent with a lag time before release from the bone marrow (chi-squared, $P < 0.01$). Radiolabelled thymidine has been administered to humans in numerous studies previously[24], but these studies were conducted primarily on cancer patients. There is no known risk to using stable isotopes[25]. Thus, this experiment now opens the door to studies of metabolism and cell tracking in humans.

In conclusion, we provide direct evidence under physiological conditions against non-random template strand segregation in the small intestine. Our approach encompasses all potential stem cell populations in the crypt, those identifiable by known markers[26–28] and putative ones not identified by a specific marker. We have demonstrated the power of MIMS to image and quantify cell metabolism in three different contexts: mammalian intestinal cell division, *Drosophila* lipid metabolism and human lymphopoiesis. We anticipate that the measuring and imaging power of MIMS will make it a major tool to study metabolism and cell fate in animals and humans.

METHODS SUMMARY

Data were acquired with the prototype (four detectors) and a second-generation instrument (NanoSIMS 50L, Cameca: seven detectors). Quantitative mass images were analysed with Open MIMS software, a plug-in to ImageJ (http://www.nrims.harvard.edu/software.php). Stable isotopes were obtained from Cambridge Isotopes (^{15}N-thymidine, ^{13}C-thymidine) and Sigma-Aldrich (^{13}C-palmitate).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 March; accepted 23 November 2011.
Published online 15 January 2012.

1. Schoenheimer, R. & Rittenberg, D. The application of isotopes to the study of intermediary metabolism. *Science* **87,** 221–226 (1938).
2. Schoenheimer, R. *et al.* The application of the nitrogen isotope N^{15} for the study of protein metabolism. *Science* **88,** 599–600 (1938).
3. Lechene, C. *et al.* High-resolution quantitative imaging of mammalian and bacterial cells using stable isotope mass spectrometry. *J. Biol.* **5,** 20 (2006).
4. Lechene, C. P., Luyten, Y., McMahon, G. & Distel, D. L. Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science* **317,** 1563–1566 (2007).
5. Lark, K. G., Consigli, R. A. & Minocha, H. C. Segregation of sister chromatids in mammalian cells. *Science* **154,** 1202–1205 (1966).
6. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).
7. Potten, C. S., Hume, W. J., Reid, P. & Cairns, J. The segregation of DNA in epithelial stem cells. *Cell* **15,** 899–906 (1978).
8. Potten, C. S., Owen, G. & Booth, D. Intestinal stem cells protect their genome by selective segregation of template DNA strands. *J. Cell Sci.* **115,** 2381–2388 (2002).
9. Shinin, V., Gayraud-Morel, B., Gomes, D. & Tajbakhsh, S. Asymmetric division and cosegregation of template DNA strands in adult muscle satellite cells. *Nature Cell Biol.* **8,** 677–682 (2006).
10. Conboy, M. J., Karasov, A. O. & Rando, T. A. High incidence of non-random template strand segregation and asymmetric fate determination in dividing stem cells and their progeny. *PLoS Biol.* **5,** e102 (2007).
11. Quyn, A. J. *et al.* Spindle orientation bias in gut epithelial stem cell compartments is lost in precancerous tissue. *Cell Stem Cell* **6,** 175–181 (2010).
12. Rando, T. A. The immortal strand hypothesis: segregation and reconstruction. *Cell* **129,** 1239–1243 (2007).
13. Kiel, M. J. *et al.* Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU. *Nature* **449,** 238–242 (2007).
14. Sotiropoulou, P. A., Candi, A. & Blanpain, C. The majority of multipotent epidermal stem cells do not protect their genome by asymmetrical chromosome segregation. *Stem Cells* **26,** 2964–2973 (2008).
15. Schepers, A. G. *et al.* Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *EMBO J.* **30,** 1104–1109 (2011).
16. Lansdorp, P. M. Immortal strands? Give me a break. *Cell* **129,** 1244–1247 (2007).
17. Pech, M. F. & Artandi, S. E. TRAPping telomerase within the intestinal stem cell niche. *EMBO J.* **30,** 986–987 (2011).
18. Cheng, H. & Leblond, C. P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine. V. Unitarian theory of the origin of the four epithelial cell types. *Am. J. Anat.* **141,** 537–561 (1974).
19. Totafurno, J., Bjerknes, M. & Cheng, H. The crypt cycle. Crypt and villus production in the adult intestinal epithelium. *Biophys. J.* **52,** 279–294 (1987).
20. Bjerknes, M. & Cheng, H. Clonal analysis of mouse intestinal epithelial progenitors. *Gastroenterology* **116,** 7–14 (1999).
21. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330,** 822–825 (2010).
22. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143,** 134–144 (2010).
23. Gutierrez, E., Wiggins, D., Fielding, B. & Gould, A. P. Specialized hepatocyte-like cells regulate *Drosophila* lipid metabolism. *Nature* **445,** 275–280 (2007).
24. Schick, P. *et al.* Labelling of human resting lymphocytes by continuous infusion of [^{3}H]thymidine. I. Characterization of cytoplasmic label. *J. Cell Sci.* **33,** 351–362 (1978).
25. Stürup, S., Hansen, H. R. & Gammelgaard, B. Application of enriched stable isotopes as tracers in biological systems: a critical review. *Anal. Bioanal. Chem.* **390,** 541–554 (2008).
26. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449,** 1003–1007 (2007).
27. Sangiorgi, E. & Capecchi, M. R. *Bmi1* is expressed *in vivo* in intestinal stem cells. *Nature Genet.* **40,** 915–920 (2008).
28. Montgomery, R. K. *et al.* Mouse telomerase reverse transcriptase (mTert) expression marks slowly cycling intestinal stem cells. *Proc. Natl Acad. Sci. USA* **108,** 179–184 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Mudge, M. Raff and A. Aperia for critical reading of the manuscript; M. Raff for numerous enlightening discussions; T. Bloom for her strong support at the origin of MIMS development; M. Wang for MIMS analysis; J. Poczatek and Z. Kaufman for MIMS analysis software development; L. Trakimas, C. MacGillivray, S. Clark and E. Hurst for histology; W. Wang for statistics advice. We thank Cambridge Isotope Laboratories for their generous gift of thymidine ($^{15}N_2$, 96–98%). M.L.S. is funded by the American Heart Association and Future Leaders in Cardiovascular Medicine. A.P.G. is funded by the Medical Research Council (U117584237). R.T.L. is funded by the National Institutes of Health (AG032977) and a grant from the Harvard Stem Cell Institute. C.P.L. is funded by the National Institutes of Health (EB001974, AG034641), the Ellison Medical Foundation and the Human Frontier Science Program.

Author Contributions M.L.S. designed the experiments to study the 'immortal strand hypothesis' in the small intestine with input from C.P.L.; M.L.S. performed *in vivo* mouse experiments with help from S.E.S.; A.P.B. and A.P.G. designed and performed the *Drosophila* experiments. M.L.S. designed the human experiment with input from R.T.L. and T.S.P.; M.L.S. and T.S.P. conducted the human protocol. S.E.S. was involved in study design. M.L.S. analysed the data with C.P.L. input. C.G. operated the instrument and assisted with analysis of *Drosophila* lipid droplets. M.L.S. and C.P.L. wrote the manuscript; A.P.B. and A.P.G. contributed the section on *Drosophila*. R.T.L. was involved in study design and provided critical feedback at all junctures. C.P.L. conceived of the general application of MIMS to metabolism, cell turnover and human experimentation. C.P.L. designed and performed *in vitro* experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.P.L. (cpl@harvard.edu).

METHODS

Multi-isotope imaging mass spectrometry (MIMS). MIMS is based on secondary ion mass spectrometry (SIMS). An ionic beam sputters the surface atomic layer of a sample, resulting in the ionization of a small atomic fraction. In a SIMS instrument, secondary ions are separated by mass, and then used to derive a quantitative atomic mass image of the surface of the analysed sample. The instrument can simultaneously measure data from multiple isotopes from the same region. Data included in this study are derived from both the NanoSims prototype (ONERA, Université Paris Sud/Cameca) and a second-generation instrument (NanoSims 50L, Cameca). The second-generation instrument is fully automated and has the capacity to quantify seven masses simultaneously.

Quantitative image data extraction. For each pixel, the number of counts of a given secondary ion is stored in a computer file. The data are reconstructed into a greyscale image in which the pixel intensity is derived from the total number of counts of a given secondary ion within the area representing a given pixel. The lateral resolution is dependent on factors, including the beam size and the number of pixels per image acquisition area. A series of images based on the analysis of successive atomic layers can be summed. The incorporation of halogenated nucleotide analogues (BrdU, iododeoxyuridine) is assessed by directly quantifying the ionized halogen atoms (^{81}Br, ^{127}I), which are found at extremely low background concentrations in most tissues. Halogen enrichment within a given region of interest is expressed as the percent enrichment over natural background.

Hue saturation intensity ratio images. The localization of tracers enriched with rare stable isotopes (for example, ^{15}N-thymidine) is done by deriving quantitative ratio images. The ratio of a rare stable isotope and the common stable isotope is quantified on a pixel by pixel basis. Incorporation of a stable isotope within a region of interest is determined when the ratio of the rare stable isotope to the common stable isotope exceeds the natural ratio. Because quantitative images acquired with MIMS generate far more data than can be represented by a greyscale, we previously developed a method of visually representing ratio data based on a hue saturation intensity (HSI) transformation. With an HSI image, the scale can be set such that the lower bounds of the scale (blue colours) are assigned to pixels in which the corresponding rare-to-common stable isotope ratio is equivalent to the natural ratio. An important advantage of an HSI image is that the scale can be extended or compressed to provide an optimal colour representation of regions of differential stable isotope enrichment. However, any image manipulations, such as those resulting from scale changes, do not modify the quantitative data underlying the colour representation.

Data processing and export. MIMS images are viewed and processed using Open MIMS software, a custom-built plugin to the ImageJ software (http://www.nrims.harvard.edu/software.php). In the experiments contained herein, nuclei were manually identified as regions of interest within the crypts of the small intestine using a combination of ^{14}N and ^{31}P mass images. Cellular identity was assigned by an observer unaware of the labelling status of the cells. Paneth cells were identified on the basis of their characteristic cytoplasmic granules seen in ^{14}N and ^{32}S images, large round nuclei seen in ^{14}N and ^{31}P images, and their anatomic location at or near the base of the crypt (Supplementary Fig. 9). ^{15}N-labelled Paneth cells (Fig. 2 and Supplementary Fig. 8 Experiment 1) had a number of granules with a median of 8 (interquartile range = 5–11, $n = 218$). The granules were not separated from the nucleus by a plasma membrane, which would have been detected by MIMS. CBC cells were identified on the basis of their location between Paneth cells at the base of the crypt and their narrow, elongated nuclei. The +4 position was assigned to the first non-Paneth cell superior to the Paneth zone at the base of the crypt. Cells in the +4 position and above (transit amplifying cells) typically had elliptical nuclei.

^{15}N-labelling is expressed as the ^{15}N:^{14}N ratio (percentage above natural ratio) obtained using the following equation:

$$^{15}\mathrm{N}:^{14}\mathrm{N}(\%\ \text{above natural ratio}) = \frac{(^{15}\mathrm{N}:^{14}\mathrm{N}\ \text{ratio}) - (\text{natural ratio})}{\text{natural ratio}} \times 100$$

Mosaic images. The maximum field size analysed in this study was 80 μm × 80 μm. The presentation of images from larger tissue swaths is made possible by the construction of mosaic images in which a series of adjacent fields are analysed. After MIMS acquisition, the full series of tiles is assembled to form a larger mosaic image.

Statistics. Data were analysed using JMP 8.0.1 and Prism 3.0 (GraphPad). Two-way *t*-tests were performed for comparisons of two experimental groups. For analyses of multiple experimental groups, one-way ANOVA and Bonferroni correction were used. Non-parametric data were analysed with a Kruskal–Wallis test and Dunn's multiple comparisons test. Binomial testing was used to compare observed versus predicted event frequencies.

Reagents. ^{15}N-thymidine and ^{13}C-thymidine were obtained from Cambridge Isotopes. ^{15}N-thymidine, used for human administration, was subjected to additional endotoxin and microbial limits testing before packaging. ^{13}C-palmitate was obtained from Sigma-Aldrich. Bromodeoxyuridine (BrdU) and iododeoxyuridine (IdU) were obtained from Sigma-Aldrich.

Cell culture. Human foreskin fibroblasts were grown in DMEM/Ham's F12 supplemented with 10% FBS, 1% penicillin/streptomycin/L-glutamine. For *in vitro* labelling experiments, fibroblasts were grown on silicon chips. Prior to MIMS analysis, cells were fixed and spun dry (Headway Research Spinner, PWM32) at 3,500 r.p.m. for 2 min.

Mice. C57B6 mice were used in accordance with the Guide for the Use and Care of Laboratory Animals and approved by the Harvard Medical School Standing Committee on Animals. Osmotic minipumps (Alzet) were implanted subcutaneously. Jejunal segments were fixed (4% paraformaldehyde) and embedded (LR white). Sections (0.5 μm) were mounted on silicon chips.

Drosophila. *Drosophila* larvae were fed a standard yeast/cornmeal diet[23] supplemented with 1% w/v ^{13}C-labelled palmitic acid-1 (Sigma-Aldrich). For chase experiments, larvae were collected at 72 h of development, following a label-free chase period of 0–24 h, using standard food supplemented with 1% unlabelled palmitic acid (Sigma-Aldrich). Following dissection, anterior midguts or fat bodies were fixed in 2% formaldehyde, post-fixed in 2% glutaraldehyde then 2% osmium tetroxide, embedded in Epon resin, sectioned (thickness = 2 μm), and mounted on silicon chips.

Human protocol. The Partners Institutional Review Board approved the protocol. After obtaining informed consent, the subject was admitted to the Brigham and Women's Hospital clinical research unit. Prior to administration, ^{15}N-thymidine was resuspended in 500 ml 0.9% NaCl and passed through a 0.2 μm filter. The subject received ^{15}N-thymidine intravenously (30 mg bolus over 5 min, then 15 mg h^{-1} for 48 h) without adverse events. Peripheral blood was collected in an EDTA-containing tube at the conclusion of the infusion and after a 4-week chase. The blood sample was centrifuged (300*g*, 5 min), the buffy coat collected, and red blood cells lysed (ACK buffer, Invitrogen). The white blood cell suspension was smeared onto silicon chips and fixed with 4% paraformaldehyde.

LETTER

doi:10.1038/nature10738

Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing

Li Ding[1,2]*, Timothy J. Ley[1,3,4]*, David E. Larson[1], Christopher A. Miller[1], Daniel C. Koboldt[1], John S. Welch[3], Julie K. Ritchey[3], Margaret A. Young[3], Tamara Lamprecht[3], Michael D. McLellan[1], Joshua F. McMichael[1], John W. Wallis[1,2], Charles Lu[1], Dong Shen[1], Christopher C. Harris[1], David J. Dooling[1,2], Robert S. Fulton[1,2], Lucinda L. Fulton[1,2], Ken Chen[1,2], Heather Schmidt[1], Joelle Kalicki-Veizer[1], Vincent J. Magrini[1,2], Lisa Cook[1], Sean D. McGrath[1], Tammi L. Vickery[1], Michael C. Wendl[1,2], Sharon Heath[3], Mark A. Watson[5], Daniel C. Link[3,4], Michael H. Tomasson[3,4], William D. Shannon[6], Jacqueline E. Payton[5], Shashikant Kulkarni[2,4,5], Peter Westervelt[3,4], Matthew J. Walter[3,4], Timothy A. Graubert[3,4], Elaine R. Mardis[1,2,4], Richard K. Wilson[1,2,4] & John F. DiPersio[3,4]

Most patients with acute myeloid leukaemia (AML) die from progressive disease after relapse, which is associated with clonal evolution at the cytogenetic level[1,2]. To determine the mutational spectrum associated with relapse, we sequenced the primary tumour and relapse genomes from eight AML patients, and validated hundreds of somatic mutations using deep sequencing; this allowed us to define clonality and clonal evolution patterns precisely at relapse. In addition to discovering novel, recurrently mutated genes (for example, *WAC*, *SMC3*, *DIS3*, *DDX41* and *DAXX*) in AML, we also found two major clonal evolution patterns during AML relapse: (1) the founding clone in the primary tumour gained mutations and evolved into the relapse clone, or (2) a subclone of the founding clone survived initial therapy, gained additional mutations and expanded at relapse. In all cases, chemotherapy failed to eradicate the founding clone. The comparison of relapse-specific versus primary tumour mutations in all eight cases revealed an increase in transversions, probably due to DNA damage caused by cytotoxic chemotherapy. These data demonstrate that AML relapse is associated with the addition of new mutations and clonal evolution, which is shaped, in part, by the chemotherapy that the patients receive to establish and maintain remissions.

To investigate the genetic changes associated with AML relapse, and to determine whether clonal evolution contributes to relapse, we performed whole-genome sequencing of primary tumour–relapse pairs and matched skin samples from eight patients, including unique patient identifier (UPN) 933124, whose primary tumour mutations were previously reported[3]. Informed consent explicit for whole-genome sequencing was obtained for all patients on a protocol approved by the Washington University Medical School Institutional Review Board. We obtained >25× haploid coverage and >97% diploid coverage for each sample (Supplementary Table 1 and Supplementary Information). These patients were from five different French–American–British haematological subtypes, with elapsed times of 235–961 days between initial diagnosis and relapse (Supplementary Table 2a, b).

Candidate somatic events in the primary tumour and relapse genomes were identified[4,5] and selected for hybridization capture-based validation using methods described in Supplementary Information. Deep sequencing of the captured target DNAs from skin (the matched normal tissue), primary tumour and relapse tumour specimens[6] (Supplementary Table 3) yielded a median of 590-fold coverage per site. The average number of mutations and structural variants was 539 (range 118–1,292) per case (Fig. 1a).

The general approach for relapse analysis is exemplified by the first sequenced case (UPN 933124). A total of 413 somatic events from tiers 1 to 3 were validated (see ref. 7 for tier designations; Supplementary Fig. 1a and Supplementary Tables 4a and 5). Of these, 78 mutations were relapse-specific (63 point mutations, 1 dinucleotide mutation, 13 indels and 1 translocation; relapse-specific criteria described in Supplementary Information and shown in Supplementary Fig. 1b), 5 point mutations were primary-tumour-specific, and 330 (317 point mutations and 13 indels) were shared between the primary tumour and relapse samples (Fig. 1a, b and Supplementary Fig. 2). The skin sample was contaminated with leukaemic cells for this case (peripheral white blood cell count was 105,000 cells mm^{-3} when the skin sample was banked), with an estimated tumour content in the skin sample of 29% (Supplementary Information). In addition to the ten somatic non-synonymous mutations originally reported for the primary tumour sample[3], we identified one deletion that was not detected in the original analysis (*DNMT3A* L723fs (ref. 8)) and three mis-sense mutations previously misclassified as germline events (*SMC3* G662C, *PDXDC1* E421K and *TTN* E14263K) (Fig. 1b, Table 1 and Supplementary Table 4b).

A total of 169 tier 1 coding mutations (approximately 21 per case) were identified in the eight patients (Table 1 and Supplementary Tables 4b and 6), of which 19 were relapse-specific. In addition to mutations in known AML genes such as *DNMT3A* (ref. 8), *FLT3* (ref. 9), *NPM1* (ref. 10), *IDH1* (ref. 7), *IDH2* (ref. 11), *WT1* (ref. 12), *RUNX1* (refs 13, 14), *PTPRT* (ref. 3), *PHF6* (ref. 15) and *ETV6* (ref. 16) in these eight patients, we also discovered novel, recurring mutations in *WAC*, *SMC3*, *DIS3*, *DDX41* and *DAXX* using 200 AML cases whose exomes were sequenced as part of the Cancer Genome Atlas AML project (Table 1, Supplementary Table 4b and Supplementary Fig. 3; T.J.L., R.K.W. and The Cancer Genome Atlas working group on AML, unpublished data). Details regarding the novel, recurrently mutated genes are provided in Table 1, Supplementary Tables 4b and 7 and Supplementary Figs 3 and 4. Structural and functional analyses of structural variants are presented in the Supplementary Information (Supplementary Figs 5–10 and Supplementary Tables 2, 8 and 9).

The generation of high-depth sequencing data allowed us to quantify accurately mutant allele frequencies in all cases, permitting estimation of the size of tumour clonal populations in each AML sample. On the basis of mutation clustering results, we inferred the identity of four clones having distinct sets of mutations (clusters) in the primary tumour of AML1/UPN 933124 (Supplementary Information). The median mutant allele frequencies in the primary tumour for clusters 1 to 4 were 46.86%, 24.89%, 16.00% and 2.39%, respectively (Fig. 1b and Supplementary Table 5c). Clone 1 is the 'founding' clone (that is, the other subclones are derived from it), containing the cluster 1 mutations; assuming that nearly all of these mutations are heterozygous, they must

[1]The Genome Institute, Washington University, St Louis, Missouri 63108, USA. [2]Department of Genetics, Washington University, St Louis, Missouri 63110, USA. [3]Department of Internal Medicine, Division of Oncology, Washington University, St Louis, Missouri 63110, USA. [4]Siteman Cancer Center, Washington University, St Louis, Missouri 63110, USA. [5]Department of Pathology and Immunology, Washington University, St Louis, Missouri 63110, USA. [6]Division of Biostatistics, Washington University, St Louis, Missouri 63110, USA.
*These authors contributed equally to this work.

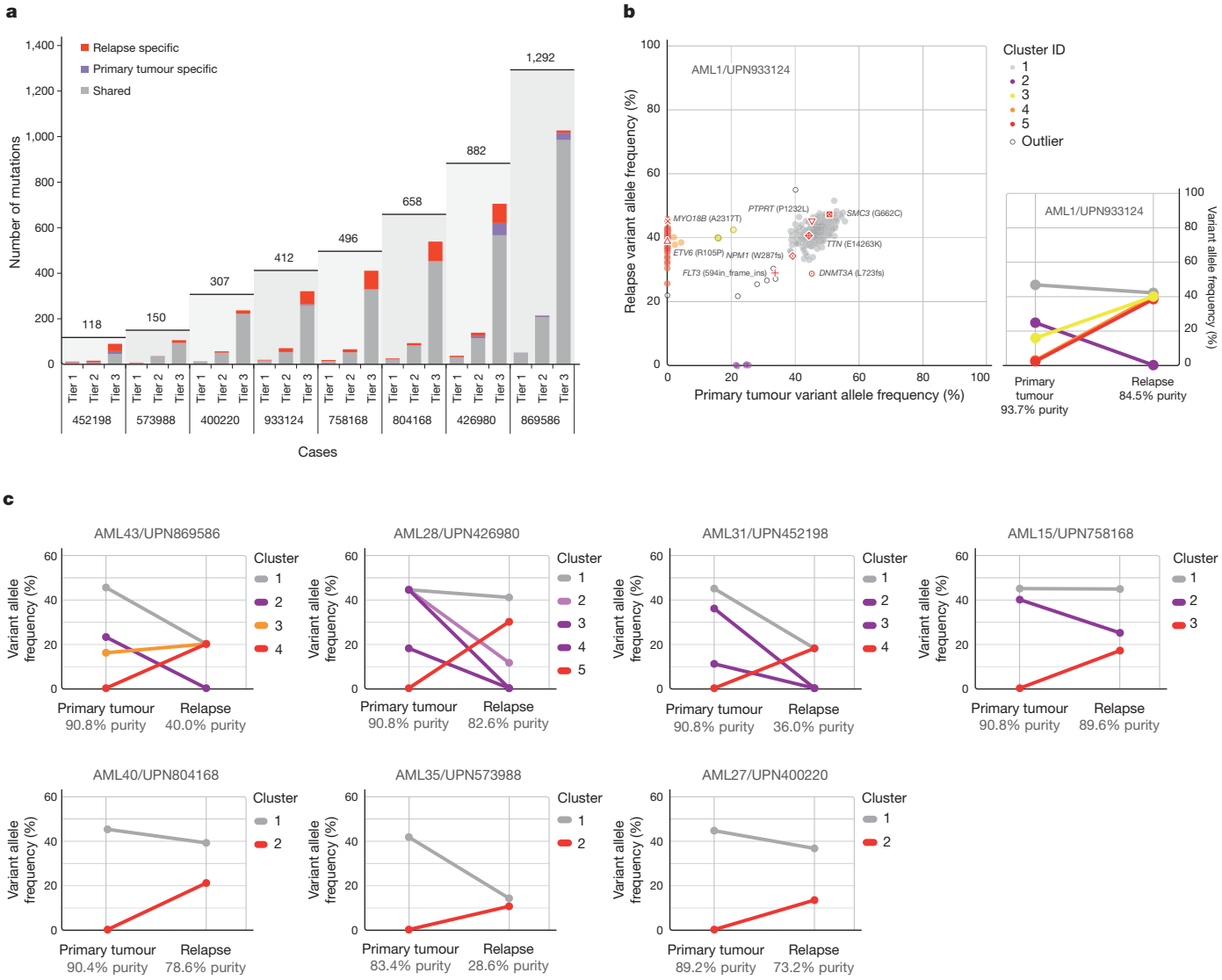


Figure 1 | Somatic mutations quantified by deep sequencing of capture validation targets in eight acute myeloid leukaemia primary tumour and relapse pairs. **a**, Summary of tier 1–3 mutations detected in eight cases (not including translocations). All mutations shown were validated using capture followed by deep sequencing. Shared mutations are in grey, primary tumour-specific mutations in blue and relapse-specific mutations in red. The total number of tier 1–3 mutations for each case is shown above the light-grey rectangle. **b**, Mutant allele frequency distribution of validated mutations from tier 1–3 in the primary tumour and relapse of case UPN 933124 (left). Mutant allele frequencies for five primary-tumour-specific mutations were obtained from a 454 deep read-count experiment. Four mutation clusters were identified in the primary tumour, and one was found at relapse. Five low-level mutations in both the primary tumour and relapse (including four residing in known copy number variable regions) were excluded from the clustering analysis. Non-synonymous mutations from genes that are recurrently mutated in AML are shown. The change of mutant allele frequencies for mutations from the five clusters is shown (right) between the primary tumour and relapse. The orange and red lines are superimposed. **c**, The mutation clusters detected in the primary tumour and relapse samples from seven additional AML patients. The relationship between clusters in the primary tumour and relapse samples are indicated by lines linking them.

Table 1 | Coding mutations identified in eight primary tumour–relapse pairs

UPN	Total tier 1 mutations (primary/relapse)	Recurrently mutated genes in primary tumour	Relapse-specific non-synonymous somatic mutations
452198	9/9	*DNMT3A*, *NPM1*, *FLT3*, *IDH1*	None
573988	6/8	*NPM1*, *IDH2*	*STOX2*
804168	22/26	*FLT3*, *WT1*, *PHF6*, *FAM5C*, *TTC39A*	*SLC25A12*, *RIPK4*, *ABCD2*
933124	14/17	*DNMT3A*, *NPM1*, *FLT3*, *TTN*, *SMC3*, *PTPRT*	*ETV6**, *MYO18B**, *WAC*†*, *STK4*
400220	12/13	*FLT3*, *RUNX1*, *WT1*, *PLEKHH1*	None
426980	32/35	*IDH2*, *MYO1F*, *DDX4*	*GBP4*, *DCLK1*, *IDH2**, *DCLK1**, *ZNF260*
758168	15/19	*DNAH9*, *DIS3*, *CNTN5*, *PML-RARA†*	ENSG00000180144, *DAGLA**
869586	51/50	*RUNX1*, *WT1*, *TTN*, *PHF6*, *NF1*, *SUZ12*, *NCOA7*, *EED*, *DAXX*, *ACSS3*, *WAC*, *NUMA1*	None

Tier 1 mutation counts exclude RNA genes.
* Recurrent mutations occurring in relapse sample.
† Translocations were not included in tier 1 mutation counts.

be present in virtually all the tumour cells at presentation and at relapse, as the variant frequency of these mutations is ~40–50%. Clone 2 (with cluster 2 mutations) and clone 3 (with cluster 3 mutations) must be derived from clone 1, because virtually all the cells in the sample contain the cluster 1 mutations (Fig. 2a). It is likely that a single cell from clone 3 gained a set of mutations (cluster 4) to form clone 4: these survived chemotherapy and evolved to become the dominant clone at relapse. We do not know whether any of the cluster 4 mutations conferred chemotherapy resistance; although none had translational consequences, we cannot rule out a relevant regulatory mutation in this cluster.

Assuming that all the mutations detected are heterozygous in the primary tumour sample (with a malignant cellular content at 93.72% for the primary bone marrow sample, see Supplementary Information), we were able to calculate the fraction of total malignant cells in each clone. Clone 1 is the founding clone; 12.74% of the tumour cells contain only this set of mutations. Clones 2, 3 and 4 evolved from clone 1. The additional mutations in clones 2 and 3 may have provided a growth or survival advantage, as 53.12% and 29.04% of the tumour cells belonged to these clones, respectively. Only 5.10% of the tumour cells were from clone 4, indicating that it may have arisen last (Fig. 2a). However, the relapse clone evolved from clone 4. A single clone containing all of the cluster 5 mutations was detected in the relapse sample; clone 5 evolved from clone 4, but gained 78 new somatic alterations after sampling at day 170. As all mutations in clone 5 appear to be present in all relapse tumour cells, we suspect that one or more of the mutations in this clone provided a strong selective advantage that contributed to relapse. The *ETV6* mutation, the *MYO18B* mutation, and/or the *WNK1-WAC* fusion are the most likely candidates, as *ETV6*, *MYO18B* and *WAC* are recurrently mutated in AML.

We evaluated the mutation clusters in the seven additional primary tumour–relapse pairs by assessing peaks of allele frequency using kernel density estimation (Supplementary Fig. 11 and Supplementary Information). We thus inferred the numbers and malignant fractions of clones in each primary tumour and relapse sample. Similar to UPN 933124, multiple mutation clusters (2–4) were present in each of the primary tumours from four patients (UPN 869586, UPN 426980, UPN 452198 and UPN 758168). However, only one major cluster was detected in each of the primary tumours from the three other patients (UPN 804168, UPN 573988 and UPN 400220) (Fig. 1c and Supplementary Table 10). Importantly, all eight patients gained relapse-specific mutations, although the number of clusters in the relapse samples varied (Fig. 1).

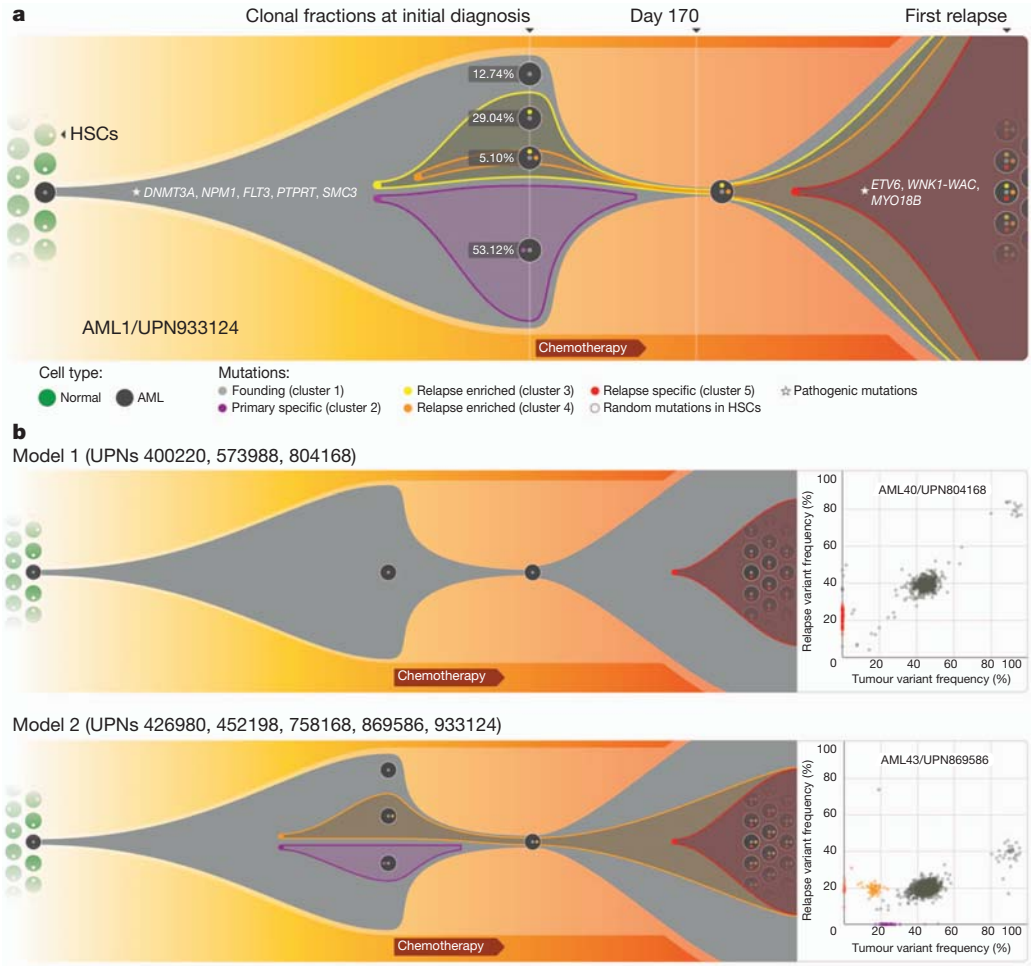


Figure 2 | Graphical representation of clonal evolution from the primary tumour to relapse in UPN 933124, and patterns of tumour evolution observed in eight primary tumour and relapse pairs. **a**, The founding clone in the primary tumour in UPN 933124 contained somatic mutations in *DNMT3A*, *NPM1*, *PTPRT*, *SMC3* and *FLT3* that are all recurrent in AML and probably relevant for pathogenesis; one subclone within the founding clone evolved to become the dominant clone at relapse by acquiring additional mutations, including recurrent mutations in *ETV6* and *MYO18B*, and a *WNK1-WAC* fusion gene. HSC, haematopoietic stem cell. **b**, Examples of the two major patterns of tumour evolution in AML. Model 1 shows the dominant clone in the primary tumour evolving into the relapse clone by gaining relapse-specific mutations; this pattern was identified in three primary tumour and relapse pairs (UPN 804168, UPN 573988 and UPN 400220). Model 2 shows a minor clone carrying the vast majority of the primary tumour mutations survived and expanded at relapse. This pattern was observed in five primary tumour and relapse pairs (UPN 933124, UPN 452198, UPN 758168, UPN 426980 and UPN 869586).

Two major patterns of clonal evolution were detected at relapse (Fig. 2b and Supplementary Fig. 3): in cases with pattern 1, the dominant clone in the primary tumour gained additional mutations and evolved into the relapse clone (UPN 804168, UPN 573988 and UPN 400220). These patients may simply be inadequately treated (for example, elderly patients who cannot tolerate aggressive consolidation, like UPN 573988), or they may have mutations in their founding clones (or germline variants) that make these cells more resistant to therapy (UPN 804168 and UPN 400220). In patients with pattern 2, a minor subclone carrying the vast majority (but not all) of the primary tumour mutations survived, gained mutations, and expanded at relapse; a subset of primary tumour mutations was often eradicated by therapy, and were not detected at relapse (UPN 758168, UPN 933124, UPN 452198, UPN 426980 and UPN 869586). Specific mutations in a key subclone may contribute to chemotherapy resistance, or the mutations important for relapse may be acquired during tumour evolution, or both. Notably, in cases 426980 and 758168, a second primary tumour clone survived chemotherapy and was also present at relapse (Fig. 1c and Supplementary Fig. 3). Owing to current technical limits in our ability to detect mutations in rare cells (mostly related to currently achievable levels of coverage with whole genome sequencing), our models represent a minimal estimate of the clonal heterogeneity in AML.

All eight patients received cytarabine and anthracycline for induction therapy, and additional cytotoxic chemotherapy for consolidation; treatment histories are summarized in Supplementary Table 2 and described in Supplementary Information. To investigate the potential impact of treatment on relapse mutation types, we compared the six classes of transition and transversion mutations in the primary tumour with the relapse-specific mutations in all eight patients (Fig. 3a). Although C•G→T•A transitions are the most common mutations found in both primary and relapse AML genomes, their frequencies are significantly different between the primary tumour mutations (51.1%) and relapse-specific mutations (40.5%) ($P = 2.99 \times 10^{-7}$). Moreover, we observed an average of 4.5%, 5.3% and 4.2% increase in A•T→C•G ($P = 9.13 \times 10^{-7}$), C•G→A•T ($P = 0.00312$) and C•G→G•C ($P = 0.00366$) transversions, respectively, in relapse-specific mutations. Notably, an increased A•T→C•G transversion rate has also been observed in cases of chronic lymphocytic leukaemia with mutated immunoglobulin genes[17]. C•G→A•T transversions are the most common mutation in lung cancer patients who were exposed to tobacco-borne carcinogens[18] (Fig. 3b and Supplementary Table 11). We examined the 456 relapse-specific mutations and 3,590 primary tumour point mutations from all eight cases as a group, and found that the transversion frequency is significantly higher for relapse-specific mutations (46%) than for primary tumour mutations (30.7%) ($P = 3.71 \times 10^{-11}$), indicating that chemotherapy has a substantial effect on the mutational spectrum at relapse. Similar results were obtained when we limited the analysis to the 213 mutations that had 0% variant frequency in the primary tumour samples (Supplementary Fig. 1b); the transversion frequency for relapse-specific mutations was 50.4%, versus 31.4% for primary tumour samples ($P = 3.89 \times 10^{-9}$). Very few copy-number alterations were detected in the eight relapse samples, suggesting that the increased transversion rate is not associated with generalized genomic instability (Supplementary Fig. 12).

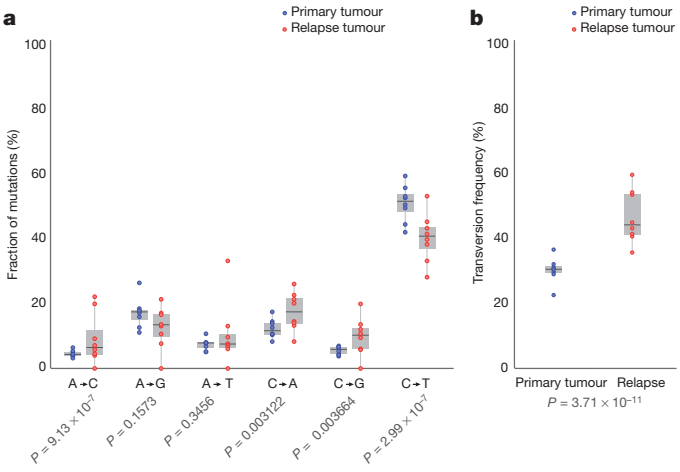


Figure 3 | Comparison of mutational classes between primary tumours and relapse samples. **a**, Fraction of the primary tumour and relapse-specific mutations in each of the transition and transversion categories. **b**, Transversion frequencies of the primary tumour and relapse-specific mutations from eight AML tumour and relapse pairs. 456 relapse-specific mutations and 3,590 primary tumour mutations from eight cases were used for assessing statistical significance using proportion tests.

We first described the use of deep sequencing to define precisely the variant allele frequencies of the mutations in the AML genome of case 933124 (ref. 3), and here have refined and extended this technique to examine clonal evolution at relapse. The analysis of eight primary AML and relapse pairs has revealed unequivocal evidence for a common origin of tumour subpopulations; a dominant mutation cluster representing the founding clone was discovered in the primary tumour sample in all cases. The relationship of the founding clone (and subclones thereof) to the 'leukaemia initiating cell' is not yet clear—purification of clonal populations and functional testing would be required to establish this relationship. We observed the loss of primary tumour subclones at relapse in four of eight cases, suggesting that some subclones are indeed eradicated by therapy (Figs 1 and 2 and Supplementary Fig. 3). Some mutations gained at relapse may alter the growth properties of AML cells, or confer resistance to additional chemotherapy. Regardless, each tumour displayed clear evidence of clonal evolution at relapse and a higher frequency of transversions that were probably induced by DNA damage from chemotherapy. Although chemotherapy is required to induce initial remissions in AML patients, our data also raise the possibility that it contributes to relapse by generating new mutations in the founding clone or one of its subclones, which then can undergo selection and clonal expansion. These data demonstrate the critical need to identify the disease-causing mutations for AML, so that targeted therapies can be developed that avoid the use of cytotoxic drugs, many of which are mutagens.

This study extends the findings of previous studies[19–21], which recently described patterns of clonal evolution in ALL patients using fluorescence *in situ* hybridization and/or copy number alterations detected by SNP arrays, and it enhances the understanding of genetic changes acquired during disease progression, as previously described for breast and pancreatic cancer metastases[22–25]. Our data provide complementary information on clonal evolution in AML, using a much larger set of mutations that were quantified with deep sequencing; this provides an unprecedented number of events that can be used to define precisely clonal size and mutational evolution at relapse. Both ALL and AML share common features of clonal heterogeneity at presentation followed by dynamic clonal evolution at relapse, including the addition of new mutations that may be relevant for relapse pathogenesis. Clonal evolution can also occur after allogeneic transplantation (for example, loss of mismatched HLA alleles via a uniparental disomy mechanism), demonstrating that the type of therapy itself can affect clonal evolution at relapse[26,27]. Taken together, these data demonstrate that AML cells routinely acquire a small number of additional mutations at relapse, and suggest that some of these mutations may contribute to clonal selection and chemotherapy resistance. The AML genome in an individual patient is clearly a 'moving target'; eradication of the founding clone and all of its subclones will be required to achieve cures.

METHODS SUMMARY

Illumina paired-end reads were aligned to NCBI build36 using BWA 0.5.5 (http://sourceforge.net/projects/bio-bwa/). Somatic mutations were identified using SomaticSniper[28] and a modified version of the SAMtools indel caller. Structural

variations were identified using BreakDancer[5]. All predicted non-repetitive somatic SNVs, indels and all structural variants were included on custom sequence capture arrays from Roche Nimblegen. Illumina 2 × 100-bp paired-end sequencing reads were produced after elution from capture arrays. VarScan[6] and a read remapping strategy using Crossmatch (P. Green, unpublished data) and BWA were used for determining the validation status of predicted SNVs, indels and structural variants. A complete description of the materials and methods is provided in Supplementary Information. All sequence variants for the AML tumour samples from eight cases have been submitted to dbGaP (accession number phs000159.v4.p2).

Received 29 March; accepted 29 November 2011.
Published online 11 January 2012.

1. Testa, J. R., Mintz, U., Rowley, J. D., Vardiman, J. W. & Golomb, H. M. Evolution of karyotypes in acute nonlymphocytic leukemia. *Cancer Res.* **39,** 3619–3627 (1979).
2. Garson, O. M. *et al.* Cytogenetic studies of 103 patients with acute myelogenous leukemia in relapse. *Cancer Genet. Cytogenet.* **40,** 187–202 (1989).
3. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
5. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6,** 677–681 (2009).
6. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).
7. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–1066 (2009).
8. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).
9. Nakao, M. *et al.* Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. *Leukemia* **10,** 1911–1918 (1996).
10. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.* **352,** 254–266 (2005).
11. Ward, P. S. *et al.* The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17,** 225–234 (2010).
12. King-Underwood, L., Renshaw, J. & Pritchard-Jones, K. Mutations in the Wilms' tumor gene WT1 in leukemias. *Blood* **87,** 2171–2179 (1996).
13. Gao, J. *et al.* Isolation of a yeast artificial chromosome spanning the 8;21 translocation breakpoint t(8;21)(q22;q22.3) in acute myelogenous leukemia. *Proc. Natl Acad. Sci. USA* **88,** 4882–4886 (1991).
14. Kirito, K. *et al.* A novel RUNX1 mutation in familial platelet disorder with propensity to develop myeloid malignancies. *Haematologica* **93,** 155–156 (2008).
15. Van Vlierberghe, P. *et al.* PHF6 mutations in adult acute myeloid leukemia. *Leukemia* **25,** 130–134 (2011).
16. Barjesteh van Waalwijk van Doorn-Khosrovani, S. *et al.* Somatic heterozygous mutations in ETV6 (TEL) and frequent absence of ETV6 protein in acute myeloid leukemia. *Oncogene* **24,** 4129–4137 (2005).
17. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
18. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
19. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322,** 1377–1380 (2008).
20. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469,** 356–361 (2011).
21. Notta, F. *et al.* Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. *Nature* **469,** 362–367 (2011).
22. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
23. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).
24. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
25. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
26. Vago, L. *et al.* Loss of mismatched HLA in leukemia after stem-cell transplantation. *N. Engl. J. Med.* **361,** 478–488 (2009).
27. Villalobos, I. B. *et al.* Relapse of leukemia with loss of mismatched HLA resulting from uniparental disomy after haploidentical hematopoietic stem cell transplantation. *Blood* **115,** 3158–3161 (2010).
28. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Analysis Pipeline group for developing the automated sequence analysis pipelines; the LIMS group for developing tools and software to manage samples and sequencing; the Systems group for providing the IT infrastructure and HPC solutions required for sequencing and analysis; and R. T. Demeter for experimental support. We also thank The Cancer Genome Atlas for allowing us to use unpublished data for this study, and the Washington University Cancer Genome Initiative for their support. This work was funded by grants to R.K.W. and the National Human Genome Research Institute (NHGRI U54 HG003079), and grants to T.J.L. from the National Cancer Institute (PO1 CA101937) and the Barnes-Jewish Hospital Foundation (00335-0505-02).

Author Contributions T.J.L., L.D., J.F.D., E.R.M. and R.K.W. designed the experiments. L.D. and T.J.L. led data analysis. L.D., D.E.L., C.A.M., D.C.K., J.S.W., M.D.M., J.W.W., C.L., D.S., C.C.H., K.C., H.S., J.K.-V., M.C.W., M.A.W., W.D.S., J.E.P. and S.K. performed data analysis. J.F.M., M.D.M. and L.D. prepared figures and tables. J.S.W., J.K.R., M.A.Y., T.L., R.S.F., L.L.F., V.J.M., L.S., L.C., S.D.M. and T.L.V. performed laboratory experiments. S.H. and P.W. provided samples and clinical data. D.J.D. provided informatics support. T.J.L., D.C.L., M.H.T., E.R.M., R.K.W. and J.F.D. developed project concept. L.D., T.J.L., M.J.W., T.A.G. and J.F.D. wrote the manuscript.

Author Information All sequence variants for the AML tumour samples from eight cases have been submitted to dbGaP under accession number phs000159.v4.p2.
 Correspondence and requests for materials should be addressed to T.J.L. (timley@wustl.edu).

LETTER

doi:10.1038/nature10744

Galectin 8 targets damaged vesicles for autophagy to defend cells against bacterial invasion

Teresa L. M. Thurston[1], Michal P. Wandel[1], Natalia von Muhlinen[1], Ágnes Foeglein[1] & Felix Randow[1]

Autophagy defends the mammalian cytosol against bacterial infection[1–3]. Efficient pathogen engulfment is mediated by cargo-selecting autophagy adaptors that rely on unidentified pattern-recognition or danger receptors to label invading pathogens as autophagy cargo, typically by polyubiquitin coating[4–9]. Here we show in human cells that galectin 8 (also known as LGALS8), a cytosolic lectin, is a danger receptor that restricts *Salmonella* proliferation. Galectin 8 monitors endosomal and lysosomal integrity and detects bacterial invasion by binding host glycans exposed on damaged *Salmonella*-containing vacuoles. By recruiting NDP52 (also known as CALCOCO2), galectin 8 activates antibacterial autophagy. Galectin-8-dependent recruitment of NDP52 to *Salmonella*-containing vesicles is transient and followed by ubiquitin-dependent NDP52 recruitment. Because galectin 8 also detects sterile damage to endosomes or lysosomes, as well as invasion by *Listeria* or *Shigella*, we suggest that galectin 8 serves as a versatile receptor for vesicle-damaging pathogens. Our results illustrate how cells deploy the danger receptor galectin 8 to combat infection by monitoring endosomal and lysosomal integrity on the basis of the specific lack of complex carbohydrates in the cytosol.

Galectins are β-galactoside-binding lectins that accumulate in the cytosol before being secreted via a leader-peptide-independent pathway[10,11]. The best-characterized functions of galectins are performed extracellularly, where they bind glycans to modulate cellular behaviour. However, the occurrence of galectins in the cytosol, which under physiological conditions is devoid of complex carbohydrates, makes them prime candidates for a role as danger and/or pattern-recognition receptors. Galectin 3 (also known as LGALS3) accumulates on damaged bacteria-containing vesicles, although the functional consequences of its recruitment remain unknown[12,13]. We screened a panel of human galectins for their ability to detect invasion by *Salmonella enterica* serovar Typhimurium. At 1 h post-infection (p.i.), galectin 3, 8 and 9 accumulated on about 10% of intracellular *S.* Typhimurium (Fig. 1a, b and Supplementary Fig. 1a), of which 90% were associated with LAMP1 (Supplementary Fig. 1b). Recruitment of galectins peaked between 1 h and 2 h p.i. (Supplementary Fig. 1c). As galectin 3, 8 and 9 were recruited to *Salmonella*-containing vesicles (SCVs), we used short interfering RNAs (siRNAs) to test whether their depletion causes hyperproliferation of *S.* Typhimurium. Cells lacking galectin 8 or NDP52, but not galectin 3 and/or 9, failed to suppress proliferation of *S.* Typhimurium (Fig. 1c and Supplementary Figs 2a–c and 3a). Microscopic analysis confirmed that the greater bacterial burden of cells lacking galectin 8 was caused by enhanced proliferation rather than differential uptake of bacteria (Supplementary Fig. 3b). Hyperproliferating bacteria in cells lacking galectin 8 appeared mainly in a LAMP1-negative compartment (Supplementary Fig. 3c), consistent with colonization of the cytosol. We conclude that galectin 8 is an antibacterial restriction factor.

As autophagy provides antibacterial protection to cells, the decoration of SCVs with galectins might be an autophagy-inducing signal, analogous to ubiquitin coating[14,15]. We therefore tested binding of galectins to autophagy receptors that restrict the proliferation of *S.* Typhimurium, that is, NDP52, p62 and optineurin[7–9]. We found in a luminescence-based mammalian interactome mapping (LUMIER) assay that galectin 8 and NDP52 interacted specifically (Fig. 2a and Supplementary Fig. 4a). Binding was confirmed by precipitating endogenous NDP52 with Flag-tagged galectin 8 (Fig. 2b).

SCVs double labelled by endogenous galectin 8 and NDP52 were prominent in *Salmonella*-infected HeLa cells (Fig. 2c). In cells expressing yellow fluorescent protein (YFP)-tagged galectins the majority of

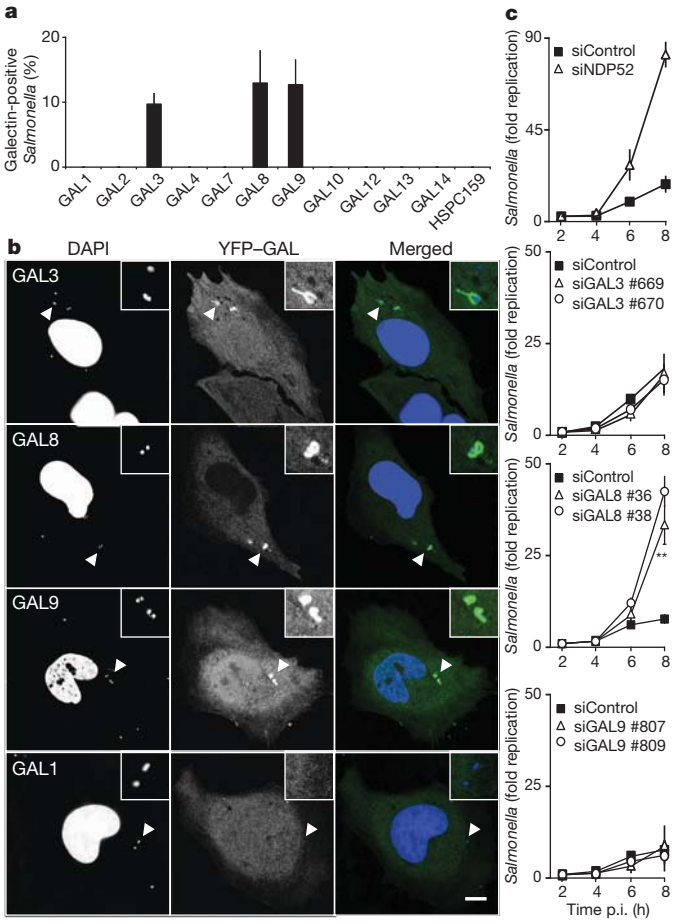


Figure 1 | Galectin 8 responds to infection by *S.* Typhimurium and restricts bacterial proliferation. a, **b**, Analysis of HeLa cells stably expressing YFP fused to the indicated galectins and infected with *S.* Typhimurium for 1 h. **a**, Percentage of bacteria coated by the indicated galectins. YFP-positive bacteria were counted by microscopy. Mean and standard deviation (s.d.) of triplicate HeLa cultures, $n > 100$ bacteria per coverslip. **b**, Confocal micrographs. Arrowheads, bacteria shown in insets. DAPI, 4′,6-diamidino-2-phenylindole. **c**, Kinetics of fold replication for *S.* Typhimurium in HeLa cells transfected with the indicated siRNAs. Bacteria were counted on the basis of their ability to form colonies on agar plates. Mean and s.d. of triplicate HeLa cultures and duplicate colony counts. siRNAs are further characterized in Supplementary Fig. 2a–c. $^{**}P < 0.01$, Student's *t*-test. Scale bar, 10 μm.

[1]MRC Laboratory of Molecular Biology, Division of Protein and Nucleic Acid Chemistry, Hills Road, Cambridge CB2 0QH, UK.

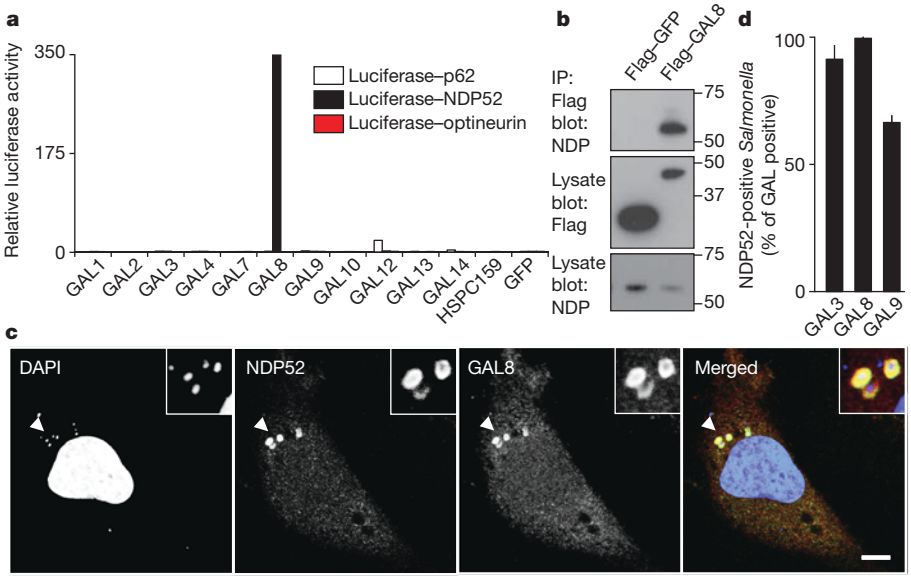


Figure 2 | Galectin 8 binds NDP52. **a**, LUMIER binding assay: normalized ratio between luciferase activity bound to beads and present in lysates. Lysates of 293ET cells expressing NDP52, p62 or optineurin each fused to luciferase, and the indicated Flag-tagged galectins were incubated with anti-Flag beads. Flag-tagged proteins are further characterized in Supplementary Fig. 4a. **b**, Lysates of 293ET cells, expressing Flag-tagged proteins as indicated, were immunoprecipitated with anti-Flag beads. Lysates and immunoprecipitates (IP) were blotted for the presence of Flag-tagged proteins and endogenous NDP52. **c**, Confocal images of HeLa cells infected with *S.* Typhimurium for 1 h and stained with antisera against NDP52 and galectin 8. Arrowheads, bacteria shown in insets. **d**, Co-localization of NDP52 with galectin-positive bacteria in HeLa cells stably expressing YFP fused to the indicated galectins, infected with *S.* Typhimurium and stained with NDP52 antiserum 1 h after infection. Mean and s.d. of duplicate HeLa cultures, $n > 100$ bacteria per coverslip, representative of two independent experiments. Scale bar, 10 μm.

galectin-positive SCVs had accumulated NDP52 (Fig. 2d and Supplementary Fig. 5a). Furthermore, at 1 h p.i. NDP52 and galectin 8 co-localized tightly in a pattern distinct from p62 or ubiquitin 'microdomains'[16] (Supplementary Fig. 5b, c).

To characterize further the interaction between galectin 8 and NDP52 we determined their respective binding sites. Galectin 8 contains two carbohydrate-recognition domains (CRD) (Supplementary Fig. 6a). NDP52 bound galectin-$8_{\Delta 1-228}$ (that is, amino acids 229–360), equivalent to the second CRD, but not galectin-8_{1-228} (that is, amino acids 1–228) (Supplementary Fig. 6b). NDP52 harbours a SKICH domain, a coiled coils-forming region, and a ubiquitin-binding zinc finger (Supplementary Fig. 6a). Galectin 8 bound $NDP52_{1-393}$ but not $NDP52_{1-370}$ (Supplementary Fig. 6c). The NDP52 fragment spanning residues 370–393 is therefore essential for binding galectin 8. This fragment, as well as $NDP52_{372-380}$, purified as GST-fusion proteins, bound galectin 8 (Supplementary Fig. 6d). A point mutation within $NDP52_{372-380}$ (L374A) abrogated binding to galectin 8, without compromising binding to ubiquitin when introduced into full-length NDP52 (Supplementary Fig. 6d, e). Binding of galectin 8 to NDP52 is direct, as the purified proteins interacted (Supplementary Fig. 6f).

To determine whether one monomer of the NDP52–galectin-8 heteromeric complex recruits the other partner, the accumulation of galectins on SCVs in cells depleted of NDP52 or TBK1 was analysed (Fig. 3a and Supplementary Fig. 2). Galectin 3, 8 and 9 re-distributed normally to SCVs in siRNA-treated cells. In contrast, in cells depleted of galectin 8 NDP52 did not localize to SCVs at 1 h p.i., a phenotype that was complemented upon expression of siRNA-resistant galectin 8 (Fig. 3b and Supplementary Fig. 7). Cells lacking galectin 3 or galectin 9 had no defect in recruiting NDP52 to SCVs. The recruitment of NDP52 to SCVs is therefore specifically mediated by galectin 8, while NDP52 and TBK1 are dispensable for the accumulation of galectins on SCVs.

Rupture of SCVs exposes the cytosol to host glycans and microbial carbohydrates, either or both of which may cause galectin 8 accumulation at SCVs. The requirement for carbohydrate binding by galectin 8 was tested using point mutations in either CRD[17]. In contrast to galectin-8(R232H), galectin-8(R69H) did not accumulate at SCVs, showing that the amino-terminal CRD is required for carbohydrate-dependent recruitment of galectin 8 to SCVs (Fig. 3c). To test whether the carbohydrates detected by galectin 8 are of microbial origin, binding of recombinant galectin 8 to bacteria *in vitro* was analysed. Galectin 8 did not bind to *S.* Typhimurium but stained blood-group-B-positive bacteria (*Escherichia coli* strain O86)[17] (Fig. 3d), suggesting that galectin 8, when accumulating on SCVs, recognizes host glycans. The occurrence of galectin-8 ligands in host cells was confirmed by staining HeLa cells with recombinant galectin 8 (Fig. 3d). Direct evidence that host glycans recruit galectins to SCVs was obtained from experiments with CHO-Lec3.2.8.1 cells[18], which lack mature glycans and in which recruitment of galectins to SCVs was severely impaired (Fig. 3e). The detection of host glycans on damaged vesicles by galectin 8 suggests that it is not a receptor specific for *S.* Typhimurium. We therefore tested whether sterile damage to vesicles is detected by galectins. Osmotic damage of endosomes induced dense puncta formed by galectin 3, 8 and 9 but not by galectin 1 (Fig. 3f and Supplementary Fig. 8). Damage to lysosomes by glycyl-L-phenylalanine 2-naphthylamide (GPN) treatment resulted in the initial loss of lysotracker staining, followed by the appearance of galectin 3, 8 and 9 speckles (Supplementary Fig. 9a). In contrast to damaged SCVs and endosomes, burst lysosomes were also detected by galectin 1, suggesting compartment-specific differences in the distribution of galectin ligands. GPN failed to induce speckles of galectin-8(R69H) (Supplementary Fig. 9b), thereby indicating that binding of glycans to the N-terminal CRD of galectin 8 is required to detect lysosomal damage. The capacity of galectin 3, 8 and 9 to detect vesicle damage by binding exposed host glycans suggests their ability to sense the invasion of cells by a wide range of vesicle-damaging pathogens. Indeed, galectin 3, 8 and 9 also accumulated around Gram-positive *Listeria monocytogenes* and Gram-negative *Shigella flexneri* (Supplementary Fig. 10), proving that these galectins detect the invasion of cells by phylogenetically distant bacteria. We conclude that galectin 3, 8 and 9 are danger receptors that sense the exposure of host glycans on ruptured membranes and thereby monitor the integrity of the endosomal/lysosomal compartment.

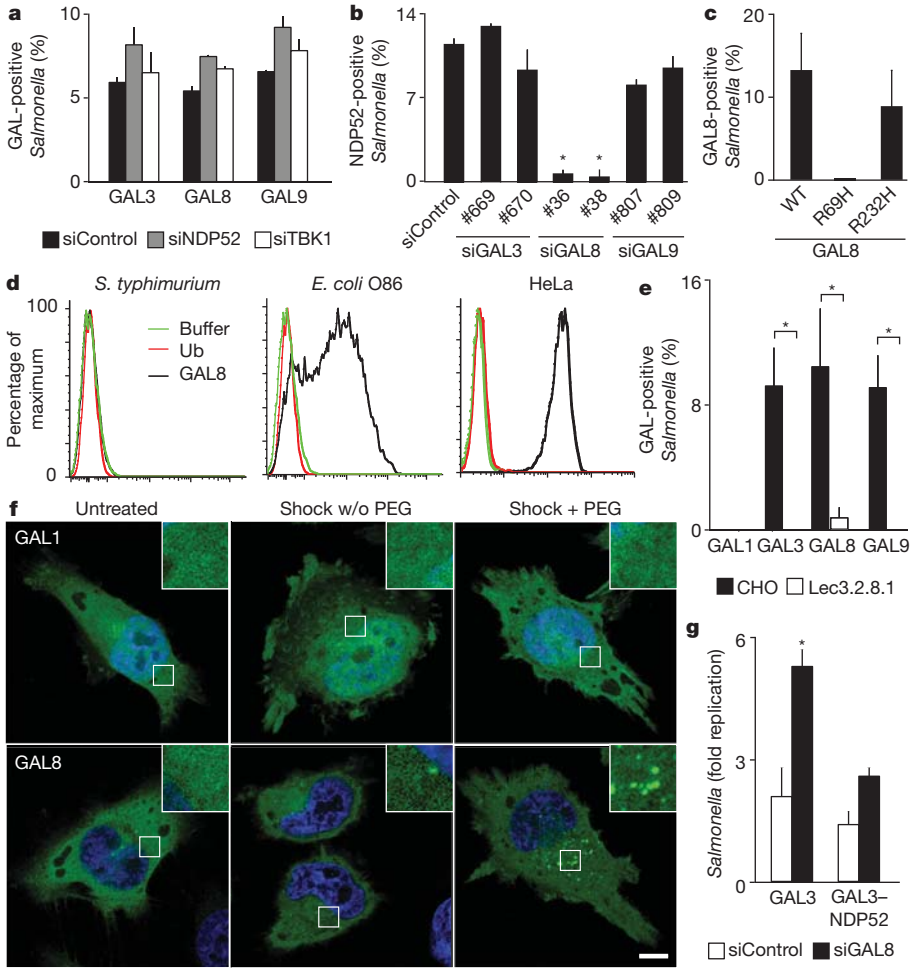


Figure 3 | Galectin 8 is a danger receptor that senses cytosolic host glycans and recruits NDP52 to restrict *Salmonella* proliferation. **a**, Percentage of *S.* Typhimurium coated by the indicated galectins. HeLa cells stably expressing YFP-tagged galectins were treated with the indicated siRNAs. YFP-positive bacteria were counted by microscopy at 1 h p.i. siRNAs are further characterized in Supplementary Fig. 2. **b**, Analysis of HeLa cells treated with the indicated siRNAs and stained with NDP52 antiserum. NDP52-positive bacteria were counted by microscopy 1 h after infection with *S.* Typhimurium. siRNAs are further characterized in Supplementary Fig. 2. **c**, Percentage of bacteria coated by the indicated galectin 8 alleles. HeLa cells stably expressing the indicated galectin 8 alleles fused to YFP were infected with *S.* Typhimurium. YFP-positive *S.* Typhimurium were counted by microscopy at 75 min p.i. WT, wild type. **d**, Binding of galectin 8 to bacteria and HeLa cells. The indicated bacteria and HeLa cells were incubated with His-GST–ubiquitin (Ub), His-GST–galectin 8 or buffer as indicated, followed by murine anti-His antibody and PE-labelled anti-mouse serum. **e**, Percentage of *S.* Typhimurium coated by the indicated galectins. Wild-type CHO cells and mutant Lec3.2.8.1 cells stably expressing YFP-tagged galectins were infected with *S.* Typhimurium. YFP-positive bacteria were counted by microscopy at 1 h p.i. **f**, Confocal images of HeLa cells expressing the indicated YFP-tagged galectins. Cells were left untreated or were exposed to hypertonic conditions, with or without (w/o) PEG as indicated, followed by hypotonic shock. **g**, Fold replication of *S.* Typhimurium in HeLa cells expressing the indicated galectin 3 variants and transfected with the indicated siRNAs. At 2 h and 6 h after infection, cells were lysed and bacteria counted on the basis of their ability to form colonies on agar plates. Galectin 3 proteins are further characterized in Supplementary Fig. 4b. Mean and s.d. of duplicate coverslips (**a**–**c**, **e**) or triplicate HeLa cultures and duplicate colony counts (**g**). >100 bacteria counted per coverslip. Data are representative of at least two repeats. $*P < 0.05$, Student's *t*-test. Scale bar, 10 μm.

To test whether the recruitment of NDP52 to SCVs is essential for the antibacterial function of galectin 8, we depleted cells of galectin 8 and targeted NDP52 artificially to SCVs by fusing it to galectin 3 (Fig. 3g and Supplementary Fig. 4b). *S.* Typhimurium hyperproliferated in galectin-8-depleted cells, while expression of galectin 3 fused to NDP52 restored the cells' restrictive capacity. Artificial targeting of NDP52 to SCVs therefore compensates for the lack of galectin 8, which strongly suggests that the recruitment of NDP52 via galectin 8 is essential to efficiently antagonize bacterial invasion.

NDP52 restricts the proliferation of *S.* Typhimurium by targeting bacteria for autophagy[8]. As galectin 8 recruits NDP52 to SCVs, we investigated whether galectin 8 is required upstream of NDP52 for the induction of antibacterial autophagy. First, we confirmed that at 1 h p.i., *S.* Typhimurium that had been sensed by galectin-8, and therefore had acquired an NDP52 coat, were taken up into LC3-positive autophagosomes (Fig. 4a). Such an outcome was predicted from the pairwise co-localization of NDP52 with galectin-8- and LC3-positive bacteria (Fig. 2d and Supplementary Fig. 11). We then tested whether depletion of galectin 8 impairs autophagy of *S.* Typhimurium. In the absence of galectin 8 fewer bacteria were targeted by LC3 (Fig. 4b) and of the remaining LC3-positive bacteria fewer had accumulated NDP52 (Supplementary Fig. 11). In contrast, galectin-8 recruitment to SCVs did not require autophagy as it occurred undisturbed in *ATG5*$^{-/-}$ fibroblasts (Supplementary Fig. 12). We conclude that the danger receptor galectin 8, by recruiting NDP52, directs autophagy towards invading bacteria.

The recruitment of NDP52 to invading bacteria is mediated by two signals, the newly discovered carbohydrate-dependent galectin-8 pathway and the previously known ubiquitin-dependent pathway[8]. Their differential contribution to the recruitment of NDP52 to *S.* Typhimurium was investigated by analysing NDP52 mutants selectively disabled to bind galectin 8 and/or ubiquitin. For accurate scoring,

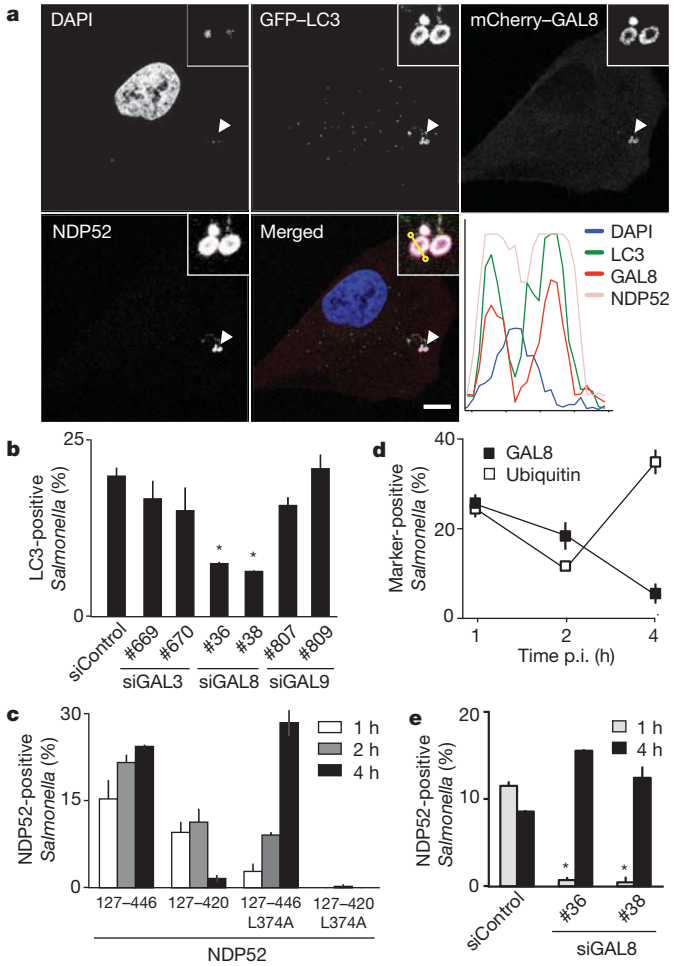


Figure 4 | The antibacterial effect of galectin 8 is mediated by autophagy. **a**, Confocal micrograph of HeLa cells expressing GFP–LC3 and mCherry–galectin-8, stained for NDP52 1 h after infection with *S.* Typhimurium. The lower right panel contains a fluorescence line scan along the yellow line in the merge inset. Arrowheads, bacteria shown in insets. **b**, Percentage of GFP–LC3-positive *S.* Typhimurium at 1 h p.i. in HeLa cells treated with the indicated siRNAs. **c**, Percentage of bacteria positive for NDP52. HeLa cells expressing the indicated NDP52 variants fused to YFP were infected with *S.* Typhimurium. **d**, Percentage of bacteria positive for the indicated markers. HeLa cells, either wild type or expressing YFP–galectin-8 as indicated, were infected with *S.* Typhimurium. Ubiquitin was detected by antibody staining. **e**, HeLa cells treated with the indicated siRNAs were infected with *S.* Typhimurium and stained for NDP52 at the indicated time points. Fluorescent bacteria were counted by microscopy at the indicated time points. Graphs, representing at least two independent repeats, show mean and s.d. of duplicate coverslips for which >200 bacteria were counted. $*P < 0.05$, Student's *t*-test. Scale bar, 10 μm.

NDP52ΔSKICH was used, as this truncated allele is distributed diffusely throughout the cytosol. $NDP52_{127–446}$ associated with SCVs at all time points investigated (Fig. 4c). Deleting the carboxy-terminal ubiquitin-binding zinc finger ($NDP52_{127–420}$) impaired the recruitment of NDP52 to *S.* Typhimurium at late but not early time points. In contrast, $NDP52_{127–446}$(L374A), which lacks affinity for galectin 8 but not ubiquitin (Supplementary Fig. 6e), did not co-localize with bacteria at 1 h p.i. but accumulated progressively over time (Fig. 4c). NDP52 is therefore recruited to SCVs in two phases—an early transient surge driven by galectin 8 and a later wave dependent on ubiquitin. The kinetics of galectin 8 and ubiquitin recruitment to SCVs support this model as SCVs are marked by galectin 8 only at early time points, whereas ubiquitin marks persist (Fig. 4d). Direct evidence for early galectin-8-dependent and late galectin-8-independent recruitment of NDP52 to *S.* Typhimurium was obtained from cells depleted of galectin 8, in which NDP52 and bacteria co-localized at 4 h but not at 1 h p.i. (Fig. 4e). $NDP52_{127–420}$(L374A), deficient in binding to ubiquitin and galectin 8, did not translocate to SCVs at any time point tested (Fig. 4c). Taken together, NDP52 relocates to SCVs in response to two signals, which are active against bacteria at different stages of invasion. The early response to invading bacteria requires the galectin-8-dependent pathway, whereas the zinc-finger-dependent pathway dominates at later time points.

The galectin-8/NDP52 pathway sheds light on why most intracellular bacteria avoid the cytosol and prefer vesicular compartments. The cytosol seems to be protected by synergistic layers of antibacterial defence that activate autophagy at distinct steps of the invasion process. An early line of defence comprises the accumulation of diacylglycerol on bacteria-containing vesicles, which subsequently become the target of autophagy[19]. Bacteria escaping the diacylglycerol pathway and exposing host glycans on their damaged vacuoles are targeted by galectin 8 and NDP52, as described in this work. A third layer of defence coats invading bacteria with polyubiquitin[6–9,20]. Neither the enzymatic machinery for nor the substrate of ubiquitylation have been identified, although LRSAM1, a RING-finger E3 ubiquitin ligase, contributes to autophagy of *S.* Typhimurium[21]. Peptidoglycan and septin cages surrounding cytosolic bacteria also contribute to autophagy[22–25]. Defects in this intricate network of autophagy-inducing defence pathways are likely to cause susceptibility to infection and promote inflammation, for example in Crohn's disease[26–29]. Galectin 8 is positioned strategically at the cellular entry point for a variety of pathogens and is therefore expected to have shaped pathogen evolution.

METHODS SUMMARY

Galectins were cloned as YFP fusions and transduced into HeLa cells. HeLa cells were infected with *S.* Typhimurium strain 12023. For confocal microscopy cells were fixed in paraformaldehyde. Bacterial growth was assessed by a gentamycin protection assay. Knockdowns were accomplished with Stealth siRNAs. LUMIER assays were performed as described[30]. For flow-cytometric analysis samples were incubated with lysates of *E. coli* expressing His-GST fusion proteins, followed by anti-His antibody and goat anti-mouse serum. Statistical testing was performed using two-tailed Student's *t*-test.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 June; accepted 1 December 2011.
Published online 15 January 2012.

1. Yang, Z. & Klionsky, D. J. An overview of the molecular mechanism of autophagy. *Curr. Top. Microbiol. Immunol.* **335,** 1–32 (2009).
2. Deretic, V. Autophagy in immunity and cell-autonomous defense against intracellular microbes. *Immunol. Rev.* **240,** 92–104 (2011).
3. Levine, B., Mizushima, N. & Virgin, H. W. Autophagy in immunity and inflammation. *Nature* **469,** 323–335 (2011).
4. Birmingham, C. L., Smith, A. C., Bakowski, M. A., Yoshimori, T. & Brumell, J. H. Autophagy controls *Salmonella* infection in response to damage to the *Salmonella*-containing vacuole. *J. Biol. Chem.* **281,** 11374–11383 (2006).
5. Johansen, T. & Lamark, T. Selective autophagy mediated by autophagic adapter proteins. *Autophagy* **7,** 279–296 (2011).
6. Perrin, A., Jiang, X., Birmingham, C., So, N. & Brumell, J. Recognition of bacteria in the cytosol of mammalian cells by the ubiquitin system. *Curr. Biol.* **14,** 806–811 (2004).
7. Zheng, Y. T. *et al.* The adaptor protein p62/SQSTM1 targets invading bacteria to the autophagy pathway. *J. Immunol.* **183,** 5909–5916 (2009).
8. Thurston, T. L. M., Ryzhakov, G., Bloor, S., von Muhlinen, N. & Randow, F. The TBK1 adaptor and autophagy receptor NDP52 restricts the proliferation of ubiquitin-coated bacteria. *Nature Immunol.* **10,** 1215–1221 (2009).
9. Wild, P. *et al.* Phosphorylation of the autophagy receptor optineurin restricts *Salmonella* growth. *Science* **333,** 228–233 (2011).
10. Houzelstein, D. *et al.* Phylogenetic analysis of the vertebrate galectin family. *Mol. Biol. Evol.* **21,** 1177–1187 (2004).
11. Rabinovich, G. A. & Toscano, M. A. Turning "sweet" on immunity: galectin–glycan interactions in immune tolerance and inflammation. *Nature Rev. Immunol.* **9,** 338–352 (2009).
12. Paz, I. *et al.* Galectin-3, a marker for vacuole lysis by invasive pathogens. *Cell. Microbiol.* **12,** 530–544 (2009).
13. Dupont, N. *et al. Shigella* phagocytic vacuolar membrane remnants participate in the cellular response to pathogen invasion and are regulated by autophagy. *Cell Host Microbe* **6,** 137–149 (2009).
14. Randow, F. How cells deploy ubiquitin and autophagy to defend their cytosol from bacterial invasion. *Autophagy* **7,** 304–309 (2011).

15. Shahnazari, S. & Brumell, J. H. Mechanisms and consequences of bacterial targeting by the autophagy pathway. *Curr. Opin. Microbiol.* **14,** 68–75 (2011).
16. Cemma, M., Kim, P. K. & Brumell, J. H. The ubiquitin-binding adaptor proteins p62/SQSTM1 and NDP52 are recruited independently to bacteria-associated microdomains to target *Salmonella* to the autophagy pathway. *Autophagy* **7,** 341–345 (2011).
17. Stowell, S. R. *et al.* Innate immune lectins kill bacteria expressing blood group antigen. *Nature Med.* **16,** 295–301 (2010).
18. Patnaik, S. K. & Stanley, P. Lectin-resistant CHO glycosylation mutants. *Methods Enzymol.* **416,** 159–182 (2006).
19. Shahnazari, S. *et al.* A diacylglycerol-dependent signaling pathway contributes to regulation of antibacterial autophagy. *Cell Host Microbe* **8,** 137–146 (2010).
20. Collins, C. A. *et al.* Atg5-independent sequestration of ubiquitinated mycobacteria. *PLoS Pathog.* **5,** e1000430 (2009).
21. Ng, A. C. Y. *et al.* Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proc. Natl Acad. Sci. USA* **108,** 4631–4638 (2011).
22. Mostowy, S. *et al.* Entrapment of intracytosolic bacteria by septin cage-like structures. *Cell Host Microbe* **8,** 433–444 (2010).
23. Yano, T. *et al.* Autophagic control of *Listeria* through intracellular innate immune recognition in *Drosophila*. *Nature Immunol.* **9,** 908–916 (2008).
24. Travassos, L. H. *et al.* Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nature Immunol.* **11,** 55–62 (2010).
25. Cooney, R. *et al.* NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nature Med.* **16,** 90–97 (2010).
26. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).
27. McCarroll, S. A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genet.* **40,** 1107–1112 (2008).
28. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39,** 596–604 (2007).
29. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nature Genet.* **39,** 207–211 (2007).
30. Ryzhakov, G. & Randow, F. SINTBAD, a novel component of innate antiviral immunity, shares a TBK1-binding domain with NAP1 and TANK. *EMBO J.* **26,** 3180–3190 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Kendrick-Jones (MRC Laboratory of Molecular Biology), A. Geerlof (European Molecular Biology Laboratory Heidelberg), N. Mizushima (Tokyo University) and P. Stanley (Albert Einstein College of Medicine) for kindly sharing reagents.

Author Contributions T.L.M.T., M.P.W., N.v.M., Á.F. and F.R. planned, performed and analysed experiments. T.L.M.T. and F.R. designed the overall research. F.R. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.R. (randow@mrc-lmb.cam.ac.uk).

METHODS

Antibodies. Antibodies were from QIAGEN (Penta-His), the Developmental Studies Hybridoma Bank (LAMP1), BD Transduction Laboratories (p62), Santa Cruz (GAL8-H80, TBK1-C100), R&D Systems (galectin 8), Transduction Laboratories (NDP52, for western blots), Enzo Life Science (ubiquitin FK2), Sigma (ATG5, Flag M2), Dabco (HRP-conjugated reagents), Jackson ImmunoResearch Laboratories (goat anti-mouse-phycoerythrin (PE)) and Invitrogen (Alexa-conjugated anti-mouse and anti-rabbit antisera). The antiserum against NDP52 used for immunofluorescence was a gift from J. Kendrick-Jones.

Plasmids. M5P or closely related plasmids were used to produce recombinant MLV for the expression of proteins in mammalian cells[31]. pETM plasmids were gifts from A. Geerlof. Open reading frames encoding human galectins, NDP52, p62, optineurin, ubiquitin, ATG5 and LC3C were amplified by PCR or have been described[8,32]. Mutations were generated by PCR and verified by sequencing.

Bacteria. *S.* Typhimurium (strain 12023), provided by D. Holden, was grown overnight in Luria broth (LB) and sub-cultured (1:33) in fresh LB for 3.5 h before infection. HeLa cells in 24-well plates were infected with 20 µl of such cultures for 15 min at 37 °C. Following two washes with warm PBS and an incubation with 100 µg ml^{-1} gentamycin for 2 h cells were cultured in 20 µg ml^{-1} gentamycin. To enumerate intracellular bacteria, cells from triplicate wells were lysed in 1 ml cold PBS containing 0.1% Triton-X-100. Serial dilutions were plated in duplicate on TYE agar.

S. flexneri M90T, provided by C. Tang, was grown overnight in Tryptic Soy Broth (TSB) and sub-cultured (1:100) in fresh TSB for 2 h before infection. Bacteria were resuspended in warm IMDM and 100 µl were added to HeLa cells in 24-well plates. Samples were centrifuged for 10 min at 670*g*. Following incubation at 37 °C for 30 min, cells were washed with warm PBS and cultured in 100 µg ml^{-1} gentamycin for 2 h and 20 µg ml^{-1} thereafter.

L. monocytogenes strain EGD (BUG 600), provided by P. Cossart, was grown overnight in Brain Heart Infusion (BHI) at 30 °C with shaking. Five-hundred microlitres of diluted cultures (1:333) were added to HeLa cells in 24-well plates, which were centrifuged at 670*g* for 10 min. Cells were incubated for 1 h at 37 °C, washed with warm PBS and cultured in media supplemented with 100 µg ml^{-1} gentamycin for the next hour and 20 µg ml^{-1} gentamycin thereafter.

Cell culture. Cells were grown in IMDM supplemented with 10% FCS at 37 °C in 5% CO_2. HeLa cells were obtained from the European Collection of Cell Cultures, CHO and Lec3.2.8.1 cells[18] were obtained from from P. Stanley, *ATG5*$^{-/-}$ MEFs[33] from N. Mizushima.

RNA interference. 5×10^4 cells per well were seeded in 24-well plates. The following day, cells were transfected with 40 pmol of siRNA (Invitrogen) using Lipofectamine 2000 (Invitrogen) in Optimem medium (Invitrogen). Optimem was replaced with complete IMDM medium after 4 h and experiments were performed after 3 days. siRNAs targeted the following sequences: siNDP52 5′-UUCAGUUGAAGCAGCUCUGUCUCCC[8]; siGAL8 #36 5′-CCCACGCCUG AAUAUUAAAGCAUUU; siGAL8 #38 5′-GGACAAAUUCCAGGUGGCUGU AAAU; siGAL3 #669 5′-AAGCCCAAUGCAAACAGAAUUGCUU; siGAL3 #670 5′-GAGAACAACAGGAGAGUCAUUGUUU; siGAL9 #807 5′-GGCUU CAGUGGAAAUGACAUUGCCU; siGAL9 #809 5′-UGUGCAACACGAGGC AGAACGGAGG; siTBK1 5′-GACAGAAGUUGUGAUCACA(TT)[34].

To render galectin 8 resistant to siGAL8 #38, silent mutations (underlined) GGATAAGTTTCAAGTCGCAGTTAAT were introduced by PCR and confirmed by sequencing.

Immunoprecipitation and western blot. Post-nuclear supernatants from 2×10^6 HeLa cells expressing Flag-tagged proteins were obtained following lysis (150 mM NaCl, 0.1% Triton-X-100, 20 mM Tris-HCl (pH 7.4), 5 mM EDTA and proteinase inhibitors). Protein complexes were immunoprecipitated for 2 h with Flag agarose before washing. Samples were eluted with Flag peptide and separated on 4–12% denaturing Bis-Tris gels (Invitrogen). Visualization following immunoblotting was performed using ECL detection reagents (Amersham Bioscience).

LUMIER assays. LUMIER binding assays[30,35] with pairs of putative interactors, one fused to luciferase and the other fused to GST or Flag, were performed in LUMIER lysis buffer (150 mM NaCl, 0.1% Triton-X-100, 20 mM Tris-HCl (pH 7.4), 5% glycerol, 5 mM EDTA and proteinase inhibitors). GST-fusion proteins were immobilized on beads before incubation with the luciferase tagged binding partner for 2 h. For Flag-based assays, both proteins were expressed in 293ET cells and immobilized using Flag-agarose. After washing in lysis buffer, proteins were eluted with glutathione or Flag peptide in *Renilla* lysis buffer (Promega). Relative luciferase activity represents the ratio of activity eluted from beads and present in lysates.

FACS. To examine the binding of galectin 8, bacteria in stationary phase or HeLa cells were washed in PBSF (PBS, 2% FCS) and incubated for 30 min at 4 °C with cleared lysates of *E. coli* expressing His-GST fusion proteins, followed by incubations with anti-His antibody and PE-conjugated goat anti-mouse serum. Bacteria were fixed in 4% paraformaldehyde before analysis.

Sterile damage to vesicles. Endosomes were lysed by exposing cells for 10 min to hypertonic medium (0.5 M sucrose in PBS, with or without 10% PEG1000), followed by two PBS washes and an incubation in 60% PBS for 3 min[36]. Cells were returned to complete medium for 20 min, before being fixed in paraformaldehyde. For live imaging of lysosomal damage, cells were labelled for 1 h with 100 nM LysoTracker Red (Invitrogen), washed with PBS, incubated in Leibovitz L15 medium and, after acquisition of the first image, exposed to 333 µM GPN[37].

Microscopy. HeLa cells were grown on glass cover slips before infection. After infection, cells were washed twice with warm PBS and fixed in 4% paraformaldehyde in PBS for 30 min. Cells were washed twice in PBS and then quenched with PBS pH 7.4 containing 1 M glycine and 0.1% Triton-X-100 for 30 min before blocking for 30 min in PBTB (PBS, 0.1% Triton-X-100, 2% BSA). Cover slips were incubated with primary followed by secondary antibodies for 1 h in PBTB before being mounting in medium containing DAPI (Vector Laboratories). At least 100 events per slide were scored in quantitative assays. Confocal images were taken with a ×63, 1.4 numerical aperture objective on either a Zeiss 710 or a Zeiss 780 microscope. Live imaging was performed on a Nikon Eclipse Ti equipped with an Andor Revolution XD system and a Yokogawa CSU-X1 spinning disk unit.

31. Randow, F. & Sale, J. E. Retroviral transduction of DT40. *Subcell. Biochem.* **40,** 383–386 (2006).
32. Bloor, S. *et al.* Signal processing by its coil zipper domain activates IKKγ. *Proc. Natl Acad. Sci. USA* **105,** 1279–1284 (2008).
33. Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432,** 1032–1036 (2004).
34. Fitzgerald, K. A. *et al.* IKKε and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunol.* **4,** 491–496 (2003).
35. Barrios-Rodiles, M. *et al.* High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307,** 1621–1625 (2005).
36. Shaughnessy, L. M., Lipp, P., Lee, K.-D. & Swanson, J. A. Localization of protein kinase C ε to macrophage vacuoles perforated by *Listeria monocytogenes* cytolysin. *Cell. Microbiol.* **9,** 1695–1704 (2007).
37. Berg, T. O., Strømhaug, P. E., Berg, T. & Seglen, P. O. Separation of lysosomes and autophagosomes by means of glycyl-phenylalanine-naphthylamide, a lysosome-disrupting cathepsin-C substrate. *Eur. J. Biochem.* **221,** 595–602 (1994).

LETTER

doi:10.1038/nature10745

Multi-isotope imaging mass spectrometry reveals slow protein turnover in hair-cell stereocilia

Duan-Sun Zhang[1]*, Valeria Piazza[1]†*, Benjamin J. Perrin[2], Agnieszka K. Rzadzinska[3], J. Collin Poczatek[4], Mei Wang[4], Haydn M. Prosser[3], James M. Ervasti[2], David P. Corey[1] & Claude P. Lechene[4]

Hair cells of the inner ear are not normally replaced during an animal's life, and must continually renew components of their various organelles[1]. Among these are the stereocilia, each with a core of several hundred actin filaments that arise from their apical surfaces and that bear the mechanotransduction apparatus at their tips. Actin turnover in stereocilia has previously been studied[2] by transfecting neonatal rat hair cells in culture with a β-actin–GFP fusion, and evidence was found that actin is replaced, from the top down, in 2–3 days. Overexpression of the actin-binding protein espin causes elongation of stereocilia within 12–24 hours, also suggesting rapid regulation of stereocilia lengths[3]. Similarly, the mechanosensory 'tip links' are replaced in 5–10 hours after cleavage in chicken and mammalian hair cells[4,5]. In contrast, turnover in chick stereocilia *in vivo* is much slower[6]. It might be that only certain components of stereocilia turn over quickly, that rapid turnover occurs only in neonatal animals, only in culture, or only in response to a challenge like breakage or actin overexpression. Here we quantify protein turnover by feeding animals with a ^{15}N-labelled precursor amino acid and using multi-isotope imaging mass spectrometry to measure appearance of new protein. Surprisingly, in adult frogs and mice and in neonatal mice, *in vivo* and *in vitro*, the stereocilia were remarkably stable, incorporating newly synthesized protein at <10% per day. Only stereocilia tips had rapid turnover and no treadmilling was observed. Other methods confirmed this: in hair cells expressing β-actin–GFP we bleached fiducial lines across hair bundles, but they did not move in 6 days. When we stopped expression of β- or γ-actin with tamoxifen-inducible recombination, neither actin isoform left the stereocilia, except at the tips. Thus, rapid turnover in stereocilia occurs only at the tips and not by a treadmilling process.

To understand protein turnover in the inner ear, we sought a method that could reveal new synthesis with high spatial resolution, in adult animals, *in vivo*, and without transfection of cells. We fed frogs with food containing L-leucine tagged with the stable isotope ^{15}N for up to 32 days, and killed them by cardiac perfusion. As leucine is an essential amino acid, newly synthesized protein would contain ^{15}N in nearly the same proportion as the food. To locate ^{15}N-tagged protein, plastic sections of frog saccules were placed onto silicon wafers or electron microscopy grids for multi-isotope imaging mass spectrometry (MIMS) imaging. We detected nitrogen as the CN^- ion, with the MIMS detectors set for the masses of ^{12}C, ^{13}C, $^{12}C^{14}N$ and $^{12}C^{15}N$, and calculated the $^{15}N/^{14}N$ ratio at each location in a field[7,8] (see Methods).

A mass-26 ($^{12}C^{14}N$) image of the saccular macula, representing total protein, shows the apical surface of a hair cell with its hair bundle (Fig. 1a, left). Lateral resolution approaches 30 nm, and depth resolution is ~2 nm (ref. 7). The mass-27 channel ($^{12}C^{15}N$; not shown) revealed newly synthesized protein, and the mass-27/mass-26 ratio indicated per cent incorporation. A ratio image (Fig. 1a, right, 32 days of feeding) shows that the stereocilia incorporate new protein more slowly than the cell body. We quantified turnover by summing the mass-26 and mass-27 counts for all pixels in each region of interest (ROI) and calculated per cent incorporation (Fig. 1c). Hair-cell cytoplasm and cuticular plates incorporated new protein at ~0.5% per day. But

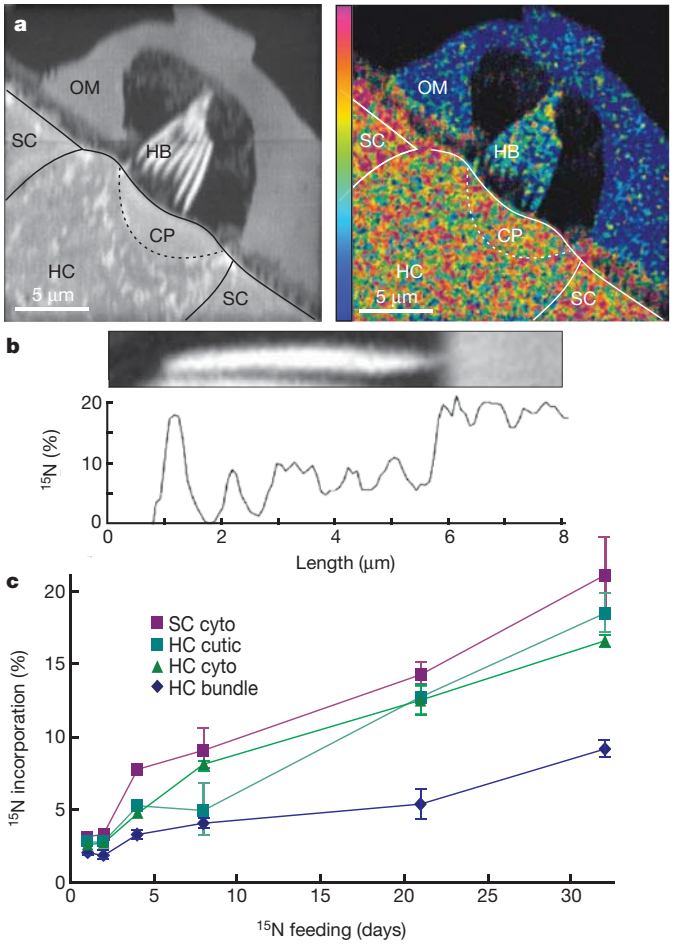


Figure 1 | Incorporation of ^{15}N into frog saccular epithelium. **a**, Left, mass-26 image for a hair cell (HC) and supporting cells (SC) at day 32. CP, cuticular plate; HB, hair bundle; OM, otolithic membrane. Right, image of mass-27/mass-26 ratio, revealing incorporation of ^{15}N. Colour scale represents 0–30% incorporation. Field is 20 × 20 μm. **b**, Top, mass-26 image of the shortest stereocilium in **a**. Bottom, profile of incorporation along its length. **c**, Turnover for days 1, 2, 4, 8, 21 and 32. cutic, cuticular plate; cyto, cytoplasm. Data are shown as mean ± standard error of the mean (s.e.m.); *N* was typically 8–10.

[1]Department of Neurobiology, Harvard Medical School and Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA. [2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA. [3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]National Resource for Imaging Mass Spectrometry, Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Cambridge, Massachusetts 02139, USA. †Present address: Georg August University, D-37077 Göttingen, Germany.
*These authors contributed equally to this work.

stereocilia incorporated new protein at <0.3% per day, far lower than the 50% per day suggested for actin in mammalian cochlea[2]. One difference is that MIMS assesses total protein; perhaps actin turns over rapidly but most other proteins are very stable. However, actin is 50–60% of the total protein in stereocilia[9,10] and would account for the majority of the MIMS signal.

Surprisingly, incorporation seemed higher at the very tips of stereocilia (Fig. 1b). Incorporation at the tip almost equalled that of cytoplasm (Fig. 1c). In this bundle, incorporation at the tips of all stereocilia was 12.5%, compared to 6.1% in the shafts. We also measured incorporation in individual ROIs along the length of a stereocilium. Of 37 frog stereocilia, there was no correlation with height in 31 and positive correlation in just one, inconsistent with movement of the ^{15}N label from top to bottom.

Protein turnover in bullfrog stereocilia was much slower than that inferred for mouse and rat hair cells[2,11], leading us to wonder whether turnover is more rapid in mammals. We fed adult mice ^{15}N-leucine and killed them after 1, 2, 8, 32, 56 or 150 days. In the mouse vestibular system, ^{15}N incorporation was faster in the cytoplasm of hair cells and supporting cells (Fig. 2a, b, e; ~7% per day initially) but still slow in stereocilia (<3% per day initially). This was not a peculiarity of actin-based structures: intestinal-brush-border microvilli showed more incorporation than the cytoplasm, and much more than stereocilia (Supplementary Fig. 2).

The MIMS instrument acquires repeated images, slowly etching the surface to create thousands of XY planes with a Z resolution of approximately 1 nm allowing the creation of a three-dimensional representation (Fig. 2c, d and Supplementary Movies 1 and 2). In mouse utricle, incorporation was again much higher in the tips than elsewhere in stereocilia. Although tip labelling would be expected in the initial phase of treadmilling, the high-turnover region never extended much below the top micrometre, which apparently represents a fixed zone that does not move down by treadmilling.

We then wondered whether turnover is more rapid in cochlear than vestibular hair cells. Apparently not: at 8, 32, 56 and 150 days (Fig. 2f–j), cochlear stereocilia had slower incorporation of new protein (~2%

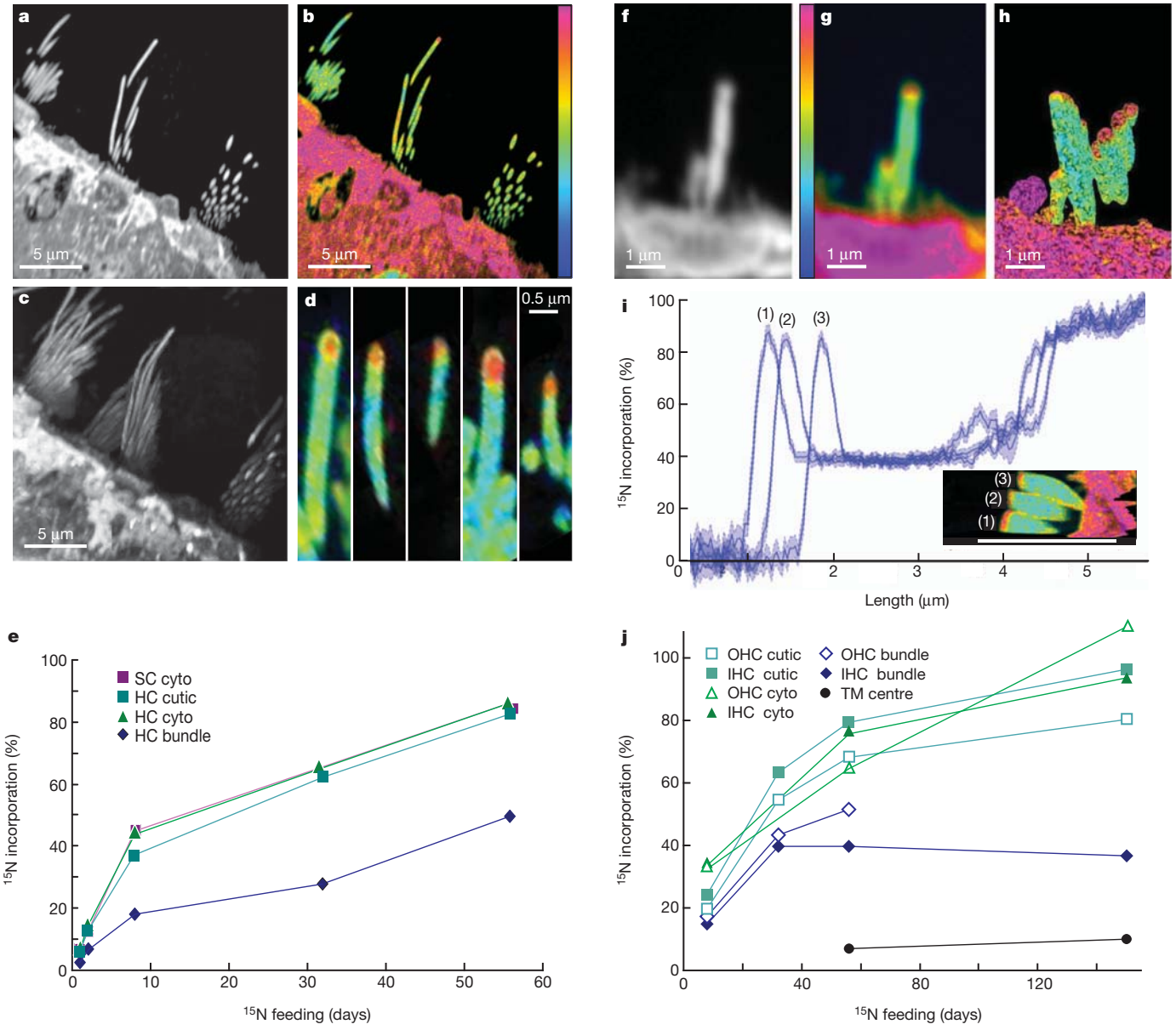


Figure 2 | Incorporation of ^{15}N into adult mouse hair cells. **a**, Mass-26 image of utricle; day 56. **b**, Mass-27/mass-26 ratio: low incorporation in stereocilia. **c**, Projection of a three-dimensional stack of **a**. **d**, Ratio image from **c**: high turnover at tips. **e**, Incorporation, days 1–56. cutic, cuticular plate; cyto, cytoplasm; HC, hair cell; SC, supporting cell. **f**, Cochlear inner hair cell; day 56; mass 26. **g**, Ratio image. **h**, Three-dimensional reconstruction from 450 images. **i**, Incorporation along axis of stereocilia from **h**. Bar: plotted length. Line widths: 95% confidence. **j**, Incorporation after 8, 32, 56 or 150 days. Colour scale 0–100%.

initially) than did cell bodies or cuticular plates (~3.5% initially). Moreover, adult stereocilia apparently have a pool of protein that is replaced exceedingly slowly: ^{15}N incorporation after 5 months was <60%.

A three-dimensional image of cochlear hair bundles (Fig. 2h and Supplementary Movie 3) also showed high turnover in the distal micrometres of the sterocilia. At 56 days, ^{15}N incorporation was 40% along most of the length of individual stereocilia, but >80% at the tips (Fig. 2i). Except for the tips, we did not see a gradient of incorporation. In addition, the resolution of MIMS revealed higher turnover in the region of the cell membrane than in stereocilia cores (Fig. 2i, inset), perhaps reflecting rapid diffusion of some membrane proteins[12].

Elsewhere in the cochlea, ^{15}N incorporation varied considerably (Supplementary Fig. 1). Like the frog otolithic membrane, the mouse tectorial membrane was exceptionally stable, with ~10% ^{15}N incorporation in 5 months. Pillar cell shafts, containing stiff microtubule bundles, had just 23% incorporation. The reticular lamina, comprising the actin-rich cuticular plates of hair cells, terminal webs of Dieter's cells and their zonulae adherens, showed low turnover. Thus, low turnover occurred in structures that convey sound to stereocilia.

Protein turnover could be more rapid in neonatal mice. To test this, mice were moved at birth to surrogate mothers that had been fed ^{15}N for 24–28 days to raise the ^{15}N content of their milk; pups were nursed for 0–15 days before being killed. In a second group, mothers were fed ^{15}N before and during pregnancy. At birth, pups were moved to control nurse mothers, and protein turnover was measured as ^{15}N loss. Pups nursed by ^{15}N mothers showed a normal rise in ^{15}N in utricular cytoplasm (Supplementary Fig. 3), with 30–50% incorporation in 4 days. In stereocilia, protein turnover was slower, with 25% in 4 days (Supplementary Fig. 3b, g). A three-dimensional rendering also showed low turnover in stereocilia, except at the tips (Supplementary Fig. 3c).

A difficulty in interpretation is that about half the hair bundles in the mouse utricle develop postnatally[13], so incorporation of new protein in the stereocilia might reflect development rather than turnover. Indeed, small bundles, presumably still developing, showed high incorporation (Supplementary Fig. 4), and were therefore excluded from the analysis in Supplementary Fig. 3g. Another difficulty is the assumption that tagged leucine is incorporated equally into proteins. Proteins with a lower abundance of leucine than the average (9.36%) would have a lower ^{15}N/^{14}N ratio, which might be interpreted as slow turnover. In fact, stereocilia proteins do have less leucine. The ten most abundant proteins of neonatal rat utricular stereocilia (P. Gillespie, personal communication), adjusted for abundance, have 8.53% leucine. This is not enough to account for low stereocilia incorporation, but we sought a different way to measure turnover.

We therefore measured the loss of ^{15}N in pups born to ^{15}N-fed mothers but nursed by controls, normalizing ^{15}N levels in each structure to that at postnatal day (P)0 (Supplementary Fig. 3h). Consistent with some postnatal bundle development, small bundles had little or no label, but full-sized hair bundles retained ^{15}N for many days, with 70% remaining at P4 (Supplementary Fig. 3d–f, h and Supplementary Fig. 4).

Perhaps turnover in stereocilia is more rapid *in vitro*. We cultured utricles from P0 pups and added ^{15}N-leucine after one day. Over four more days, hair bundles showed relatively low ^{15}N incorporation (~8% per day versus 12–16% per day in cytoplasm; Supplementary Fig. 5).

Perhaps the fluorescence previously observed[2] was due to newly synthesized β-actin–GFP that was only a small fraction of the actin in the core—in essence a tracer for treadmilling, the bulk of which was driven by recycling of actin monomers that are only slowly replaced. To assess treadmilling with an alternative to MIMS, and with steady rather than transient expression of a tracer, we constructed mouse lines expressing β-actin–GFP under the myosin-7a promoter (Supplementary Fig. 6). Stereocilia were brightly fluorescent (Fig. 3). We dissected utricles at ages P0–P5 and cultured them for 1 day. In a confocal microscope, we selected a field in which the epithelium was tilted sideways, and bleached the GFP in a 1-μm-wide line mostly perpendicular to the bundle axis, for 5–8 bundles (Fig. 3b, c). A three-dimensional confocal stack was acquired, and the utricle placed back into culture. Two, four and six days later, each utricle was returned for imaging (Fig. 3d). The bleached line was sometimes obscured if the bundle axis changed in culture, but the three-dimensional image could be used to reconstruct it. In all experiments, we saw no significant movement of the bleached line (Fig. 3f). On average, the line moved by 2 ± 3% of the bundle height in 6 days (Fig. 3g).

It is possible that our cultures were not healthy, and that treadmilling was impaired. However, in the same utricles, we observed the appearance and growth of new bundles. Moreover, both bleached and unbleached hair cells took up the fluorescent dye RH414, a marker for functional transduction (data not shown). Thus an independent method gave no evidence of treadmilling on a timescale of days.

The GFP tag could have inhibited treadmilling in Myo7a::β-actin:GFP mice. We therefore studied treadmilling by terminating β- or γ-actin expression and following the redistribution of actin with antibodies. Mice expressing a tamoxifen-inducible Cre under the β-actin promoter were crossed with mice carrying either a floxed β- or γ-actin gene[14]. At 3

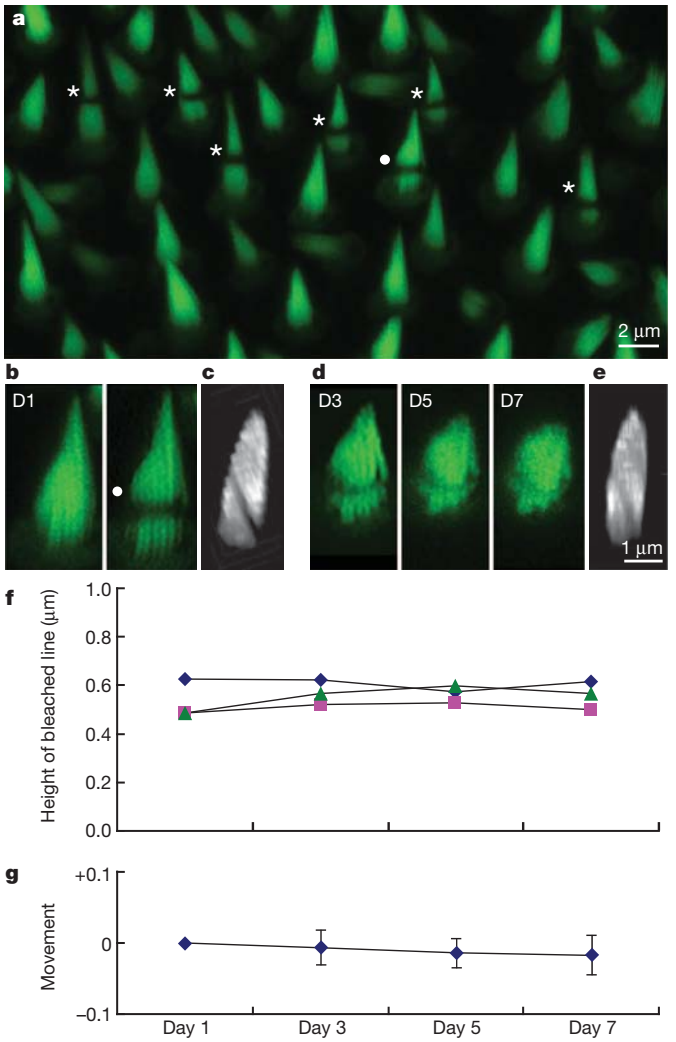


Figure 3 | Tracking treadmilling by bleaching GFP-tagged β-actin. **a**, Neonatal utricular macula in culture, hair cells expressed β-actin–GFP. Bleached bundles are indicated by asterisks. **b**, Projection of confocal planes before and after bleaching. D1, day 1. **c**, Three-dimensional stack rotated to show oblique bleached line. **d**, Same bundle on days (D)3, 5 and 7. A shift in bundle orientation partially obscured the line. **e**, Rotation of the day-7 stack revealed the bleached line. **f**, Line positions for three bundles, as a proportion of bundle height. **g**, Average movement of bleached lines, as a proportion of bundle height. $N = 24$; mean ± s.e.m.

weeks of age, after bundles had developed, mice were given tamoxifen to promote excision. One, three, eighteen or thirty-four weeks later, we fixed and removed cochleas, and used antibodies specific for β- and γ-actin.

In control mice without tamoxifen, β- and γ-actin were uniformly distributed (Fig. 4a). If there is rapid treadmilling in adults, β-actin protein should be lost from stereocilia soon after excision prevented β-actin gene expression. Similarly, γ-actin should be lost in mice with a floxed γ-actin gene. This occurred, but only for a 0.4-μm region at stereocilia tips (Fig. 4a, b). The rest of the stereocilium retained the deleted actin isoform, and the region of isoform loss did not progress towards stereocilia bases, even after 34 weeks.

The MIMS method provides a new way to study protein stability in the inner ear[7]. MIMS is based on Schoenheimer's pioneering use of stable isotopes to reveal the dynamics of cellular components[1], but adds imaging with exceptional spatial resolution. ^{15}N incorporation can be calculated with an accuracy that depends only on acquisition time. It does not rely on tagging any one protein species, or transfecting a specific cell type.

In contrast to previous studies[2], MIMS revealed unexpectedly low turnover of protein in hair-cell stereocilia. Incorporation of ^{15}N in stereocilia occurred at <0.3% per day in frog, <2% per day in adult mouse and <9% per day in neonatal mice, incompatible with treadmilling in 2–3 days. By bleaching fluorescently labelled actin in bundles, we saw no significant movement of the actin. Finally, excision of the β- and γ-actin genes in mature mice showed loss of the excised actin from the stereocilia tips, but not a progressive loss from tip to base. We found no evidence for treadmilling of stereocilia cores, over weeks or even months. It is possible that the previously observed treadmilling was produced by overexpression of β-actin, if the exogenous actin significantly exceeds endogenous actin. Or—at least in the vestibular system—addition of actin to hair bundles developing postnatally might have confused interpretation.

Nevertheless, there is a conspicuously dynamic turnover of protein at the tips of stereocilia, in the distal 0.3–0.5 μm. This could represent turnover of transduction components, which can be replaced in hours when damaged[4,5]. It could represent movement of proteins to the tips by hair-cell myosins, most of which climb from bases to tips. Or it may reflect a rapid regulation of actin filaments, which can change length by ~0.1 μm when tip links are cut[2]. If so, there might be both polymerization and depolymerization at the barbed end. Rapid but local regulation of filament length is not inconsistent with the limited growth of stereocilia induced by overexpression of espin[3].

In long-lived vertebrates, the actin of stereocilia must be replaced. It may be that treadmilling does occur, but at a timescale far longer than that studied here. It may instead be that individual actin filaments do not extend the length of the stereocilium—that free ends can accept or shed monomers and that unsynchronized treadmilling occurs on a submicroscopic scale. Unlike filopodia, stereocilia seem to develop once, elaborating an intricate staircase of heights that persists for the life of an animal; in most of their length, they then replace individual proteins slowly and without disrupting this exquisite structure.

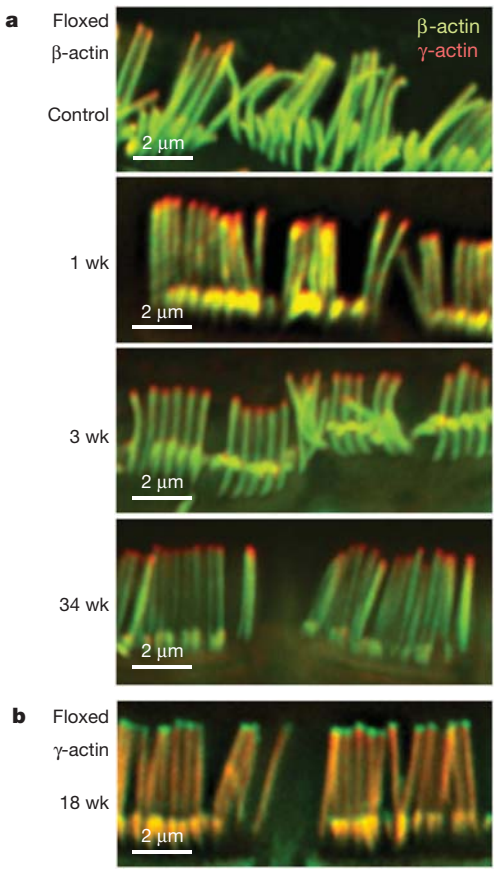


Figure 4 | Loss of actin from tips following knockout of actin genes. **a**, **b**, Adult (3-week-old) cochlear inner hair cells, labelled with antibodies to β-actin (green) or γ-actin (red). **a**, Floxed β-actin mouse crossed to Cagg-CreER, without tamoxifen (control), or 1, 3 or 34 weeks (wk) after tamoxifen. Loss of β-actin occurred only at tips of stereocilia. In control mice, a few hair cells displayed loss of β-actin at tips without tamoxifen, suggesting a leaky promoter. **b**, Floxed γ-actin mouse 18 weeks after tamoxifen. Loss of γ-actin occurred only at tips of stereocilia.

METHODS SUMMARY

Tissue preparation and ^{15}N feeding. Adult bullfrogs (*Rana catesbeiana*) were maintained on a ^{15}N-leucine diet, then fixed and treated with EDTA to dissolve otoconia before embedding and sectioning.

Adult mice (*Mus musculus*, CD1 strain) were maintained on a ^{15}N-leucine diet before being killed and processed. For neonatal mice, mothers were maintained on ^{15}N food before and during pregnancy; at birth, the pups were moved to a control surrogate mother. A second set was born to control mothers and moved at birth to a surrogate mother that had been maintained on ^{15}N food.

Utricles were dissected, cultured overnight in DMEM-F12 with 1 or 6% FBS[16] and maintained in ^{15}N-enriched medium for up to 5 days.

MIMS analysis. MIMS was performed as described previously[7]. ^{15}N incorporation was calculated by comparing counts for masses 27 and 26 to the natural abundance of ^{15}N (0.37%) and the relative abundance of ^{15}N in the food (1–2%).

Photobleaching of cultured mouse utricles. We constructed mice expressing a β-actin–GFP fusion under the myosin-7a promoter, integrated in the HPRT locus on the X chromosome (see Supplementary Fig. 6). Utricles from neonatal pups (mosaic for the transgene) were cultured as above. Hair bundles were bleached on day 1; images were taken on days 3, 5 and 7.

Conditional deletion of β- and γ-actin. *Actg1*^{flox} and *Actb*^{flox} mice, in which exons 3 and 4 were flanked by *loxP* sites[14] were crossed to mice expressing Cre/Esr1 recombinase driven by the chicken β-actin promoter coupled to the cytomegalovirus immediate-early enhancer. Mice were dosed with tamoxifen to induce recombination. Anti-γ-actin antibody clone 1-37 (ref. 15) IgG was conjugated to Alexa fluorescent dyes. FITC-labelled anti-β-actin antibodies (clone AC-15) were obtained from Abcam.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 July; accepted 1 December 2011.
Published online 15 January 2012.

1. Schoenheimer, R. *The Dynamic State of Body Constituents* (Harvard Univ. Press, 1942).
2. Rzadzinska, A. K. *et al.* An actin molecular treadmill and myosins maintain stereocilia functional architecture and self-renewal. *J. Cell Biol.* **164,** 887–897 (2004).
3. Rzadzinska, A. *et al.* Balanced levels of espin are critical for stereociliary growth and length maintenance. *Cell Motil. Cytoskeleton* **62,** 157–165 (2005).
4. Zhao, Y., Yamoah, E. N. & Gillespie, P. G. Regeneration of broken tip links and restoration of mechanical transduction in hair cells. *Proc. Natl Acad. Sci. USA* **93,** 15469–15474 (1996).
5. Jia, S., Yang, S., Guo, W. & He, D. Z. Fate of mammalian cochlear hair cells and stereocilia after loss of the stereocilia. *J. Neurosci.* **29,** 15277–15285 (2009).
6. Pickles, J. O., Billieux-Hawkins, D. A. & Rouse, G. W. The incorporation and turnover of radiolabelled amino acids in developing stereocilia of the chick cochlea. *Hear. Res.* **101,** 45–54 (1996).

7. Lechene, C. *et al.* High-resolution quantitative imaging of mammalian and bacterial cells using stable isotope mass spectrometry. *J. Biol.* **5,** 20 (2006).
8. Lechene, C. P., Luyten, Y., McMahon, G. & Distel, D. L. Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science* **317,** 1563–1566 (2007).
9. Shepherd, G. M. G., Barres, B. A. & Corey, D. P. "Bundle Blot" purification and initial protein characterization of hair-cell stereocilia. *Proc. Natl Acad. Sci. USA* **86,** 4973–4977 (1989).
10. Shin, J. B. *et al.* Hair bundles are specialized for ATP delivery via creatine kinase. *Neuron* **53,** 371–386 (2007).
11. Schneider, M. E., Belyantseva, I. A., Azevedo, R. B. & Kachar, B. Rapid renewal of auditory hair bundles. *Nature* **418,** 837–838 (2002).
12. Grati, M. *et al.* Rapid turnover of stereocilia membrane proteins: evidence from the trafficking and mobility of plasma membrane Ca^{2+}-ATPase 2. *J. Neurosci.* **26,** 6386–6395 (2006).
13. Kirkegaard, M. & Nyengaard, J. R. Stereological study of postnatal development in the mouse utricular macula. *J. Comp. Neurol.* **492,** 132–144 (2005).
14. Perrin, B. J., Sonnemann, K. J. & Ervasti, J. M. β-actin and γ-actin are each dispensable for auditory hair cell development but required for stereocilia maintenance. *PLoS Genet.* **6,** e1001158 (2010).
15. Hanft, L. M. *et al.* Cytoplasmic γ-actin contributes to a compensatory remodeling response in dystrophin-deficient muscle. *Proc. Natl Acad. Sci. USA* **103,** 5385–5390 (2006).
16. Holt, J. R. *et al.* A chemical-genetic strategy implicates myosin-1c in adaptation by hair cells. *Cell* **108,** 371–381 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements C.P.L. thanks M. Raff for numerous discussions; T. Bloom for her insight, foresight and support at the origin of MIMS development. We thank D. Cotanche and G. Benichou for providing additional mouse cochlear samples, L. Trakimas for histological assistance, Z. Kaufman for assisting data analysis, and J. Hill for electronic and mechanical maintenance of the prototype instrument. G. McMahon contributed to operating the instrument and data analysis. This work was supported by National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB) grants P41RR14579, P41EB001974, NIH grants R01DC00033, R01DC03463, R01DC04179, R37DK39773, R01EY12963, R01GM47214 and R01D K58762, and National Science Foundation Division of Integrative Biology and Neuroscience (NSF/IBN) grant IBN-998298 to C.P.L., by NIH grant R01DC02281 to D.P.C., and by NIH grants F32DC009539 to B.J.P. and R01AR049899 to J.M.E., and by Wellcome Trust grant WT079643 to the Wellcome Trust Sanger Institute. Development of the SIMS instrument was supported by ONERA, CNRS, Université Paris Sud and Cameca (France). The work was also helped in part by software funded by the NIH National Center for Research Resources (NIH/NCRR) Center for Integrative Biomedical Computing, 2P41 RR0112553-12 and the Department of Energy SciDAC Visualization and Analytics Center for Enabling Technologies, DEFC0206ER25781. D.-S.Z. is a Research Associate, V.P. was a Research Associate and D.P.C. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions D.-S.Z. carried out the ^{15}N experiments and the preparation of tissue for MIMS imaging. V.P. and A.K.R. conceived the photobleaching experiments; A.K.R. and H.M.P. made the β-actin–GFP mouse and V.P. did the experiments. B.J.P. and J.M.E. conceived the conditional actin deletion experiments; B.J.P. made the mice and carried out the experiments. C.P.L. and D.P.C. conceived the MIMS study of hair cells. C.P.L. developed the MIMS method, oversaw the MIMS imaging with M.W. and analysed the MIMS images with J.C.P. D.P.C. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.P.L. (cpl@harvard.edu) or D.P.C. (dcorey@hms.harvard.edu).

METHODS

Tissue preparation and ^{15}N feeding. Adult bullfrogs (*Rana catesbeiana*) were fed an unlabelled control diet for 2–3 days, and were transferred at day 0 to a ^{15}N-rich diet. Food pellets contained ^{15}N-labelled leucine so that the ^{15}N content was 1.28%. Bullfrogs were fixed via cardiac perfusion with 4% formaldehyde + 0.1% glutaraldehyde. After decapitation, the saccules were exposed and the head and carcass stored in 4% formaldehyde + 1% glutaraldehyde. Heads were transferred to PBS before dissection of the saccules. Saccules were then treated with EDTA for 2–10 days to dissolve otoconia before embedding in Epon and sectioning.

Adult mice (*Mus musculus*, CD1 strain) ate control food that contained no added ^{15}N, and then were switched to food with ^{15}N-leucine (either 1.28 or 1.23%) for up to 150 days. Neonatal mice were labelled in two ways. For one set, mothers were fed ^{15}N food for 4–8 days, mated, and continued on ^{15}N food until pups were delivered (^{15}N mother). At birth, the pups were moved to a surrogate mother that had been maintained on control food (^{14}N mother). A second set was born to ^{14}N mothers and moved at birth to a surrogate mother that had been maintained for 24–28 days and was continued on ^{15}N food. Pups were nursed by surrogate mothers for 0, 2, 4, 8 or 15 days before being killed and fixed. At least two animals (four inner ears) were prepared for each condition.

Cultured utricles were dissected by standard methods[17]. After removing the otolithic membranes, the maculae were plated on a sterile portion of ACLAR film (EMS) and incubated in DMEM-F12 containing 100 μg ml^{-1} carbenicillin, with either 1 or 6% FBS. After an overnight incubation, the medium was replaced with ^{15}N-enriched medium (^{15}N, Cambridge Isotopes, 2.44 or 1.35% ^{15}N). The cultures were then incubated at 37 °C for up to 5 days until fixation. Tissues were fixed in 4% formaldehyde + 0.1% glutaraldehyde in cacodylate buffer for 15 min at 22 °C, then overnight at 4 °C. Cultures were then moved to cacodylate buffer and held for 2–3 days at 4 °C, before processing in batches for plastic embedding, as above.

MIMS analysis. In the multi-isotope imaging mass spectrometer[7], a ~30 nm beam of Cs^+ ions, scanned across the sample, sputters atoms, atomic clusters or molecules from the surface, of which a small fraction are ionized. Negative secondary ions are accelerated back, then shaped with ion optics and analysed with a double sector mass spectrometer employing four detectors, each tuned to a specific mass. Because N_2 is not charged, ^{15}N and ^{14}N were detected as cyanide anions, $^{12}C^{15}N^-$ and $^{12}C^{14}N^-$, mass 27 and mass 26, respectively, and their ratio was calculated for each point in the raster scan.

We calculated the proportion of newly synthesized protein, for each pixel of the field, from the ratio of counts in the $^{12}C^{15}N$ and $^{12}C^{14}N$ images according to

$$\text{incorporation} = \frac{\frac{N_{27}}{N_{26}} - A_n}{A_f - A_n}$$

where N_{27} and N_{26} are the counts in each pixel for masses 27 and 26, A_n is the natural abundance of ^{15}N (0.37%), and A_f is the relative abundance of ^{15}N in the food (1–2%). Incorporation was displayed with a hue saturation intensity transformation (HSI) in which the hue corresponds to the ratio value, and the intensity is an index of statistical reliability[8]. Image analysis performed in ImageJ/OpenMIMS (NIH/NRIMS). 3D reconstructions made in ImageVis3D (NIH/NCRR SCI/CIBC).

Photobleaching of cultured mouse utricles. To label endogenous actin, we constructed mice expressing a β-actin–GFP fusion under the myosin-7a promoter, integrated in the HPRT locus on the X chromosome. Correct integration of the BAC was confirmed Southern blot using flanking probes. Details of the construction, using methods previously described[17], are in Supplementary Fig. 6.

Neonatal pups (heterozygous and therefore mosaic for the transgene) were killed between P0 and P4, their utricles dissected and cultured as above. Dissection was performed on experimental day 0, then stereocilia were bleached on day 1 and images were taken on experimental days 3, 5 and 7. Each imaging day, the cells were transferred to DMEM-F12 without phenol red and then to the stage of an Olympus FV1000 confocal (upright configuration, LUMFL ×60 immersion objective, 1.1 NA) with a warmed chamber. On day 1, a Z-stack of the selected field containing 20–50 hair cells was collected using confocal imaging of the GFP with a 488 nm laser. Target hair bundles were selected and bleached with one to three 300-ms flashes of a focused 405 nm laser at 20 μs pixel^{-1}. Each selected bundle in the field was separately bleached. After bleaching, another Z-stack for the whole field was collected to confirm the bleached locations. The following experimental days one Z-stack per field was acquired. Movement of the bleached line was assessed by rotating the image stack to observe the line relative to the full length of the bundle.

Conditional deletion of β- and γ-actin. The *Actg1*^{flox} and *Actb*^{flox} mice, in which exons 3 and 4 were flanked by *loxP* sites[14] were crossed to transgenic mice expressing Cre/Esr1 recombinase driven by the chicken β-actin promoter coupled to the cytomegalovirus immediate-early enhancer (Jackson Labs; stock number 004682). All mice were on the C57BL/6 background. Mice were dosed with tamoxifen once a day for 4 days to induce recombination (20 mg ml^{-1} in corn oil, administered as 0.15 mg tamoxifen per gram of mouse by oral gavage). Esr1 control of Cre recombinase seemed to be leaky, in that 18-week-old mice that had not been given tamoxifen had the same actin-distribution phenotype as mice given tamoxifen; both were different from mice lacking CAG-Cre/Esr1. The phenotype was seen in a few cells of 3-week-old mice even in the absence of tamoxifen, so we restricted quantitative analysis to mice dosed with tamoxifen at 3 weeks.

Monoclonal mouse anti-γ-actin antibody clone 1-37 IgG (ref. 15) was purified from ascites using a T-gel Purification kit (Pierce) and conjugated to either Alexa-488 or Alexa-568 fluorescent dyes using a Monoclonal Antibody Labelling Kit (Invitrogen) following the manufacturer's instructions. FITC-labelled anti-β-actin antibodies (clone AC-15) were obtained from Abcam (validated in ref. 14).

For labelling, mice were perfused with 4% paraformaldehyde in PBS. Cochleas were then dissected and incubated in the same fixative for 2 h at room temperature (22 °C), washed in PBS and decalcified in 170 mM EDTA in PBS at 4 °C for 16 h. The organ of Corti was dissected, postfixed in 100% methanol at −20 °C for 10 min, rinsed in PBS and permeabilized in 0.5% Triton X-100 in PBS for 20 min at room temperature. Tissue was blocked for 1 h in 5% goat serum in PBS before incubation with the indicated antibodies. Samples were mounted in ProLong anti-fade reagent and viewed on a Deltavision PersonalDV deconvolution microscope equipped with a ×100 1.4 NA objective (Applied Precision). Stacks of images were collected at 0.20-μm intervals and deconvolved using Resolve3d software (Applied Precision).

17. Prosser, H. M., Rzadzinska, A. K., Steel, K. P. & Bradley, A. Mosaic complementation demonstrates a regulatory role for myosin VIIa in actin dynamics of stereocilia. *Mol. Cell. Biol.* **28,** 1702–1712 (2008).

LETTER

doi:10.1038/nature10754

Neuron-type-specific signals for reward and punishment in the ventral tegmental area

Jeremiah Y. Cohen[1]*, Sebastian Haesler[1]*, Linh Vong[2], Bradford B. Lowell[2] & Naoshige Uchida[1]

Dopamine has a central role in motivation and reward. Dopaminergic neurons in the ventral tegmental area (VTA) signal the discrepancy between expected and actual rewards (that is, reward prediction error)[1–3], but how they compute such signals is unknown. We recorded the activity of VTA neurons while mice associated different odour cues with appetitive and aversive outcomes. We found three types of neuron based on responses to odours and outcomes: approximately half of the neurons (type I, 52%) showed phasic excitation after reward-predicting odours and rewards in a manner consistent with reward prediction error coding; the other half of neurons showed persistent activity during the delay between odour and outcome that was modulated positively (type II, 31%) or negatively (type III, 18%) by the value of outcomes. Whereas the activity of type I neurons was sensitive to actual outcomes (that is, when the reward was delivered as expected compared to when it was unexpectedly omitted), the activity of type II and type III neurons was determined predominantly by reward-predicting odours. We 'tagged' dopaminergic and GABAergic neurons with the light-sensitive protein channelrhodopsin-2 and identified them based on their responses to optical stimulation while recording. All identified dopaminergic neurons were of type I and all GABAergic neurons were of type II. These results show that VTA GABAergic neurons signal expected reward, a key variable for dopaminergic neurons to calculate reward prediction error.

Dopaminergic neurons fire phasically (100–500 ms) after unpredicted rewards or cues that predict reward[1–3]. Their response to reward is reduced when a reward is fully predicted. Furthermore, their activity is suppressed when a predicted reward is omitted. From these observations, previous studies hypothesized that dopaminergic neurons signal discrepancies between expected and actual rewards (that is, they compute reward prediction error (RPE)), but how dopaminergic neurons compute RPE is unknown.

Dopaminergic neurons make up about 55–65% of VTA neurons; the rest are mostly GABAergic inhibitory neurons[4–6]. Many addictive drugs inhibit VTA GABAergic neurons, which increases dopamine release (called disinhibition), a potential mechanism for reinforcing the effects of these drugs[7–12]. Despite the known role of VTA GABAergic neurons inhibiting dopaminergic neurons *in vitro*[13], little is known about their role in normal reward processing. One obstacle has been the difficulty of identifying different neuron types with extracellular recording techniques. Conventionally, spike waveforms and other firing properties have been used to identify presumed dopaminergic and GABAergic neurons[1,2,14,15], but this approach has been questioned recently[5,16]. We thus aimed to observe how dopaminergic and GABAergic neurons process information about rewards and punishments.

We classically conditioned mice with different odour cues that predicted appetitive or aversive outcomes. The possible outcomes were big reward, small reward, nothing, or punishment (a puff of air delivered to the animal's face). Each behavioural trial began with a conditioned stimulus (CS; an odour, 1 s), followed by a 1-s delay and an unconditioned stimulus (US; the outcome). Within the first two behavioural sessions, mice began licking towards the water-delivery tube in the delay before rewards arrived, indicating that they quickly learned the CS–US associations (Fig. 1). The lick rate was significantly higher preceding big rewards than small ones (paired *t*-tests between lick rates for big versus small rewards for each session, $P < 0.05$ for each mouse).

We recorded the activity of VTA neurons while mice performed the conditioning task. All 95 neurons showed task-related responses (analysis of variance (ANOVA), all $P < 0.001$), thus all recorded neurons were used in the following analyses. Observing the temporal profiles of responses in trials with rewards, we found neurons that showed firing patterns that resemble those of dopaminergic neurons found in non-human primates[1,2,15]. These neurons were excited phasically by reward-predicting stimuli or reward (Fig. 2a, top). We also found many neurons with firing patterns distinct from typical dopaminergic neurons. These neurons showed persistent excitation during the delay before rewards, in response to reward-predicting odours (Fig. 2a, middle). Other neurons showed persistent inhibition to reward-predicting odours (Fig. 2a, bottom). To characterize the responses of the population, we measured the temporal response profile of each neuron during big-reward trials by quantifying firing rate changes from baseline in 100-ms bins using a receiver operating characteristic (ROC) analysis (Fig. 2b and Supplementary Fig. 1). We calculated the area under the ROC curve (auROC) at each time bin. Values greater than 0.5 indicate increases in firing rate relative to baseline, whereas values less than 0.5 indicate decreases.

To classify these response profiles, we used principal component analysis (PCA) followed by unsupervised, hierarchical clustering. This yielded three clusters of neurons that were separated according to (1) the magnitude of activity during the delay between CS and US, and (2) the magnitude of responses to the CS or US (Fig. 2c). Forty-nine neurons (52%) were classified as type I, which showed phasic responses. Twenty-nine neurons (31%) were classified as type II, which showed sustained excitation to reward-predicting odours, whereas 17 neurons (18%) were classified as type III, which showed sustained inhibition (Fig. 2d).

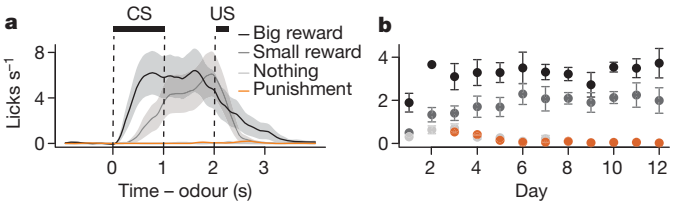


Figure 1 | Odour-outcome association task in mice. a, Licking behaviour from a representative experimental session. Black bars indicate CS and US delivery. Shaded regions around lick traces denote standard error of the mean (s.e.m.). **b**, Mean ± s.e.m. licks during the delay between CS and US as a function of days of the experiment across animals.

[1]Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Division of Endocrinology, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA.
*These authors contributed equally to this work.

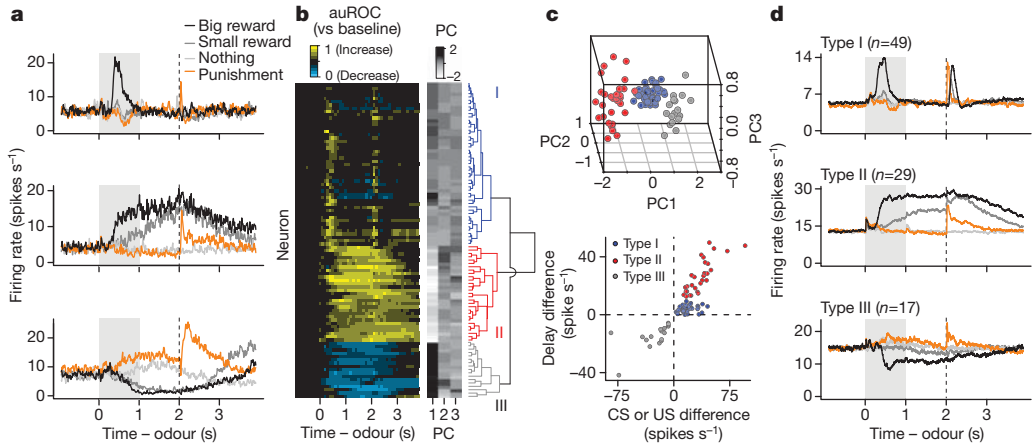
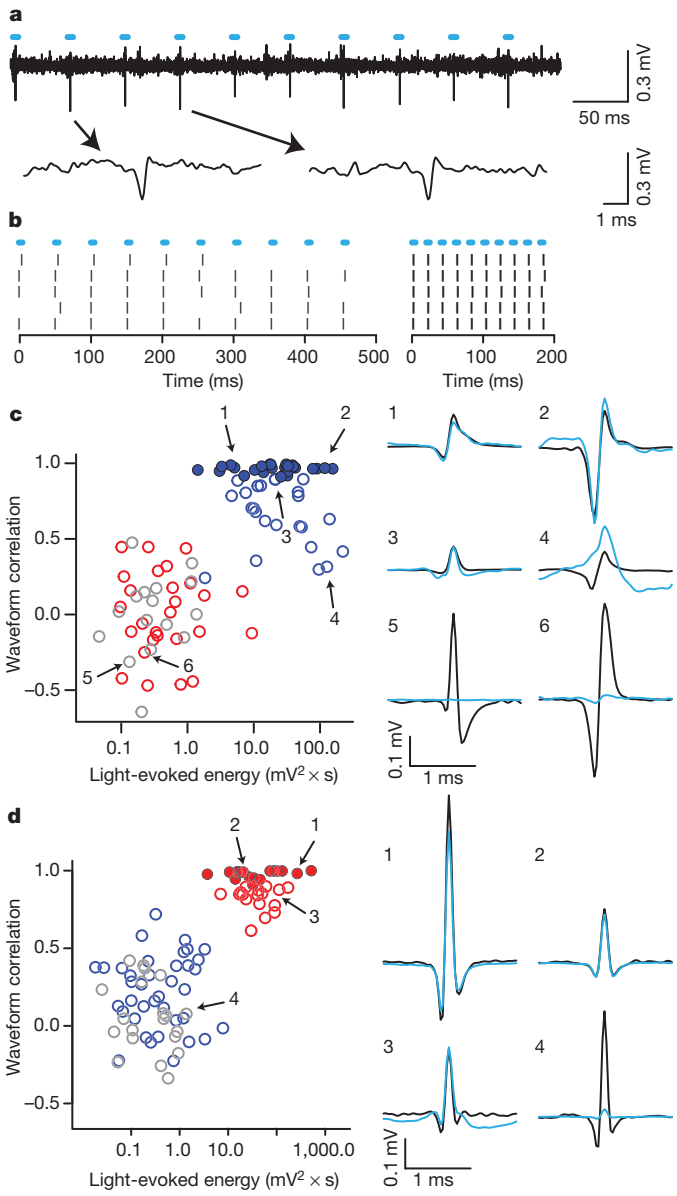


Figure 2 | VTA neurons show three distinct response types. **a**, Responses of example neurons. **b**, Responses of all neurons. Yellow, increase from baseline; cyan, decrease from baseline. Each row represents one neuron. The similarity order of the three main clusters is arranged to match the order presented in **a**. **c**, Top: the first three principal components of the auROC curves. Points are coloured based on hierarchical clustering from the dendrogram. Bottom: classification of neurons based on response differences between big-reward and no-reward trials during the delay versus during the CS or US. **d**, Average firing rates from type I–III neurons.



To identify dopaminergic neurons, we expressed channelrhodopsin-2 (ChR2), a light-gated cation channel[17,18], in dopaminergic neurons (see Methods). We confined expression to dopaminergic neurons by injecting adeno-associated virus containing FLEX-ChR2 (AAV-FLEX-ChR2)[19] into transgenic mice expressing Cre recombinase under the control of the promoter of the dopamine transporter (*DAT*; also called *Slc6a3*) gene (Supplementary Figs 2 and 3). For each neuron, we measured the response to light pulses and the shape of spontaneous spikes. We observed many neurons that fired after light pulses (Fig. 3a, b). We calculated the correlation between the spontaneous spike waveform and light-evoked voltage response and plotted it against the energy of light-evoked responses for each recording (Fig. 3c). This yielded two distinct clusters: one that showed significant responses to light pulses and one that did not. To identify dopaminergic neurons stringently, we applied the criterion that the light-evoked waveform must look almost identical to the spontaneous waveform (correlation coefficient >0.9). Twenty-six neurons met this criterion (filled blue points in Fig. 3c). Consistent with direct light activation rather than indirect, synaptic activation, all 26 neurons showed light-evoked spikes within a few milliseconds of light onset with small jitter, and followed high-frequency light stimulation of 50 Hz (Supplementary Fig. 4). These properties strongly indicate that these 26 neurons expressed ChR2. We therefore designate these 26 neurons as identified dopaminergic neurons. All identified dopaminergic neurons were of type I. Conversely, none of type II or III neurons was activated by light (red and grey points in Fig. 3c).

Next, we asked whether GABAergic neurons could be mapped to type II or III neurons. We recorded from 92 VTA neurons in mice expressing Cre recombinase under the control of the endogenous vesicular γ-aminobutyric acid (GABA) transporter (*Vgat*; also called *Slc32a1*) gene. These mice showed similar licking behaviour to *DAT*-Cre mice (Supplementary Fig. 5). We applied the PCA parameters

Figure 3 | Identifying dopaminergic and GABAergic neurons. **a**, Voltage trace from 10 pulses of 20 Hz light stimulation (cyan bars). Two light-triggered spikes are shown below. **b**, Response from this neuron to 20 Hz (left) and 50 Hz (right) stimulation. Ticks represent spikes. **c**, Quantification of light-evoked responses and mapping of response types in *DAT*-Cre mice. Blue, type I; red, type II; grey, type III neurons. Identified dopaminergic neurons are indicated by filled circles. Abscissa: energy (integral of the squared voltage values, $\int v^2 dt$) of the light-evoked response from each neuron. Ordinate: cross-correlation between the mean spontaneous spike and the light-evoked response. Example neurons are shown to the right (black, spontaneous spikes; cyan, light-evoked voltages). **d**, Light-evoked responses in *Vgat*-Cre mice. Conventions are the same as in **c**.

from the 95 neurons from *DAT*-Cre mice to the 92 neurons from *Vgat*-Cre mice. This yielded 38 type I neurons, 34 type II neurons and 20 type III neurons. Using the same criteria for GABAergic neurons as we used for dopaminergic neurons, we identified 17 GABAergic neurons (Fig. 3d and Supplementary Fig. 4). All 34 type II neurons fell in the upper cluster in Fig. 3d. We also found type I neurons that were inhibited by optical stimulation, consistent with local GABAergic stimulation (Supplementary Fig. 6).

Our data set of identified dopaminergic neurons allows us to characterize their diversity. We observed that some were excited by reward, some were excited by a reward-predicting CS, and some were excited by both (Fig. 4a–c). Although previous studies in non-human primates found similar variability[20,21] (Supplementary Fig. 7), this result may suggest that some dopaminergic neurons do not strictly follow canonical RPE coding. However, the US responses may be due to the delay between CS and US, known to increase the US response due to temporal uncertainty[20]. In addition, this diversity was correlated with the effect of training that occurred over several days across the population of dopaminergic neurons, even after animals had reached asymptotic behavioural performance (Fig. 1b). Soon after reaching a behavioural performance criterion, many dopaminergic neurons showed stronger responses to US over CS but the preference gradually shifted to CS over several days (Fig. 4d; Pearson correlation, $r = 0.42$, $P < 0.05$). This is consistent with a previous study in non-human primates that showed US responses gradually disappear over >1 month of training[21]. Thus, identified dopaminergic neurons appear to respond to CS and US similarly to those reported in non-human primate studies.

Another important response property that supports RPE coding in dopaminergic neurons is their decrease in firing rate when an expected reward is omitted[1,3]. We thus omitted reward unexpectedly on 10% of big-reward trials in some sessions. Fifteen of seventeen dopaminergic neurons showed a decrease in firing rate upon reward omission relative to reward delivery (Fig. 4f, g). The two dopaminergic neurons that were not modulated by reward omission were excited by big-reward CS, but fired close to 0 spikes s^{-1} otherwise; the low firing rate at the time of reward left little room to 'dip' further. We obtained similar results when we compared the firing rate upon reward omission to the baseline firing rate (9 of 17 neurons $P < 0.05$, *t*-test; mean auROC = 0.407, $t_{16} = 2.56$, $P < 0.05$; Supplementary Fig. 8a, b). Thus, most dopaminergic neurons coded RPE when expected reward was omitted.

GABAergic neurons showed persistent activity during the delay period, which parametrically encoded the value of upcoming outcomes (paired *t*-tests between no-, small- and big-reward trials, all $P < 0.001$ for 16 of 17 identified GABAergic neurons, Supplementary Fig. 7a; regression slopes, Supplementary Fig. 10i). This suggests that these neurons encode expectation about rewards. If this is the case, one prediction is that the activity of these neurons is not modulated by delivery or omission of reward. Indeed, GABAergic (and unidentified type II) and type III neurons were not significantly modulated by the presence or absence of reward itself (Fig. 4f, g and Supplementary Fig. 8), in contrast to identified dopaminergic neurons. None of the identified GABAergic neurons, and only 2 of 17 unidentified type II neurons, showed significant decreases in firing rate relative to when reward was delivered. None of the 11 type III neurons showed significant modulation by reward omission. Thus, the activity of type II and III neurons was modulated predominantly by reward-predicting cues but not actual reward.

Recent studies have revealed a diversity of responses of dopaminergic neurons to aversive stimuli: some are excited, others inhibited[15]. To test whether this diversity exists in dopaminergic and GABAergic VTA neurons, we delivered air puffs in some sessions. Identified dopaminergic neurons showed some diversity: although most significant responses were inhibitory, some were excitatory (Fig. 4h, i and Supplementary Fig. 9). In contrast, most type II and III neurons (and 13 of 14 identified GABAergic neurons) were excited by air puffs.

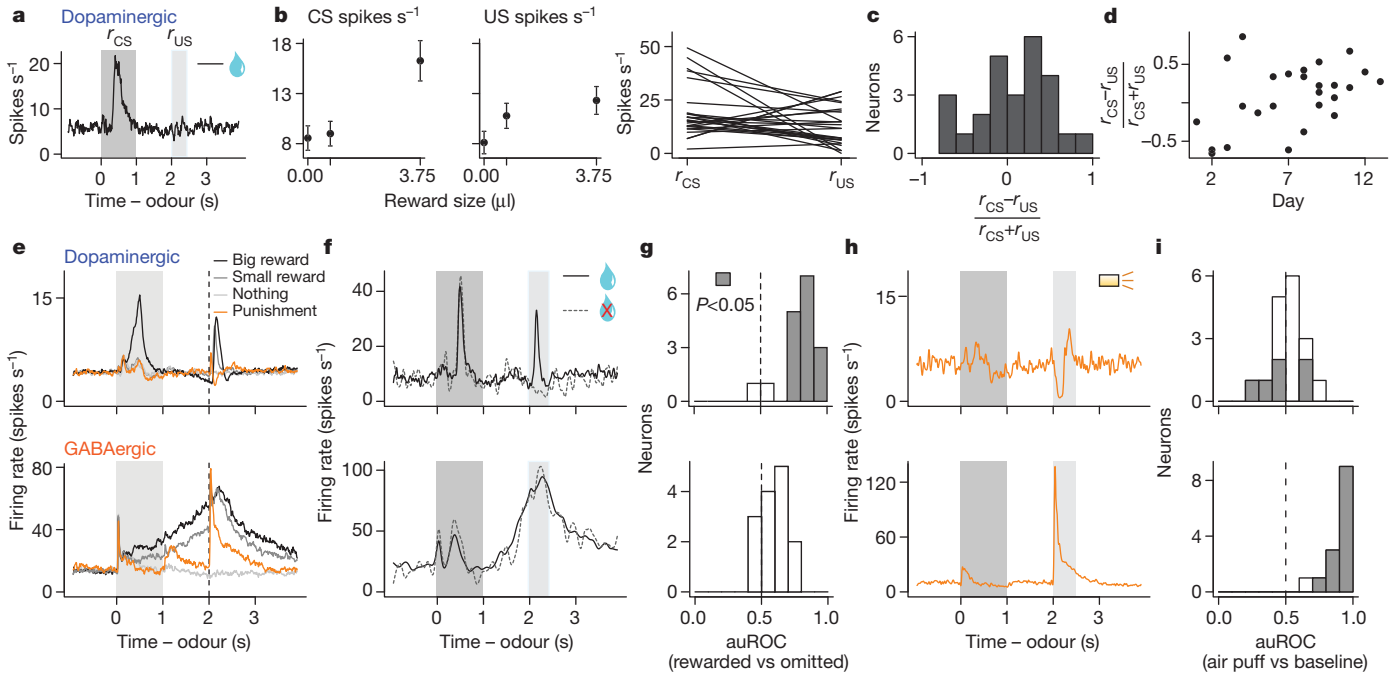


Figure 4 | Response variability based on CS–US preference, reward omission and air puffs. **a**, Response of a dopaminergic neuron during big-reward trials. **b**, Firing rate (mean ± s.e.m.) versus reward size (left) and in response to big-reward-predicting CS and big-reward US for each dopaminergic neuron (right). **c**, Histogram of CS–US index for dopaminergic neurons. **d**, CS–US index versus day after the behaviour was learned. **e**, Average responses of dopaminergic and GABAergic neurons. **f**, Responses of a dopaminergic and GABAergic neuron for reward present (solid) and unexpectedly absent (dashed) on big-reward trials. **g**, Histograms of differences in firing rates during the outcome period (2–2.5 s) between rewarded and reward-omitted trials for dopaminergic (top) and GABAergic (bottom) neurons. Values are represented using auROC (<0.5, rewarded < omitted; 0.5, no difference; >0.5, rewarded > omitted). Significant values are filled (*t*-test, $P < 0.05$). **h**, Responses of a dopaminergic and GABAergic neuron during punishment trials. **i**, Histograms of auROC values during the air puff (2–2.5 s) relative to baseline (<0.5, decrease; >0.5, increase from baseline).

Detecting the discrepancy between expected and actual outcomes has a critical role in optimal learning[1,22,23]. Although phasic firing of VTA dopaminergic neurons may act as such an error signal, how this is computed remains largely unknown. Models have postulated the existence of value-dependent, inhibitory input to dopaminergic neurons that persists during the delay between a CS and US (Supplementary Fig. 11a)[1,23]. Our data indicate that VTA GABAergic neurons provide such an inhibitory input that counteracts excitatory drive from primary reward when the reward is expected. In addition, these neurons were excited by aversive stimuli, potentially contributing to suppression of firing in some dopaminergic neurons in response to aversive events (Fig. 4). Previous work showed that VTA GABAergic neurons receive inputs from prefrontal cortex and subcortical areas that could provide reward-related signals[24–29]. Phasic excitation of VTA GABAergic neurons could be driven by inputs from lateral habenula neurons that are phasically excited by aversive stimuli[29]. These habenular neurons do not show sustained activity between CS and US, so it is unlikely that they provide reward expectation signals to VTA GABAergic neurons. Instead, these signals may come from the pedunculopontine nucleus[25] or orbitofrontal cortex[27] (Supplementary Fig. 11b). VTA GABAergic neurons form synapses preferentially onto dendrites of dopaminergic neurons[28], whereas other inhibitory inputs form synapses onto their somata[29]. Dendritic inhibition is thought to be weaker than somatic 'shunting' inhibition[28] but appears well suited for deriving graded outputs by 'arithmetically' combining excitatory and inhibitory inputs.

A major effect of drugs of addiction is inhibition of VTA GABAergic neurons[7,8]. If VTA GABAergic neurons are involved in computation of RPE, inhibition of GABAergic neurons by addictive drugs could lead to sustained RPE even after the learned effects of drug intake are well established, thereby resulting in sustained reinforcement of drug taking[30]. Understanding local circuits in VTA in the context of learning theory may thus provide crucial insights into normal as well as abnormal functions of reward circuits.

METHODS SUMMARY

All surgical and experimental procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Institutional Animal Care and Use Committee. We injected male *DAT*-Cre and *Vgat*-Cre mice with adeno-associated virus carrying FLEX-ChR2 into the VTA and implanted a head plate and a microdrive containing six tetrodes and an optical fibre. While mice performed a classical conditioning task, we recorded spiking activity from VTA neurons. We delivered pulses of light to activate ChR2 and classified neurons as dopaminergic, GABAergic or unidentified. After experiments, we performed immunohistochemistry to localize recording sites amid dopaminergic neurons.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 May; accepted 2 December 2011.
Published online 18 January 2012.

1. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).
2. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47,** 129–141 (2005).
3. Schultz, W. Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.* **57,** 87–115 (2006).
4. Swanson, L. W. The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.* **9,** 321–353 (1982).
5. Margolis, E. B., Lock, H., Hjelmstad, G. O. & Fields, H. L. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *J. Physiol.* **577,** 907–924 (2006).
6. Nair-Roberts, R. G. *et al.* Stereological estimates of dopaminergic, GABAergic and glutamatergic neurons in the ventral tegmental area, substantia nigra and retrorubral field in the rat. *Neuroscience* **152,** 1024–1031 (2008).
7. Hyman, S. E., Malenka, R. C. & Nestler, E. J. Neural mechanisms of addiction: the role of reward-related learning and memory. *Annu. Rev. Neurosci.* **29,** 565–598 (2006).
8. Lüscher, C. & Malenka, R. C. Drug-evoked synaptic plasticity in addiction: from molecular changes to circuit remodeling. *Neuron* **69,** 650–663 (2011).
9. Johnson, S. W. & North, R. A. Opioids excite dopamine neurons by hyperpolarization of local interneurons. *J. Neurosci.* **12,** 483–488 (1992).
10. Mansvelder, H. D., Keath, J. R. & McGehee, D. S. Synaptic mechanisms underlie nicotine-induced excitability of brain reward areas. *Neuron* **33,** 905–919 (2002).
11. Szabo, B., Siemes, S. & Wallmichrath, I. Inhibition of GABAergic neurotransmission in the ventral tegmental area by cannabinoids. *Eur. J. Neurosci.* **15,** 2057–2061 (2002).
12. Tan, K. R. *et al.* Neural bases for addictive properties of benzodiazepines. *Nature* **463,** 769–774 (2010).
13. Dobi, A., Margolis, E. B., Wang, H.-L., Harvey, B. K. & Morales, M. Glutamatergic and nonglutamatergic neurons of the ventral tegmental area establish local synaptic contacts with dopaminergic and nondopaminergic neurons. *J. Neurosci.* **30,** 218–229 (2010).
14. Steffensen, S. C., Svingos, A. L., Pickel, V. M. & Henriksen, S. J. Electrophysiological characterization of GABAergic neurons in the ventral tegmental area. *J. Neurosci.* **18,** 8003–8015 (1998).
15. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459,** 837–841 (2009).
16. Lammel, S. *et al.* Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron* **57,** 760–773 (2008).
17. Nagel, G. *et al.* Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc. Natl Acad. Sci. USA* **100,** 13940–13945 (2003).
18. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8,** 1263–1268 (2005).
19. Atasoy, D., Aponte, Y., Su, H. H. & Sternson, S. M. A. FLEX switch targets Channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28,** 7025–7030 (2008).
20. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nature Neurosci.* **11,** 966–973 (2008).
21. Takikawa, Y., Kawagoe, R. & Hikosaka, O. A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *J. Neurophysiol.* **92,** 2520–2529 (2004).
22. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black, A. H. & Wagner, A. R.) 64–99 (New York, 1972).
23. Houk, J. C., Adams, J. L. & Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 249–270 (MIT Press, 1995).
24. Carr, D. B. & Sesack, S. R. Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. *J. Neurosci.* **20,** 3864–3873 (2000).
25. Okada, K., Toyama, K., Inoue, Y., Isa, T. & Kobayashi, Y. Different pedunculopontine tegmental neurons signal predicted and actual task rewards. *J. Neurosci.* **29,** 4858–4870 (2009).
26. Matsumoto, M. & Hikosaka, O. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447,** 1111–1115 (2007).
27. Takahashi, Y. K. *et al.* Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neurosci.* **14,** 1590–1597 (2011).
28. Omelchenko, N. & Sesack, S. R. Ultrastructural analysis of local collaterals of rat ventral tegmental area neurons: GABA phenotype and synapses onto dopamine and GABA cells. *Synapse* **63,** 895–906 (2009).
29. Jhou, T. C., Fields, H. L., Baxter, M. G., Saper, C. B. & Holland, P. C. The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. *Neuron* **61,** 786–800 (2009).
30. Redish, A. D. Addiction as a computational process gone awry. *Science* **306,** 1944–1947 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Meister, V. N. Murthy, J. D. Schall and R. P. Heitz for comments, C. Dulac for sharing resources, C. I. Moore, J. Ritt and J. Siegle for advice about microdrives, K. Deisseroth for the AAV-FLEX-ChR2 construct, and E. Soucy and J. Greenwood for technical support. This work was supported by a Howard Hughes Medical Institute Fellowship from the Helen Hay Whitney Foundation (J.Y.C.); the Human Frontiers Science Program (S.H.); a Howard Hughes Medical Institute Collaborative Innovation Award, a Smith Family New Investigator Award, the Alfred Sloan Foundation, the Milton Fund (N.U.); F32 DK078478, P30 DK046200 (L.V.); and R01 DK075632, R01 DK089044, P30 DK046200, P30 DK057521 (B.B.L.).

Author Contributions J.Y.C. and S.H. collected and analysed data. J.Y.C., S.H. and N.U. designed experiments and wrote the paper. L.V. and B.B.L. generated *Vgat*-Cre mice.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.U. (uchida@mcb.harvard.edu).

METHODS

Animals. We used six adult male mice, backcrossed with C57/BL6J mice, heterozygous for Cre recombinase under the control of the *DAT* gene (B6.SJL-Slc6a3$^{tm1.1(cre)Bkmn}$/J, The Jackson Laboratory)[31] and six adult male mice, backcrossed with C57/BL6J mice, heterozygous for Cre recombinase under the control of the *Vgat* gene (*Vgat*-ires-Cre)[32]. Animals were housed on a 12-h dark/12-h light cycle (dark from 06:00 to 18:00) and each performed the conditioning task at the same time of day, between 07:00 and 18:00.

Surgery and viral injections. Mice were surgically implanted with a head plate and a microdrive containing electrodes and an optical fibre. During surgery, we injected 200–500 nl adeno-associated virus (AAV), serotype 5, carrying an inverted ChR2 (H134R)-EYFP flanked by double *loxP* sites[18,19,33] into the VTA stereotactically (from bregma: 3.1 mm posterior, 0.8 mm lateral, 4–4.5 mm ventral). Our expression was highly selective (<1% of ChR2-positive neurons were negative for the dopaminergic neuron marker tyrosine hydroxylase, TH) and efficient (>90% of TH-expressing neurons co-expressed ChR2; Supplementary Figs 2 and 3).

For each unidentified type I neuron in *DAT*-Cre mice, we were unable to identify it as dopaminergic using our stringent criteria not because there was no response on the electrode after light pulses, but because the response did not match the shape of the spontaneous spike waveforms (open blue points in Fig. 3c). This may have been from changing the spike shape by opening many ChR2 or due to 'spillover' from nearby neurons, contaminating the signal.

After injections in *Vgat*-Cre mice, we confirmed that our ChR2 expression was selective (<1% of ChR2-positive neurons were negative for the GABAergic neuron marker glutamic acid decarboxylase, GAD65/67) and efficient (but less efficient than in *DAT*-Cre mice; >60% of GAD65/67-expressing neurons co-expressed ChR2) (Supplementary Fig. 2).

For fluorogold tracer injections (Fluorochrome), we injected two animals each with 10–25 nl into the following areas: prefrontal cortex (from bregma: 2.1 mm anterior, 0.25 mm lateral, 1.6 mm ventral), medial striatum (1.4 mm anterior, 1.0 mm lateral, 4.0 mm ventral) and basolateral amygdala (1.4 mm posterior, 2.8 mm lateral, 4.3 mm ventral).

All surgery was performed under aseptic conditions with animals under ketamine/medetomidine anaesthesia (60/0.5 mg kg^{-1}, intraperitoneal, respectively). Analgesia (ketoprofen, 5 mg kg^{-1}, intraperitoneal) was administered postoperatively.

Behavioural task. After >1 week of recovery, mice were water-deprived in their home cage. Weight was maintained above 90% of their full body weight. Animals were head-restrained using a custom-made metal plate and habituated for 1–2 d while head-restrained before training on the task. Odours were delivered with a custom-made olfactometer[34]. Each odour was dissolved in paraffin oil at 1/100 dilution. Thirty microlitres of diluted odour was placed inside a filter-paper housing. Odours were isoamyl acetate, 1-butanol, *N*-butyl acetate, citral, eugenol, (+) limonene and (−) carvone, and differed for different animals. Odourized air was further diluted with filtered air by 1:10 to produce a 500 ml min^{-1} total flow rate. Licks were detected by breaks of an infrared beam placed in front of the water tube.

We delivered one of four odours, selected pseudorandomly, for 1 s, followed by a delay of 1 s and an outcome. Each odour predicted a different outcome: a large drop of water (3.75 µl; valve open for 100 ms), a small drop of water (0.75 µl; valve open for 20 ms), no outcome, or an air puff delivered to the animal's face. Air puff trials were included for 17 identified dopaminergic neurons, 14 identified GABAergic neurons, and 15 type III neurons. Intertrial intervals (ITIs) were drawn from an exponential distribution with a rate parameter of 10 (that is, ITI = 10exp(−10*x*)). This resulted in a flat ITI hazard function, ensuring that expectation about the start of the next trial did not increase over time. Data from *DAT*-Cre mice were obtained from 85 sessions (9–23 sessions per animal, 14 ± 4.8 sessions, mean ± s.d.). For 17 identified dopaminergic neurons, 14 identified GABAergic neurons, and 11 type III neurons, we omitted rewards during 10% of big-reward trials. Animals performed between 400 and 1,000 trials per day (533 ± 120 trials, mean ± s.d.). Data from *Vgat*-Cre mice were obtained from 71 sessions.

Electrophysiology. We recorded extracellularly from multiple neurons simultaneously using a custom-built 200-µm-fibreoptic-coupled screw-driven microdrive with six implanted tetrodes (four wires wound together). For three *Vgat*-Cre mice, we used a 105-µm fibre-optic. Tetrodes were glued to the fibre optic with epoxy. The ends of the tetrodes were 350–600 µm from the end of the fibre optic. Neural signals and time stamps for behaviour were recorded using a DigiLynx recording system (Neuralynx). Broadband signals from each wire filtered between 0.1 and 9,000 Hz were recorded continuously at 32 kHz. To extract the timing of spikes, signals were band-pass-filtered between 300–6,000 Hz and sorted online and off-line using custom software.

To verify that our recordings targeted dopaminergic or GABAergic neurons, we used ChR2 to observe stimulation-locked spikes[35]. The optical fibre was coupled with a diode-pumped solid-state laser with analogue amplitude modulation (Laserglow Technologies). Before and after each behavioural session, we delivered trains of 10 light pulses, each 5-ms long, at 1, 5, 10, 20 and 50 Hz at 473 nm at 5–20 mW mm^{-2}. We used frequencies of 1, 10, 20, 50 and 100 Hz in *Vgat*-Cre mice. Higher intensities typically resulted in light-evoked spike waveforms that did not match spontaneous ones (open blue points in Fig. 3c and open red points in Fig. 3d). Therefore, we adjusted the light intensity after observing the responses. The increasing latency of light-evoked spiking as a function of stimulation frequency indicates that the response was not due to photochemical artefact (Supplementary Fig. 4f, j). Spike shape was measured using a broadband signal (0.1– 9,000 Hz) sampled at 32 kHz. This ensured that particular features of the spike waveform were not missed. A previous study found decreasing probability of activation with increasing frequency[33]. This discrepancy may be explained by the low levels of light we used to identify neurons and the difference in experimental preparation, *in vitro* versus *in vivo*[36].

We used two criteria to include a neuron in our data set. First, the neuron must have been well isolated (signal-to-noise ratio of >5 dB). Second, the neuron must have been recorded within 500 µm of an identified dopaminergic neuron (or a type I neuron in *Vgat*-Cre mice), to ensure that all neurons came from VTA. Using distance cutoffs of 400, 300 or 200 µm yielded similar estimates of the proportion of type I neurons as the full data set (49 of 91, 49 of 86 and 49 of 82, respectively, in *DAT*-Cre mice).

Recording sites were further verified histologically with electrolytic lesions using 15–20 s of 100 µA direct current and from the optical fibre track (Supplementary Fig. 12).

Data analysis. To measure firing rates, peristimulus time histograms (PSTHs) were constructed using 10-ms bins. To calculate spike density functions, PSTHs were convolved with a function resembling a postsynaptic potential[37]. To determine whether a neuron showed a significant task-related response, we calculated an ANOVA on the trial-by-trial firing rates (spikes s^{-1}) during the baseline period (1 s before odour onset), CS period (from odour onset to odour offset), delay (from odour offset to outcome onset), and US period (from outcome onset to 500 ms after outcome onset). The factors were task epoch (that is, baseline, CS, delay, or US) and outcome type. All neurons showed a main effect of time; 183 of 187 neurons showed a main effect of outcome type; 184 of 187 neurons showed a time–outcome interaction ($P < 0.001$).

Light-evoked spikes were detected during the 10 ms after light onset. If <20% of light pulses had a spike (defined as a waveform that matched that of the isolated unit) during the 10 ms after light onset (lower left points in Fig. 3c, d), the maximum voltage during that interval was used as the light-evoked 'response'. Cross-correlations between spontaneous and light-evoked spike waveforms were calculated by aligning the positive peak of each waveform, averaging separately, and aligning the peaks of the averages. The cross-correlation coefficients reported in Fig. 3 are from a lag of 0 ms. The correlation was calculated using the full duration of the spontaneous spike (spike duration was measured as the first time until the last time at which the voltage was significantly different from baseline using Wilcoxon rank sum tests). The energy of the light-evoked waveform is defined as the integral of the squared voltage values ($\int v^2 dt$). To ensure that measurements of the fidelity of light-evoked responses were not biased by poorly isolated neurons, we plotted the probability of firing at 50 Hz light stimulation against L-ratio, a measure of isolation quality[38] (Supplementary Fig. 13). L-ratio approximates the fraction of 'contaminated' spikes; smaller L-ratios indicate better isolation. One hundred and eighty of 187 neurons in the data set had L-ratios <0.05.

CS–US indices were calculated as (CS − US)/(CS + US), where CS is the difference between the peak firing rate (maximum value of the PSTH) in the 1 s after odour onset and the baseline firing rate, and US is the difference between the peak firing rate in the 0.5 s after reward onset and the baseline firing rate. The baseline firing rate was calculated as the mean of the PSTH in the 0.5 s before odour onset.

ROC curves were calculated by comparing the distribution of firing rates across trials in 100-ms bins to the distribution of baseline firing rates (900 ms before odour onset; Supplementary Fig. 1). PCA was calculated using the singular value decomposition of the auROC. Learning was measured using the lick responses (mean licks s^{-1}) across consecutive days of behavioural sessions. Curves were fit with logistic functions, $k/(1+\exp(-t))$. The task was considered fully learned if the learning rate was within 95% of the carrying capacity of the logistic (k). All 12 animals learned the task within the first 3 days.

Hierarchical clustering was done using the first three PCs of the auROC curves using a Euclidean distance metric and complete agglomeration method.

All statistical tests were done with corrections for multiple comparisons (Bonferroni or Tukey). Analyses were done with R (http://www.r-project.org/).

Immunohistochemistry. After recording, which lasted between 9 and 23 days, mice were given an overdose of ketamine/medetomidine, exsanguinated with saline, perfused with paraformaldehyde, and brains were cut in 50 µm coronal or horizontal sections. Sections were immunostained with antibodies to TH and

secondary antibodies labelled with Cy3 (Jackson Immunoresearch). Sections were further stained with 4′,6-diamidino-2-phenylindole (DAPI) or TO-PRO-3 (Invitrogen) to visualize nuclei. Recording sites were identified and verified to be amid EYFP staining and TH staining in VTA. Brain slices from additional virus-injected *DAT*-Cre and *Vgat*-Cre animals were stained for TH and GAD-65/67 (Millipore) with secondary antibodies labelled with Cy3 or Alexa 594 (Invitrogen). Cellular fluorescence intensities of ChR2-EYFP were measured using Volocity Image Analysis Software (Perkin Elmer).

31. Bäckman, C. M. *et al.* Characterization of a mouse strain expressing Cre recombinase from the 3′ untranslated region of the dopamine transporter locus. *Genesis* **44,** 383–390 (2006).
32. Vong, L. *et al.* Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71,** 142–154 (2011).
33. Tsai, H. C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324,** 1080–1084 (2009).
34. Uchida, N. & Mainen, Z. F. Speed and accuracy of olfactory discrimination in the rat. *Nature Neurosci.* **6,** 1224–1229 (2003).
35. Lima, S. Q., Hromádka, T., Znamenskiy, P. & Zador, A. M. PINP: a new method of tagging neuronal populations for identification during in vivo electrophysiological recording. *PLoS ONE* **4,** e6099 (2009).
36. Zhao, S. *et al.* Cell-type specific channelrhodopsin-2 transgenic mice for optogenetic dissection of neural circuitry function. *Nature Methods* **8,** 745–752 (2011).
37. Thompson, K. G., Hanes, D. P., Bichot, N. P. & Schall, J. D. Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *J. Neurophysiol.* **76,** 4040–4055 (1996).
38. Schmitzer-Torbert, N. & Redish, A. D. Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the Multiple T Task. *J. Neurophysiol.* **91,** 2259–2272 (2004).

LETTER

doi:10.1038/nature10758

Exercise-induced BCL2-regulated autophagy is required for muscle glucose homeostasis

Congcong He[1,2,3]*, Michael C. Bassik[4]†*, Viviana Moresi[5], Kai Sun[2,6], Yongjie Wei[1,2,3], Zhongju Zou[1,2,3], Zhenyi An[1,2], Joy Loh[7], Jill Fisher[4], Qihua Sun[1,2], Stanley Korsmeyer[4]‡, Milton Packer[8], Herman I. May[2], Joseph A. Hill[2], Herbert W. Virgin[7], Christopher Gilpin[9], Guanghua Xiao[8], Rhonda Bassel-Duby[5], Philipp E. Scherer[2,6] & Beth Levine[1,2,3,10]

Exercise has beneficial effects on human health, including protection against metabolic disorders such as diabetes[1]. However, the cellular mechanisms underlying these effects are incompletely understood. The lysosomal degradation pathway, autophagy, is an intracellular recycling system that functions during basal conditions in organelle and protein quality control[2]. During stress, increased levels of autophagy permit cells to adapt to changing nutritional and energy demands through protein catabolism[3]. Moreover, in animal models, autophagy protects against diseases such as cancer, neurodegenerative disorders, infections, inflammatory diseases, ageing and insulin resistance[4–6]. Here we show that acute exercise induces autophagy in skeletal and cardiac muscle of fed mice. To investigate the role of exercise-mediated autophagy *in vivo*, we generated mutant mice that show normal levels of basal autophagy but are deficient in stimulus (exercise- or starvation)-induced autophagy. These mice (termed BCL2 AAA mice) contain knock-in mutations in BCL2 phosphorylation sites (Thr69Ala, Ser70Ala and Ser84Ala) that prevent stimulus-induced disruption of the BCL2–beclin-1 complex and autophagy activation. BCL2 AAA mice show decreased endurance and altered glucose metabolism during acute exercise, as well as impaired chronic exercise-mediated protection against high-fat-diet-induced glucose intolerance. Thus, exercise induces autophagy, BCL2 is a crucial regulator of exercise- (and starvation)-induced autophagy *in vivo*, and autophagy induction may contribute to the beneficial metabolic effects of exercise.

To investigate whether autophagy is induced by exercise, we analysed tissues of mice that transgenically express a green fluorescent protein (GFP)-labelled marker of autophagosomes, GFP–LC3 (LC3 is also known as MAP1LC3) (ref. 7), after treadmill exercise. In both skeletal and cardiac muscle, autophagosome (GFP–LC3 puncta) numbers increased after 30 min (~300 m) of running and reached a plateau at 80 min (~900 m) (Fig. 1a–c). This was observed in several muscle groups, including the vastus lateralis (thigh muscle) (Fig. 1a–c), soleus, tibialis anterior and extensor digitorum longus (Supplementary Fig. 1). Exercise also resulted in biochemical evidence of skeletal and cardiac muscle autophagy, including conversion of the non-lipidated form of LC3, LC3-I, to the autophagosome-membrane-associated lipidated form, LC3-II, and degradation of the autophagy substrate protein p62 (Fig. 1d). In addition, exercise induced autophagy in other organs involved in glucose and energy homeostasis, such as liver and pancreas, including islet β-cells (Supplementary Fig. 2a–f) and (based on LC3-II conversion) adipose tissue (Supplementary Fig. 2g). Thus, exercise is a newly defined stimulus that induces autophagy *in vivo*.

We found that exercise-induced autophagy *in vivo* involves disruption of the BCL2–beclin-1 complex. BCL2 is an anti-apoptotic and anti-autophagy protein that inhibits autophagy through a direct interaction with the BH3 domain of the autophagy protein beclin 1 at the endoplasmic reticulum[8]. Disruption of the BCL2–beclin-1 complex is crucial for stimulus-induced autophagy in mammalian cells[8]. Beclin 1 immunoprecipitation with BCL2 decreased in muscle within 15 min of exercise and was barely detectable after 30 min (Fig. 1e). Dissociation of the BCL2–beclin-1 complex was not associated with phosphorylation of kinases previously linked to BCL2 phosphorylation, such as JNK and p38 MAPK[9] (Supplementary Fig. 3).

To study the physiological functions of exercise-induced autophagy *in vivo*, we constructed mutant mice that are deficient in stimulus-induced but not basal autophagy. As phosphorylation of three sites in the non-structured loop of human BCL2 is critical for stimulus-induced autophagy *in vitro*[8], we generated knock-in ($Bcl2^{AAA}$) mice

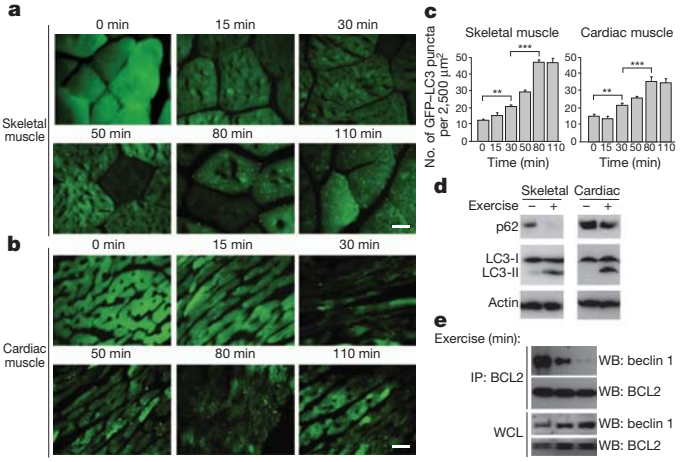


Figure 1 | Exercise induces autophagy in skeletal and cardiac muscle. **a**, **b**, Representative images of GFP–LC3 puncta (autophagosomes) in skeletal (vastus lateralis) (**a**) and cardiac (**b**) muscle from GFP–LC3 transgenic mice at serial time points after exercise. Scale bar, 20 μm. **c**, Quantification of data (mean ± s.e.m. of 10 tissue sections) in **a** and **b**. $^{**}P < 0.01$, $^{***}P < 0.001$ (one-way ANOVA). **d**, Western blot analysis of LC3-I/II (non-lipidated and lipidated forms of MAP1LC3, respectively) and p62 levels in indicated tissue from mice at rest (−) or after maximal exercise (+). Skeletal and cardiac indicate skeletal and cardiac muscle, respectively. **e**, Co-immunoprecipitation of beclin 1 with BCL2 in muscle tissue from mice at indicated time points after exercise. IP, immunoprecipitate; WB, western blot; WCL, whole cell lysates.

[1]Center for Autophagy Research, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [2]Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [3]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [4]Dana-Farber Cancer Institute, Massachusetts 02115, USA. [5]Department of Molecular Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [6]Touchstone Diabetes Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [7]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 6110, USA. [8]Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [9]Department of Cell Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. [10]Department of Microbiology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. †Present address: University of California at San Francisco, San Francisco, California 94158, USA.
*These authors contributed equally to this work.
‡Deceased.

lacking three conserved phosphorylation residues in the non-structured loop region of mouse BCL2: Thr 69, Ser 70 and Ser 84 (homologous to human Ser 87) were replaced by alanines (BCL2 AAA) (Supplementary Fig. 4a–c). Homozygous *Bcl2*^{AAA} mice were viable and fertile, born in the expected Mendelian ratio, of normal size and weight, and displayed normal histology of major organs (data not shown). Steady-state levels of BCL2 AAA in *Bcl2*^{AAA} mice were similar to those of wild-type (WT) BCL2 in *Bcl2*^{WT} mice in skeletal and cardiac muscle, liver, adipose tissue and pancreas (Supplementary Fig. 4d).

Murine embryonic fibroblasts (MEFs) derived from BCL2 AAA mice showed a defect in stimulus-induced autophagy (Fig. 2a, b). In response to starvation, BCL2 underwent phosphorylation, BCL2 dissociated from beclin 1, and autophagy was induced in MEFs from isogenic wild-type mice. However, in BCL2 AAA MEFs, BCL2 phosphorylation was absent (as measured by p32 labelling), the BCL2–beclin-1 complex was not disrupted, and less autophagy was induced. Levels of basal autophagy were similar in wild-type and BCL2 AAA MEFs. Similar findings were observed *in vivo* in skeletal and cardiac muscle (Supplementary Fig. 5); autophagosome numbers were similar at baseline in wild-type and BCL2 AAA mice expressing GFP–LC3 but failed to increase in GFP–LC3 BCL2 AAA mice in response to 48 h starvation.

To evaluate whether BCL2 AAA mice are deficient in exercise-induced autophagy, we exercised GFP–LC3 wild-type mice and GFP–LC3 BCL2 AAA mice for a fixed time and fixed distance (80 min (~900 m)) and at 75% of their maximal running capacity (see Fig. 3a). Under both conditions, BCL2 AAA mice displayed marked impairment of exercise-induced upregulation of skeletal and cardiac muscle (as measured by numbers of GFP–LC3 puncta) (Fig. 2c, d and Supplementary Fig. 1b). BCL2 AAA mice also exhibited less exercise-induced LC3-II conversion and p62 degradation in skeletal and cardiac muscle (Fig. 2e), impaired autophagic responses in liver and pancreatic β-cells (Supplementary Fig. 2c, f), and defective exercise-induced dissociation of the BCL2–beclin-1 complex in muscle (Fig. 2f). Thus, non-phosphorylatable BCL2 does not alter basal autophagy *in vivo*, but prevents autophagy activation in response to starvation and exercise. This blockade of autophagy activation was not associated with increased cell death (Supplementary Fig. 6).

The deficient exercise-induced autophagy in BCL2 AAA mice was accompanied by lower maximal exercise capacity (Fig. 3a). BCL2 AAA mice had similar levels of serum lactate after maximal exercise, similar baseline muscle strength, and similar baseline cardiac function and size as wild-type mice (Supplementary Fig. 7). Characterization of various muscle groups (soleus, tibialis anterior and extensor digitorum longus) from wild-type and BCL2 AAA mice at baseline and after 80 min of exercise revealed no differences in weight, fibre cross-sectional area, morphology (haematoxylin and eosin staining), fibre type analysis, glycogen content, or mitochondrial content and functionality (Fig. 3b and Supplementary Figs 8–10). These analyses suggest that differences in baseline cardiac and skeletal muscle properties do not account for decreased exercise endurance in BCL2 AAA mice.

During strenuous exercise, mammals undergo metabolic changes to increase skeletal muscle glucose uptake/utilization efficiency, including increased insulin sensitivity and redistribution of glucose transporters, such as GLUT4 (also known as SLC2A4), to the plasma membrane[10,11], which is essential for exercise-stimulated glucose uptake[12]. However, BCL2 AAA mice demonstrated impaired exercise-induced increase in insulin sensitivity, as reflected by less of a decline in plasma glucose (Fig. 3c) and plasma insulin (Fig. 3d) levels than in wild-type mice. Also, unlike wild-type mice, BCL2 AAA mice failed to exhibit increased plasma membrane GLUT4 localization in the vastus lateralis and soleus muscles following maximal or 80 min exercise, respectively (Fig. 3e and Supplementary Figs 11a and 12a) and they exhibited lower levels of radiolabelled glucose uptake in soleus muscle (Fig. 3h).

Adenosine monophosphate-activated protein kinase (AMPK) has a central role in enhanced GLUT4 plasma membrane localization and skeletal muscle glucose uptake during exercise[13]. BCL2 AAA mice had a notable decrease in AMPK activation, as measured by levels of phosphorylation at residue threonine 172, in both the vastus lateralis (Fig. 3f and Supplementary Fig. 11b) and in the soleus (Supplementary Fig. 12b) muscles. Phosphorylation of the downstream AMPK target acetyl-CoA carboxylase (ACC) paralleled AMPK phosphorylation (Supplementary Fig. 13), providing additional support for a defect in exercise-induced AMPK activation in BCL2 AAA mice.

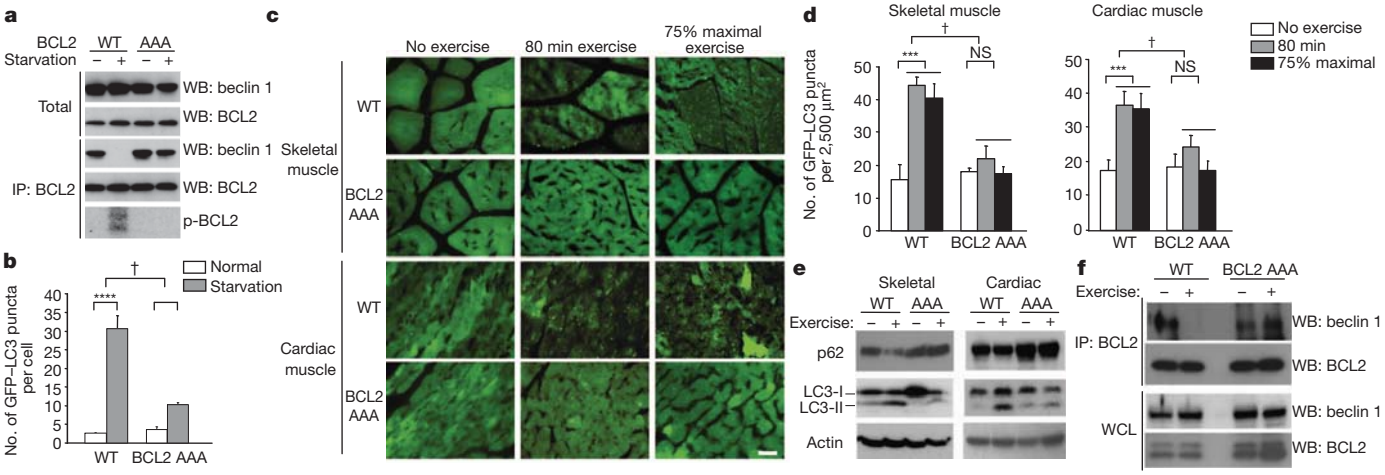


Figure 2 | Non-phosphorylatable BCL2 AAA knock-in mutations block BCL2 phosphorylation, BCL2–beclin 1 dissociation, and starvation- and exercise-induced autophagy. **a**, Analysis of BCL2 phosphorylation (detected by anti-BCL2 immunoprecipitation and autoradiography of ^{32}P-labelled cells) and beclin 1 co-immunoprecipitation with BCL2 in wild-type (WT) or BCL2 AAA MEFs grown in normal media or subjected to 4 h Earle's balanced salt solution (EBSS) starvation. p-BCL2, phospho-BCL2. **b**, Quantification of GFP–LC3 puncta (autophagosomes) in MEFs of indicated genotype in normal growth conditions or starvation conditions. Data represent mean ± s.e.m. for 100 cells per well of triplicate samples per condition. Similar results were observed in three independent experiments. **c**, Representative images of GFP–LC3 puncta (autophagosomes) in skeletal and cardiac muscle of GFP–LC3 wild-type and GFP–LC3 BCL2 AAA mice before exercise, after 80 min exercise, or after 75% of maximal exercise capacity. Scale bar, 20 μm. **d**, Quantification of data (mean ± s.d. of 4 mice per experimental group) in **c**. **e**, Western blot analysis of LC3-I/II and p62 levels in indicated tissue from mice of indicated genotype at rest (−) or after maximal exercise (+). Skeletal and cardiac indicate skeletal and cardiac muscle, respectively. **f**, Co-immunoprecipitation of beclin 1 with BCL2 in muscle tissue from mice of indicated genotype at rest (−) or after 30 min of exercise. WCL, whole cell lysates. NS, not significant. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, one-way ANOVA for comparison between groups; $\dagger P < 0.001$, two-way ANOVA for comparison of magnitude of changes between different groups in mice of different genotypes.

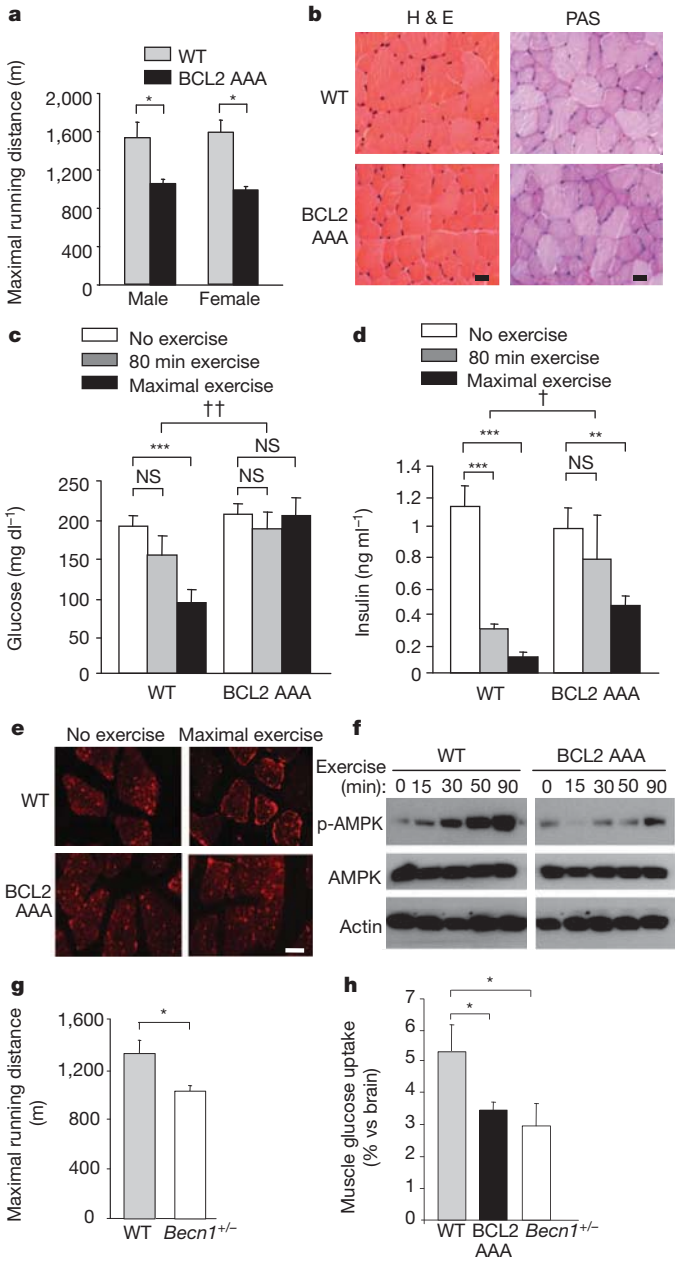


Figure 3 | BCL2 AAA mice show deficient exercise endurance and alterations in muscle glucose metabolism. **a**, Maximal treadmill running distance for mice of indicated genotype. Data represent mean ± s.e.m. of 5 mice per group. **b**, Representative haematoxylin and eosin (H & E) and periodic acid-Schiff (PAS) staining in tibialis anterior muscle sections from mice of indicated genotype. Scale bar, 20 μm. **c**, **d**, Plasma glucose (**c**) and insulin (**d**) levels in mice of indicated genotype at rest, after 80 min exercise (~900 m), or maximal exercise. Data represent combined mean ± s.e.m. for 9–11 mice per group from three independent cohorts; similar results were observed in each cohort. **e**, Representative images of GLUT4 immunofluorescence staining in vastus lateralis muscle of mice of indicated genotype before and after maximal exercise. Scale bar, 20 μm. For **b** and **e**, similar results were observed in 3 mice per group. **f**, Western blot analysis of AMPK phosphorylation (p-AMPK (Thr 172)) in vastus lateralis muscle lysates from mice of indicated genotype at indicated time after exercise. **g**, Maximal treadmill running distance for mice of indicated genotype. Data represent mean ± s.e.m. of 4–6 male mice per group. **h**, Soleus muscle ^{14}C-deoxyglucose uptake during treadmill exercise in mice of indicated genotype. Data represent mean ± s.e.m. of 3 mice per group. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, one-way ANOVA for comparison between groups; $\dagger P < 0.05$, $\dagger\dagger P < 0.01$, two-way ANOVA for comparison of magnitude of changes between different groups in mice of different genotypes. NS, not significant.

No differences were observed in phosphorylation of CaMK1, a downstream target of the Ca^{2+}-stimulated kinase CaMKKβ (which is upstream of AMPK in certain contexts)[13] (Supplementary Fig. 13b), suggesting that differences in levels of AMPK activation between wild-type and BCL2 AAA mice are not due to altered Ca^{2+} homeostasis. They are also probably unrelated to possible effects of BCL2 on mitochondrial function, because phosphorylated BCL2 resides at the endoplasmic reticulum[14] and markers of mitochondrial function did not differ in the muscles of BCL2 AAA and wild-type mice (Supplementary Fig. 10). Differences in insulin signalling molecules also did not explain the altered glucose homeostasis in exercised BCL2 AAA mice, because at late time points during exercise, decreased (not increased) AKT activation and mTOR activation was observed in muscle of both wild-type and BCL2 AAA mice (Supplementary Fig. 13a). No differences were observed in muscle phosphorylation of the AKT substrate AS160 (also known as TBC1D4), or levels of the muscle-specific forkhead transcription factor target atrogin 1 (also known as FBXO32), in response to exercise or between genotypes.

To confirm that the observed defects in exercise-induced endurance and muscle glucose metabolism are due to defective autophagy, we examined mice with allelic loss of the autophagy gene beclin 1 (*Becn1*), which have decreased beclin 1 protein expression and decreased starvation-induced autophagy in skeletal muscle[15]. Similar to BCL2 AAA mice, *Becn1*$^{+/-}$ mice exhibited deficient exercise-induced autophagy in skeletal muscle (Supplementary Fig. 14a, b), normal muscle strength (Supplementary Fig. 14c), decreased maximal treadmill running distance (Fig. 3g), impaired exercise-induced GLUT4 plasma membrane localization (Supplementary Fig. 14d) and muscle glucose uptake (Fig. 3h), and lower levels of exercise-induced muscle AMPK activation (phospho-AMPK and its target, phospho-ACC) (Supplementary Fig. 14e) than wild-type controls. The similarity of the phenotypes of BCL2 AAA and *Becn1*$^{+/-}$ mice provides strong support for a role of deficient beclin 1 activity, rather than other BCL2-regulated functions, in the impairment of exercise endurance, glucose uptake and AMPK activation in BCL2 AAA mice.

AMPK activates autophagy in response to low energy conditions by sensing the cellular ratio of AMP to ATP[13]. Deficient AMPK activation in BCL2 AAA and *Becn1*$^{+/-}$ mice during exercise raised the possibility that, similar to observations *in vitro*[16], the autophagy pathway may function *in vivo* in a feed-forward manner in AMPK activation. To evaluate whether impaired exercise-induced AMPK activation *in vivo* is due to deficient autophagy, rather than potential autophagy-independent effects of the BCL2 AAA mutation or monoallelic loss of beclin 1, we examined exercise-induced AMPK activation in *Atg16l1*^{HM} mice that are hypomorphic for ATG16L1 (ref. 17), an essential autophagy protein that specifies the LC3 lipidation site for autophagosomal membrane biogenesis[18]. *Atg16l1*^{HM} mice had a significant reduction in ATG16L1 protein expression in skeletal muscle and a defect in exercise-induced autophagy in muscle lysates after 80 min of exercise (Supplementary Fig. 15a, b). This defect in exercise-induced autophagy was associated with decreased exercise-induced AMPK phosphorylation and ACC phosphorylation (Supplementary Fig. 15c).

Together, our *in vivo* studies in BCL2 AAA, *Becn1*$^{+/-}$ and *Atg16l1*^{HM} mice suggest that cellular autophagy function is partially required for normal levels of exercise-induced muscle AMPK activation. The mechanism underlying this requirement is not yet known, but is unlikely to be caused by non-specific effects of the mouse mutations on AMPK activation as the direct AMPK activator, the AMP analogue 5-aminoimidazole-4-carboxamide riboside (AICAR), induced AMPK activation normally in BCL2 AAA, *Becn1*$^{+/-}$ and *Atg16l1*^{HM} MEFs (Supplementary Fig. 16).

We next examined whether exercise-induced autophagy is required for the beneficial metabolic effects of long-term exercise training, using a high-fat diet (HFD) model of obesity and impaired glucose tolerance[6,19]. Age-matched cohorts of wild-type and BCL2 AAA mice were fed a HFD

for four weeks; the HFD was continued for an additional 8 weeks either combined with 50 min of daily treadmill exercise or with no exercise (Supplementary Fig. 17a). In both wild-type and BCL2 AAA mice, HFD led to body weight gain (Supplementary Fig. 17b), which was probably due to increased fat mass (Supplementary Fig. 18a), as no differences in lean mass or muscle fibre size were observed (Supplementary Fig. 18). HFD led to modest suppression of baseline autophagy in skeletal and cardiac muscle as measured by levels of LC3-II conversion and p62 in tissue lysates (Supplementary Fig. 19). Ultrastructural analyses of the liver and pancreas did not reveal any differences in HFD versus regular diet animals, in wild-type versus BCL2 AAA mice, or in non-exercised versus exercised animals, except for hepatic lipid droplet accumulation in all HFD-fed groups (Supplementary Fig. 20). Although BCL2 AAA mice were more sensitive to HFD-induced obesity, 8 weeks of daily exercise reduced the magnitude of weight gain similarly in BCL2 AAA and wild-type mice (Supplementary Fig. 17b). Thus, the BCL2 AAA mutation did not alter the response of mice to HFD with respect to muscle fibre size, the morphology of liver and pancreas, or the effect of exercise on HFD-induced obesity.

However, BCL2 AAA mice failed to exhibit normal exercise-induced protection against HFD-induced impaired glucose tolerance. Before the administration of HFD, wild-type and BCL2 AAA mice had similar clearance rates in oral glucose tolerance tests (Fig. 4a). Four weeks of HFD led to impaired glucose tolerance in both genotypes versus regular diet control groups (Fig. 4b); the magnitude of this impairment was similar in BCL2 AAA and wild-type mice despite more weight gain in the BCL2 AAA mice. However, exercise training markedly improved glucose tolerance in HFD-fed wild-type mice but not in autophagy-deficient BCL2 AAA mice (Fig. 4c and Supplementary Fig. 21). This failure of BCL2 AAA mice to show improved glucose tolerance is unlikely to be due to deficient insulin production, because at the end of the study HFD-fed wild-type and BCL2 AAA mice displayed similar circulating fasting insulin levels (Supplementary Fig. 22a), similar levels of insulin secretion in response to oral glucose challenge (Supplementary Fig. 22b), and similar pancreatic β-cell morphology (Supplementary Fig. 20b). These results suggest that BCL2-regulated functions are essential for chronic exercise-mediated protection against HFD-induced glucose intolerance. We speculate that, as in acute exercise, this may involve the dynamic interplay of autophagy induction and AMPK activation.

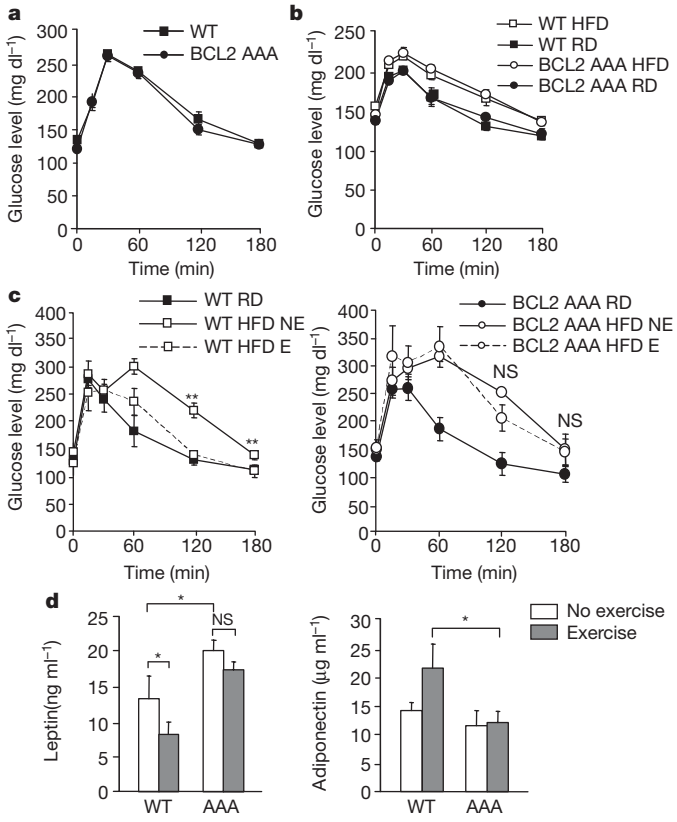


Figure 4 | Long-term exercise training protects wild-type but not BCL2 AAA mice from HFD-induced glucose intolerance. **a**, **b**, Oral glucose tolerance test (OGTT) before (**a**, week 0) and after (**b**, week 4) 4 weeks of HFD. **c**, **d**, OGTT (**c**) and serum leptin and adiponectin levels (**d**) after 8 weeks of daily exercise. For **a**–**d**, results represent the mean ± s.e.m. for 4–5 mice per group. E, exercise; NE, no exercise; RD, regular diet . $*P < 0.05$, $**P < 0.01$, (**c**, one-way ANOVA; **d**, Wilcoxon rank test). NS, not significant.

HFD-induced obesity is associated with several other metabolic changes, including increased levels of serum leptin, an appetite-inhibiting adipokine;[20] reduced levels of circulating adiponectin[21], an antidiabetic adipokine[22–24]; and increased serum triglycerides and cholesterol. In HFD-fed mice, baseline levels of leptin were higher in BCL2 AAA mice than in wild-type mice, and 8 weeks of exercise decreased serum leptin in wild-type but not in BCL2 AAA mice (Fig. 4d). Although adiponectin levels in HFD-fed mice were similar in the non-exercised wild-type and BCL2 AAA groups, they were significantly increased in the exercised wild-type mice versus the exercised BCL2 AAA mice (Fig. 4d). Both serum triglycerides and cholesterol increased in wild-type and BCL2 AAA HFD-fed mice; with exercise, the values returned to those observed in animals fed a regular diet in wild-type but not BCL2 AAA mice (Supplementary Fig. 23). Thus, the BCL2 AAA mutation impaired the beneficial effects of exercise on metabolic changes induced by a HFD.

The HFD-fed, daily exercised wild-type mice were more metabolically active than their BCL2 AAA counterparts, as demonstrated by elevated levels of oxygen consumption, CO_2 production and heat generation during a 12 h night interval when the mice were not exercised (Supplementary Fig. 24a). In addition, in vastus lateralis muscle, wild-type mice had elevated induction of messenger RNA for UCP1, *Ucp1*, a mitochondrial uncoupling protein mainly expressed in brown adipose tissue that contributes to thermogenesis and energy expenditure[25] (Supplementary Fig. 24d). These changes were probably due to improved metabolic fitness in response to exercise rather than alterations in food intake, spontaneous physical activity, or general mitochondrial function. No differences were observed in daily food intake in exercised HFD-fed wild-type and BCL2 AAA mice (Supplementary Fig. 24b). Spontaneous physical activity was similar in HFD-fed non-exercised wild-type and BCL2 AAA mice, although there was a trend towards decreased spontaneous physical activity in the exercised BCL2 AAA mice (Supplementary Fig. 24c). No changes were observed with exercise in either genotype in mRNA levels of the skeletal muscle-expressed uncoupling proteins UCP2 or UCP3, or the mitochondrial proteins cytochrome *b* and *c* (Supplementary Fig. 24d, e).

The HFD study suggests that increased autophagy triggered by exercise may be critical for improving impaired glucose tolerance and metabolism in diet-induced obesity. However, we cannot definitely conclude that lack of exercise-induced improvement in glucose tolerance in HFD-fed BCL2 AAA mice is caused by deficient exercise-induced autophagy; it is possible that other effects of the BCL2 AAA mutation are responsible for this phenotype. Nonetheless, given our findings in acute exercise (which demonstrate impaired muscle glucose uptake, GLUT4 plasma membrane localization and AMPK activation in autophagy-deficient animals), it seems plausible that alterations in exercise-induced skeletal muscle glucose metabolism in autophagy-deficient animals may also contribute to the failure of exercise to reverse HFD-induced metabolic abnormalities.

Our findings demonstrate that exercise is a potent inducer of autophagy, and that acute and chronic exercise enhances glucose metabolism in mice capable of inducing autophagy but not in autophagy-deficient mice. These beneficial metabolic effects (as well as exercise- and starvation-induced autophagy) are blocked by a

mutation in BCL2 that prevents its release from an inhibitory interaction with the autophagy protein beclin 1. Thus, BCL2 has previously undescribed essential roles in the *in vivo* regulation of stimulus-induced autophagy as well as glucose metabolism. We propose that BCL2-regulated autophagy activation contributes to the beneficial metabolic effects of exercise, and that manipulation of the autophagy pathway and/or the function of the autophagy inhibitory BCL2 protein may be a logical strategy to mimic the health effects of exercise and to prevent or treat impaired glucose metabolism. More broadly, on the basis of this newly discovered link between exercise, autophagy and altered metabolism, we speculate that autophagy may represent a cellular mechanism by which exercise prolongs life and protects against cancer, cardiovascular disorders and inflammatory diseases[1].

METHODS SUMMARY

Mouse strains. GFP–LC3 transgenic[7], *Becn1*$^{+/-}$ (ref. 15) and *Atg16l1*^{HM} (ref. 17) mice have been described. The generation of mice with a knock-in mutation in *Bcl2* that results in Thr69Ala, Ser70Ala and Ser84Ala mutations in BCL2 (BCL2 AAA) is described in Methods.

Exercise studies. Acute and chronic exercise studies were performed using a treadmill protocol described in Methods.

Autophagy assays. Autophagy was measured by visualizing GFP–LC3 puncta by fluorescence microscopy in MEFs or tissue sections from mice with transgenic expression of GFP–LC3, or by western blot analysis of LC3-II conversion and p62 levels in tissue lysates.

Metabolic analyses. Metabolic and blood parameters in the acute and chronic exercise studies were measured as described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 September 2010; accepted 6 December 2011.
Published online 18 January 2012.

1. Handschin, C. & Spiegelman, B. M. The role of exercise and PGC1α in inflammation and chronic disease. *Nature* **454,** 463–469 (2008).
2. Mizushima, N. & Levine, B. Autophagy in mammalian development and differentiation. *Nature Cell Biol.* **12,** 823–830 (2010).
3. Kuma, A. & Mizushima, N. Physiological role of autophagy as an intracellular recycling system: with an emphasis on nutrient metabolism. *Semin. Cell Dev. Biol.* **21,** 683–690 (2010).
4. Levine, B. & Kroemer, G. Autophagy in the pathogenesis of disease. *Cell* **132,** 27–42 (2008).
5. Yang, L., Li, P., Fu, S., Calay, E. S. & Hotamisligil, G. S. Defective hepatic autophagy in obesity promotes ER stress and causes insulin resistance. *Cell Metab.* **11,** 467–478 (2010).
6. Ebato, C. *et al.* Autophagy is important in islet homeostasis and compensatory increase of β cell mass in response to high-fat diet. *Cell Metab.* **8,** 325–332 (2008).
7. Mizushima, N., Yamamoto, A., Matsui, M., Yoshimori, T. & Ohsumi, Y. *In vivo* analysis of autophagy in response to nutrient starvation using transgenic mice expressing a fluorescent autophagosome marker. *Mol. Biol. Cell* **15,** 1101–1111 (2004).
8. He, C. & Levine, B. The Beclin 1 interactome. *Curr. Opin. Cell Biol.* **22,** 140–149 (2010).
9. Wei, Y., Pattingre, S., Sinha, S., Bassik, M. & Levine, B. JNK1-mediated phosphorylation of Bcl-2 regulates starvation-induced autophagy. *Mol. Cell* **30,** 678–688 (2008).
10. Goodyear, L. J. & Kahn, B. B. Exercise, glucose transport, and insulin sensitivity. *Annu. Rev. Med.* **49,** 235–261 (1998).
11. Ryder, J. W., Chibalin, A. V. & Zierath, J. R. Intracellular mechanisms underlying increases in glucose uptake in response to insulin or exercise in skeletal muscle. *Acta Physiol. Scand.* **171,** 249–257 (2001).
12. Zisman, A. *et al.* Targeted disruption of the glucose transporter 4 selectively in muscle causes insulin resistance and glucose intolerance. *Nature Med.* **6,** 924–928 (2000).
13. Hardie, D. G. AMP-activated protein kinase: an energy sensor that regulates all aspects of cell function. *Genes Dev.* **25,** 1895–1908 (2011).
14. Bassik, M. C., Scorrano, L., Oakes, S. A., Pozzan, T. & Korsmeyer, S. J. Phosphorylation of BCL-2 regulates ER Ca^{2+} homeostasis and apoptosis. *EMBO J.* **23,** 1207–1216 (2004).
15. Qu, X. *et al.* Promotion of tumorigenesis by heterozygous disruption of the beclin 1 autophagy gene. *J. Clin. Invest.* **112,** 1809–1820 (2003).
16. Malik, S. A. *et al.* BH3 mimetics activate multiple pro-autophagic pathways. *Oncogene* **30,** 3918–3928 (2011).
17. Cadwell, K. *et al.* A key role for autophagy and the autophagy gene *Atg16l1* in mouse and human intestinal Paneth cells. *Nature* **456,** 259–263 (2008).
18. Fujita, N. *et al.* The Atg16L complex specifies the site of LC3 lipidation for membrane biogenesis in autophagy. *Mol. Biol. Cell* **19,** 2092–2100 (2008).
19. Asterholm, I. W. & Scherer, P. E. Enhanced metabolic flexibility associated with elevated adiponectin levels. *Am. J. Pathol.* **176,** 1364–1376 (2010).
20. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372,** 425–432 (1994).
21. Scherer, P. E., Williams, S., Fogliano, M., Baldini, G. & Lodish, H. F. A novel serum protein similar to C1q, produced exclusively in adipocytes. *J. Biol. Chem.* **270,** 26746–26749 (1995).
22. Frederich, R. C. *et al.* Leptin levels reflect body lipid content in mice: evidence for diet-induced resistance to leptin action. *Nature Med.* **1,** 1311–1314 (1995).
23. Van Heek, M. *et al.* Diet-induced obese mice develop peripheral, but not central, resistance to leptin. *J. Clin. Invest.* **99,** 385–390 (1997).
24. Shetty, S., Kusminski, C. M. & Scherer, P. E. Adiponectin in health and disease: evaluation of adiponectin-targeted drug development strategies. *Trends Pharmacol. Sci.* **30,** 234–239 (2009).
25. Almind, K., Manieri, M., Sivitz, W. I., Cinti, S. & Kahn, C. R. Ectopic brown adipose tissue in muscle provides a mechanism for differences in risk of metabolic syndrome in mice. *Proc. Natl Acad. Sci. USA* **104,** 2366–2371 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the UT Southwestern Mouse Metabolic Phenotyping Core and E. Berglund for assistance with metabolic measurements, J. Shelton for assistance with muscle stains, N. Mizushima for critical reagents, and B. D. Levine for expert advice. This work was supported by National Institutes of Health grants RO1 CA109618 (B.L.), ROI HL080244 (J.A.H.), ROI HL090842 (J.A.H.), ROI AI084887 (H.W.V.), RCI DK086629 (P.E.S.), RO1 CA112023 (P.E.S.) and 1PO1 DK0887761 (P.E.S.).

Author Contributions C.H., M.C.B., V.M., K.S., S.K., M.P., J.A.H., H.W.V., R.B.-D., P.E.S. and B.L. designed the experiments. C.H., M.C.B., V.M., K.S., Y.W., Z.Z., Z.A., J.L., J.F., Q.S., H.I.M. and C.G. performed the experiments. G.X. performed statistical analyses. C.H. and B.L. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.L. (beth.levine@utsouthwestern.edu).

METHODS

Mouse strains. GFP–LC3 transgenic, *Becn1*$^{+/-}$ and *Atg16l1*^{HM} mice have been previously described[7,15,17]. For the construction of a mouse strain with knock-in mutations in the phosphorylation sites in the non-structured loop of BCL2 (*Bcl2*^{AAA} mice), BAC clones (Incyte) were screened for the presence of *Bcl2* using the following primers: BCL2 5′, GTGGGGCGGGAGTCGGGACT; BCL2 3′, GACCCAGAATCCACTCACAC. The *Bcl2* BAC clone was digested by BglII, subcloned into pSP72, and a puromycin resistance marker flanked by FRT sites (SalI fragment of pPGKPuro, cloned into pFRT) was blunt cloned into a BsaBI site 3′ of exon II. A mutant allele was generated by subcloning a PstI fragment from the coding region. Amino acid residues Thr 69, Ser 70 and Ser 84 (homologous to human Ser 87) were changed to alanine in two steps using the Quickchange PCR (Stratagene) and the following primer sets: TSAA 5′, GAGATGGCTGCCA GGGCGGCTCCTCTCAGGCCCC; TSAA 3′, GGGGCCTGAGAGGAGCC GCCCTGGCAGCCATCTC; S84A 5′, GCTGGGCCTGCGCTCGCCCC TGTGC CACCATG; S84A 3′, CATGGTGGCACAGGGGCGAGCGCAGGCCCAGC. The targeting construct 7.1AAA was electroporated into RW4 embryonic stem cells (129 Sv/J), and 36 h later, clones were selected with puromycin, picked and screened by Southern blot analysis with the probes indicated in Supplementary Fig. 4a. A 1.7 kb shift upwards indicative of the mutant allele was detected on Southern blots in 2/300 clones as the targeted locus (Supplementary Fig. 4b). These clones were tested for normal karyotype and used to inject blastocysts from C57BL/6J donors. Mice with germline transmission were bred to mice expressing FLP from the β-actin promoter (Jackson Laboratories: B6; SJLTgN(ACTFLPe)9205Dym), and offspring were screened for deletion of the puromycin selection marker. To screen for the presence of the mutation following excision of the selectable marker, PCR with 5′ and 3′ *Bcl2* primers was performed, and digestions were performed with BglI or AatII enzymes; BglI cuts the amplified fragment only if the T69AS70A site is present, while AatII cuts only the wild-type sequence for the same region (Supplementary Fig. 4c). Using this scheme, homozygous *Bcl2*^{AAA} mutants were identified by the presence of PCR fragments cleaved only by BglI and not AatII, whereas wild-type mice were identified by the presence of PCR fragments cleaved only by AatII and not BglI. *Bcl2*^{AAA} mice were backcrossed for more than ten generations to C57/B6 mice (Jackson Laboratories), and homozygous *Bcl2*^{AAA} and *Bcl2*^{WT} offspring were used in all studies.

Cell lines. Primary MEFs were established either from *Bcl2*^{AAA} and *Bcl2*^{WT} mice, the offspring of *Bcl2*^{AAA} and *Bcl2*^{WT} mice crossed with GFP–LC3 transgenic mice, *Becn1*$^{+/-}$ and *Becn1*$^{+/+}$ mice, and *Atg16l1*^{HM} and *Atg16l1*^{WT} at embryonic day 13.5 and cultured as described[26].

Radiolabelling and co-immunoprecipitation from MEFs. MEFs derived from *Bcl2*^{AAA} and *Bcl2*^{WT} mice were cultured overnight in labelling medium (phosphate-free DMEM with 10% dialysed FBS) and then cultured in starvation medium (EBSS minus phosphate) or labelling medium for 4 h containing 2 mCi ml^{-1} ^{32}P-orthophosphate. Cells were lysed in lysis buffer containing 50 mM Tris (pH 7.9), 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, proteinase inhibitor cocktail (Roche Applied Sciences) and halt phosphatase inhibitor cocktail (Thermo Scientific), and were subjected to immunoprecipitation with a monoclonal anti-BCL2 antibody (Santa Cruz Biotechnology, 1:50). Eluates were separated by SDS–PAGE and detected by anti-beclin 1 antibody (Santa Cruz Biotechnology, 1:200), anti-BCL2–HRP antibody (C2, Santa Cruz Biotechnology, 1:100) and autoradiography.

Co-immunoprecipitations from muscle tissue. Vastus lateralis (thigh muscle) was dissected and homogenized in lysis buffer containing 50 mM Tris (pH 7.9), 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, proteinase inhibitor cocktail (Roche Applied Sciences) and halt phosphatase inhibitor cocktail (Thermo Scientific), and subjected to immunoprecipitation with a monoclonal anti-BCL2 antibody (Santa Cruz Biotechnology, 1:50). Eluates were separated by SDS–PAGE and detected by anti-beclin 1 antibody (Santa Cruz Biotechnology, 1:200) and anti-BCL2–HRP antibody (C2, Santa Cruz Biotechnology, 1:100) and autoradiography.

Mouse exercise studies. For acute exercise studies, 8-week-old (wild type, BCL2 AAA and *Atg16l1*^{HM}) or 12-week-old (*Becn1*$^{+/-}$ and wild-type *Becn1*$^{+/+}$ littermate) mice were acclimated to and trained on a 10° uphill Exer 3/6 open treadmill (Columbus Instruments) for 2 days. On day 1 mice ran for 5 min at 8 m min^{-1} and on day 2 mice ran for 5 min at 8 m min^{-1} followed by another 5 min at 10 m min^{-1}. On day 3, mice were subjected to a single bout of running starting at the speed of 10 m min^{-1}. Forty minutes later, the treadmill speed was increased at a rate of 1 m min^{-1} every 10 min for a total of 30 min, and then increased at the rate of 1 m min^{-1} every 5 min until mice were exhausted. Exhaustion was defined as the point at which mice spent more than 5 s on the electric shocker without attempting to resume running. Total running time was recorded and total running distance was calculated for each mouse.

For long-term exercise training, male *Bcl2*^{WT} or *Bcl2*^{AAA} mice (8 weeks old) were randomly divided into three cohorts, including: (1) mice fed a regular diet (Harlan Teklad) without daily exercise, (2) mice fed a HFD containing 60% fat (Research Diets) without daily exercise, and (3) mice fed a HFD with daily exercise. Prior to initiation of exercise, mice were fed a HFD for 4 weeks. At the end of the fourth week, mice in the exercise groups were acclimated to treadmill running for 2 days as in the acute exercise studies, and then were trained on the treadmill with 10° uphill incline for 50 min d^{-1}, 5 d week^{-1} at 17 m min^{-1} for 8 weeks. Mice were given a HFD during the 8-week training period. All animal procedures were performed in accordance with institutional guidelines and with approval from the Institutional Animal Care and Use Committee.

Autophagy analyses. MEFs expressing GFP–LC3 were cultured in normal or starvation (EBSS, Earle's balanced salt solution) medium for 4 h, and GFP–LC3 puncta were quantified by fluorescence microscopy as described[27]. For assessment of autophagy *in vivo* following starvation or exercise, 8-week-old *Bcl2*^{WT} or *Bcl2*^{AAA} GFP–LC3 mice or *Becn1*$^{+/+}$ or *Becn1*$^{+/-}$ GFP–LC3 mice were either subjected to starvation for 48 h or exercised for the indicated time period, anaesthetized by isoflurane, and perfused with 4% paraformaldehyde (PFA). Skeletal muscle (vastus lateralis, tibialis anterior, extensor digitorum longus and soleus), heart (left ventricle), pancreas and liver tissues were fixed in 4% PFA overnight, 15% sucrose for 4 h and 30% sucrose overnight before frozen sections were prepared. The number of GFP–LC3 puncta per unit area of tissue was quantified by fluorescence microscopy as described[15]. Autophagy in skeletal and cardiac muscle, liver, pancreas and adipose tissue during baseline conditions and after exercise was also analysed by western blot analysis of tissue extracts with antibodies against LC3 and p62 (see below for details).

Immunofluorescence studies. For immunofluorescence staining of frozen muscle and pancreatic sections, slides were heated at 50 °C for 10 min, rehydrated in 100%, 95% and 70% ethanol, washed 3 × 5 min in PBS, blocked with 1% goat serum and immunostained with a rabbit anti-mouse GLUT4 antibody (Alpha Diagnostic, 1:100 dilution) (muscle samples) and an Alexa Fluor 594 donkey anti-rabbit secondary antibody (Invitrogen, 1:1,000 dilution) or with a guinea pig anti-swine insulin antibody (Dako, 1:500 dilution) (pancreatic samples) and an Alexa Fluor 594 goat anti-guinea pig secondary antibody (Invitrogen, 1: 500 dilution). Immunofluorescence studies on skeletal muscle sections with anti-laminin and anti-type I myosin heavy chain antibodies were performed as previously described[28]. Immunofluorescence images were taken using a Zeiss Axioplan2 microscope or Leica TCS SP5 confocal microscope.

Muscle morphology and function studies. Sections were stained with haematoxylin and eosin to visualize tissue architecture following standard protocols. Metachromatic ATPase staining and measurement of fibre cross-sectional area were performed as previously described[28]. Periodic acid-Schiff staining for carbohydrates was performed according to standard protocols[29]. Succinic dehydrogenase staining on cryosections was carried out using a 0.2 M phosphate buffer at pH 7.6. For staining, 270 mg succinic acid and 10 mg nitro blue tetrazolium were freshly dissolved in 10 ml phosphate buffer and incubated with sections for 1 h.

To measure muscle strength, a modified mesh grip assay was applied[30]. Briefly, mice were placed on the wired mesh of cage tops. When they firmly grabbed the mesh, the cage top was flipped and held at a height of 30 cm above the bench until the mice released their grip. The time for each mouse to remain on the mesh was recorded. TUNEL staining of muscle sections to detect apoptosis was performed according to the manufacturer's instructions (ApopTag Peroxidase *In situ* Apoptosis Detection Kit, Millipore), using Sigma *FAST* 3,3′-diaminobenzidine (DAB) tablets as the peroxidase substrate.

Electron microscopy. Mice were euthanized, and liver and pancreas were rapidly fixed in 2.5% glutaraldehyde in 0.1 M cacodylate buffer. Electron microscopy was performed as previously described[31].

Mouse metabolic studies and blood tests. Food intake, O_2 consumption, CO_2 production, heat production and physical activity were measured in TSE metabolic chambers (TSE Systems) over a 4 day period. Mice in the exercise group were removed from the chambers and allowed to run for 50 min daily as in the above protocol during the day (light period). Metabolic measurements were performed for 12 h daily during the dark period. Fat mass and lean tissue mass were determined using the Bruker Minispec mq10 scan system. Plasma glucose, lactate, cholesterol and triglycerides were measured using the Vitros 250 system (Ortho-Clinical Diagnostics). Plasma insulin (Crystal Chem), serum leptin (Crystal Chem) and adiponectin (Millipore) levels were determined using commercial ELISA kits. Glucose tolerance tests were performed by oral glucose injection (1 g kg^{-1}) after a 3 h fast. Blood was drawn from tail veins at the indicated time points after glucose injection and serum glucose levels were analysed using commercial glucose assay reagents (Sigma, Cat# 6918, P7119, F5803).

Echocardiography. Cardiac function was assessed by echocardiography, using a Visualsonic Vevo 2100 ultrasound machine equipped with a 30-Mhz transducer applied to the chest wall. Ventricular dimensions and ejection fraction were assessed using short axis view in two-dimensional and three-dimensional modes using the onboard VisualSonics cardiac analysis package.

Muscle glucose uptake assays. Surgical catheterization of the jugular vein was performed on each mouse. At day 4 after catheterization, mice were transferred to clean cages at 9:00 a.m. to begin a 4 h fast. Mice then ran on a treadmill for 50 min using the exercise protocol described above. At $t = 50$ min, a 13 μCi bolus of ^{14}C-deoxyglucose was injected intravenously and mice resumed running for another 25 min according to the exercise protocol described above before euthanasia. Skeletal muscle (soleus) and brain tissues were dissected and homogenized in 1.5 ml 0.5% perchloric acid. ^{14}C-deoxyglucose counts were determined in homogenized samples as previously described[32] and muscle values were normalized to the counts in the brain.

Western blot analyses. Mouse tissue extracts were prepared by homogenizing tissues in lysis buffer containing 50 mM Tris (pH 7.9), 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, proteinase inhibitor cocktail (Roche Applied Sciences) and halt phosphatase inhibitor cocktail (Pierce), and subjected to western blot analysis with anti-LC3 (Novus Biologicals, 1:500), anti-p62 (BD Biosciences, 1:500 dilution; Progen (C-terminal specific), 1:1,000 dilution). Anti-p-CaMKI (Santa Cruz Biotechnology, 1:100), anti-CaMKI (Santa Cruz Biotechnology, 1:300), anti-BCL2 (Santa Cruz Biotechnology, 1: 100), anti-AMPK (Cell Signaling, 1:1,000), anti-p-AMPK (Cell Signaling, 1:1,000), anti-JNK (Cell Signaling, 1:400), anti-p-JNK (Cell Signaling, 1:500) anti-AS160 (Millipore, 1:400) anti-p-AS160 (Millipore, 1:100), anti-4E-BP1 (Cell Signaling, 1:1,000), anti-p-4E-BP1 (Cell Signaling, 1:1,000), anti-p-ACC (Millipore, 1:500) anti-AKT (Cell Signaling, 1:600), anti-p-AKT (Cell Signaling, 1:500), anti-p38 (Cell Signaling, 1:500), anti-p-p38 (Cell Signaling, 1:400), anti-atrogin-1 (ECM Biosciences, 1:400), ATG16L1 (Sigma, 1:1,000) and anti-actin (Santa Cruz Biotechnology, 1:3,000) antibodies.

For *in vitro* assessment of AMPK activation, MEFS derived from *Bcl2*^{AAA} and *Bcl2*^{WT} mice, *Becn1*$^{+/-}$ and *Benc1*$^{+/+}$ mice, and *Atg16l1*^{HM} and *Atg16l1*^{WT} mice were treated with 2 mM of the AMP analogue AICAR, or DMSO vehicle for 2 h. Whole-cell lysates were prepared, and AMPK and phospho-AMPK levels were determined by western blot analysis. For quantification, the relative intensity values of western blot bands were normalized to that of the first lane (set as 1) in the wild-type non-exercised group.

Statistical analyses. ANOVA approaches were used to compare values among different experimental groups for data that met the normality assumption. One-way ANOVA was used for comparison between two groups. Two-way ANOVA was used for comparison of the magnitude of changes between two different groups in mice of two different genotypes. The normality assumption for the ANOVA model was checked using residual plots. When the assumption was violated, the data were log-transformed in order to meet the assumption. For data sets in which log transformation was inadequate to meet the analysis assumption, the non-parametric Wilcoxon rank-sum test was used.

Real-time PCR analyses. RNA isolation from muscle tissues and real-time RT–PCR were performed as previously described[28]. The following Sybr Green primers were used: β-actin: forward, CTGGCTCCTAGCACCATGAAGAT; reverse, GGTGGACAGTGAGGCCAGGAT; UCP1: forward, TCAGGATTGGCCTCTACGAC; reverse, TAAGCCGGCTGAGATCTTGT; UCP2: forward, TGCCCGTAATGCCATTGTC; reverse, AGTGGCAAGGGAGGTCATCT; cytochrome *c*: forward, GTCTGTTCGGGCGGAAGACAG; reverse, GGGGAGAGGATACCCTGATGG; and cytochrome *b*: forward, ATTCATTGACCTACCTGCCC; reverse, TCTGATGTGTAGTGTATGGC. The following Tagman primers were used: PPARGC1α (PGC1α): Mm00447183_m1; myoglobin: Mm00442968_m1; PPARGC1β (PGC1β): Mm01258518_m1; MYH7 (MHC Type I): Mm00600555_ml; MYH1 (MHC Type IId/x): Mm01332488_g1; MYH2 (MHC Type IIa): Mm01332564_m1; and GAPDH: Mm99999915_g1.

26. Su, T. *et al.* Deletion of histidine triad nucleotide-binding protein 1/PKC-interacting protein in mice enhances cell growth and carcinogenesis. *Proc. Natl Acad. Sci. USA* **100,** 7824–7829 (2003).
27. Furuya, N., Yu, J., Byfield, M., Pattingre, S. & Levine, B. The evolutionarily conserved domain of Beclin 1 is required for Vps34 binding, autophagy and tumor suppressor function. *Autophagy* **1,** 46–52 (2005).
28. Moresi, V. *et al.* Myogenin and class II HDACs control neurogenic muscle atrophy by inducing E3 ubiquitin ligases. *Cell* **143,** 35–45 (2010).
29. Sheehan, D. C. & Hrapchak, D. C. *Theory and Practice of Histotechnology* 2nd edn, 162–166 (Battelle, 1980).
30. Butchbach, M., Edwards, J. & Burghes, A. Abnormal motor phenotype in the SMNΔ7 mouse model of spinal muscular atrophy. *Neurobiol. Dis.* **27,** 207–219 (2007).
31. Liang, X. H. *et al.* Induction of autophagy and inhibition of tumorigenesis by beclin 1. *Nature* **402,** 672–676 (1999).
32. Berglund, E. D. *et al.* Fibroblast growth factor 21 controls glycemia via regulation of hepatic glucose flux and insulin sensitivity. *Endocrinology* **150,** 4084–4093 (2009).

ARTICLE

doi:10.1038/nature10774

Complete subunit architecture of the proteasome regulatory particle

Gabriel C. Lander[1]*, Eric Estrin[2]*, Mary E. Matyskiela[2]*, Charlene Bashore[2], Eva Nogales[1,3,4] & Andreas Martin[2,4]

The proteasome is the major ATP-dependent protease in eukaryotic cells, but limited structural information restricts a mechanistic understanding of its activities. The proteasome regulatory particle, consisting of the lid and base subcomplexes, recognizes and processes polyubiquitinated substrates. Here we used electron microscopy and a new heterologous expression system for the lid to delineate the complete subunit architecture of the regulatory particle from yeast. Our studies reveal the spatial arrangement of ubiquitin receptors, deubiquitinating enzymes and the protein unfolding machinery at subnanometre resolution, outlining the substrate's path to degradation. Unexpectedly, the ATPase subunits within the base unfoldase are arranged in a spiral staircase, providing insight into potential mechanisms for substrate translocation through the central pore. Large conformational rearrangements of the lid upon holoenzyme formation suggest allosteric regulation of deubiquitination. We provide a structural basis for the ability of the proteasome to degrade a diverse set of substrates and thus regulate vital cellular processes.

The ubiquitin–proteasome system is the major pathway for selective protein degradation in eukaryotic cells. Covalent modification with a polyubiquitin chain targets damaged, misfolded and short-lived regulatory proteins for ATP-dependent destruction by the 26S proteasome, a massive 1.5 MDa proteolytic machine. The proteasome thus controls a myriad of essential cellular processes, including the cell cycle, transcription and protein quality control[1]. Despite intensive study, however, the structural basis for substrate recognition and processing by the proteasome remains poorly understood.

The proteasome contains at least 32 different subunits that form a barrel-shaped 20S proteolytic core capped on either end by a 19S regulatory particle. The active sites of the peptidase are sequestered in an internal chamber, and access is controlled by the regulatory particle, which functions in substrate recognition, deubiquitination, unfolding and translocation of the unfolded chains into the core[2–5].

The regulatory particle is composed of 19 subunits and can be divided into two subcomplexes, the lid and the base. The lid consists of nine non-ATPase proteins (Rpn3, Rpn5– Rpn9, Rpn11, Rpn12 and Sem1 in yeast), including the deubiquitinating enzyme (DUB) Rpn11, whose activity is essential for efficient substrate degradation[6,7]. The base contains six distinct AAA+ ATPases, Rpt1–Rpt6, that form a hetero-hexameric ring (in the order Rpt1, Rpt2, Rpt6, Rpt3, Rpt4, Rpt5; ref. 8) and constitute the molecular motor of the proteasome. The ATPases are predicted to use the energy of ATP binding and hydrolysis to exert a pulling force on substrate proteins, unfold them, and translocate the polypeptides through a narrow central pore into the peptidase chamber. In the presence of ATP, the carboxy termini of the ATPases bind dedicated sites on the α-subunit ring (α1–α7) of the 20S core, triggering the opening of a gated access channel and facilitating substrate entry[5,9–11]. Besides Rpt1–Rpt6, the base contains four non-ATPase subunits: Rpn1, Rpn2 and the ubiquitin receptors Rpn10 and Rpn13. Additional ubiquitin shuttle receptors (Rad23, Ddi1 and Dsk2) are recruited to the base through interactions with Rpn1, which also binds a second, non-essential DUB, Ubp6 (refs 12–14).

Whereas the proteolytic core has been well studied, there is only limited structural characterization of the regulatory particle[11,15–17]. None of the 13 non-ATPase subunits, including the ubiquitin receptors and deubiquitinating enzymes, have been localized within this assembly. Although it has been shown that efficient degradation depends on the length, linkage type and placement of an ubiquitin chain, as well as the presence of an unstructured initiation site on a substrate[3,18,19], we are missing the topological information needed to explain these requirements. Thus, elucidating the architecture of the regulatory particle and the spatial arrangement of individual subunits is crucial to understanding the molecular mechanisms for substrate recognition and processing.

Here, we present the electron microscopy structure of the proteasome holoenzyme and the lid subcomplex. A new heterologous expression system for the lid facilitated the localization of all subunits within the regulatory particle, providing a complete architectural picture of the proteasome. The resulting structural understanding offers novel insight into the mechanisms of ubiquitin binding, deubiquitination, substrate unfolding and translocation by this major eukaryotic proteolytic machine.

Recombinant expression of yeast lid in *Escherichia coli*

We developed a system for the heterologous coexpression of all nine lid subunits from *Saccharomyces cerevisiae* in *Escherichia coli*. This system allowed us to generate truncations, deletions and fusion constructs that were used to localize individual subunits and delineate their boundaries within the lid. The recombinant, purified lid was analysed in its subunit composition and stoichiometry by SDS–polyacrylamide gel electrophoresis (SDS–PAGE; Supplementary Figs 1 and 2) and tandem mass spectrometry. The small, non-essential subunit Sem1 could not be detected, neither for the recombinant nor the endogenous lid that was isolated from yeast. All other subunits were present with the expected stoichiometry, and gel-filtration analyses showed indistinguishable elution profiles for the heterologously expressed lid and its endogenous counterpart (data not shown). Furthermore, atomic emission spectroscopy confirmed that the essential Zn^{2+} ion was incorporated in Rpn11, indicating proper folding in *E. coli*.

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, California 94720, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [3]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [4]QB3 Institute, University of California, Berkeley, California 94720, USA.
*These authors contributed equally to this work.

To compare the functionalities of recombinant and endogenous lid, we established conditions for their *in vitro* reconstitution with base and 20S core subcomplexes from yeast to yield 26S holoenzyme. These reassembled particles were assayed for their activity in ubiquitin-dependent substrate degradation by using a polyubiquitinated green fluorescent protein (GFP)–cyclin fusion protein and following the decrease in GFP fluorescence. Proteasome reconstituted with *E. coli*-expressed lid supported robust substrate degradation (Supplementary Fig. 3). Importantly, the three-dimensional electron microscopy reconstructions from negative-stained samples of both lid subcomplexes are practically identical (Fig. 1a and Supplementary Fig. 4), establishing this recombinant system as an ideal tool for our structural studies of the regulatory particle.

Localization of regulatory particle subunits

As a first step in elucidating the architecture of the regulatory particle, we compared the single-particle electron microscopy reconstructions of the yeast holoenzyme and the isolated lid subcomplex obtained at 9- and 15-Å resolution, respectively (Fig. 1b, Supplementary Figs 5–7 and Supplementary Movie 1). Docking the five-lobed, hand-shaped structure of the lid into the electron density of the holoenzyme revealed the lid's position on one side of the regulatory particle, forming extensive interactions with the base subcomplex, but also contacting the 20S core. The lid subunits Rpn3, Rpn5, Rpn6, Rpn7, Rpn9 and Rpn12 contain a C-terminal PCI (Proteasome–CSN–eIF3) domain that is assumed to have scaffolding functions and allow inter-subunit contacts[1]. Our reconstruction provided sufficient resolution to unambiguously locate the winged-helix fold and the flanking helical segments of individual PCIs (Fig. 1c and Supplementary Movie 1). The C-terminal PCI domains of the six Rpn subunits thus interact laterally to form a horseshoe-shaped anchor from which the amino-terminal domains extend radially. This arrangement demonstrates the scaffolding function of PCI domains in the lid, and we predict that similar interactions underlie the architecture of other PCI-containing complexes.

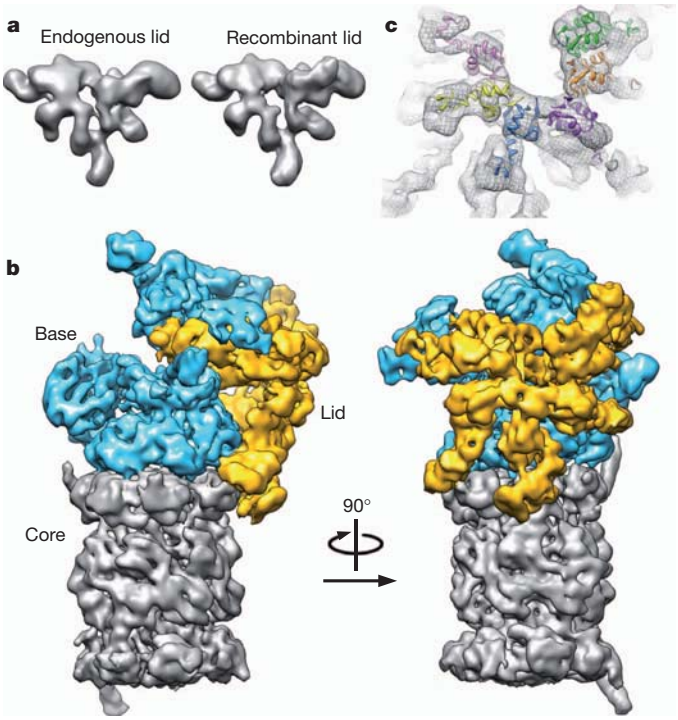


Figure 1 | The lid subcomplex within the holoenzyme assembly. a, Negative-stain three-dimensional reconstruction at approximately 15-Å resolution shows resemblance between endogenous (left) and recombinant (right) lid. **b**, Locations of lid (yellow) and base (cyan) within the subnanometre holoenzyme reconstruction. **c**, Six copies of the crystal structure of a PCI domain (PDB ID: 1RZ4) are docked into the lid electron density, showing a horseshoe-shaped arrangement of the winged-helix domains. Each domain is coloured according to its respective lid subunit (Fig. 2).

To determine the subunit topology of the lid, we used our heterologous *E. coli* expression system, fused maltose-binding protein (MBP) to the N or C terminus of individual subunits (Supplementary Fig. 1), and localized the MBP within the tagged lid particles by negative-stain electron microscopy (Supplementary Fig. 8a). None of the MBP fusions notably affected the lid structure, and we were able to identify the positions of all eight essential lid subunits and the relative orientation of their N and C termini. In combination with the PCI docking, the resolution of secondary structures in the cryoelectron density and known molecular weights, this information allowed us to delineate approximate subunit boundaries (Fig. 2a and Supplementary Movie 1).

Overall, Rpn3, Rpn7, Rpn6, Rpn5 and Rpn9 form the fingers of the hand-shaped lid structure. Rpn8 shows an extended conformation that connects Rpn3 and Rpn9, and thus closes the PCI horseshoe. In addition, it interacts with Rpn11, the only essential DUB of the proteasome, which lies in the palm of the hand and makes extensive contacts with Rpn8, Rpn9 and Rpn5.

Using the topology determined for the isolated lid subcomplex, we delineated the individual lid subunits in the context of the holoenzyme (Fig. 2b). To complete the subunit assignment for the entire regulatory particle, the positions of Rpt1–Rpt6 in the base subcomplex were assigned according to established interactions with the core particle[15,20], whose crystal structure could be docked unambiguously into the electron microscopy density (Supplementary Fig. 9). We localized the two large non-ATPases Rpn1 and Rpn2 of the base subcomplex by antibody-labelling of a C-terminal Flag tag and N-terminal fusion of glutathione-*S*-transferase (GST), respectively (Supplementary Figs 2 and 10a–c). Rpn1 and Rpn2 had been predicted to contain numerous tetratricopeptide repeat (TPR)-like motifs and adopt α-solenoid structures[21]. Indeed, we found a high structural resemblance between Rpn1 and Rpn2, both consisting of a strongly curled solenoid that transitions into an extended arm towards the C terminus (Fig. 3a). Rpn1 contacts the C-terminal helix of the 20S core subunit α4 and, based on the variability observed in our electron microscopy images, is likely to be flexible or loosely attached to the side of the base. Previous crystallography studies of the archaeal proteasome homologue PAN revealed that the N-terminal domains of the ATPases form a separate hexameric ring (N-ring) that consists of OB domains and three protruding coiled-coil segments[17,22]. Each coiled coil is formed by the far N-terminal residues of two neighbouring ATPases in the hexamer. Although Rpt1 and Rpt2 do not seem to form an extended coiled coil, we find that the N-terminal helical portion of Rpt1 interacts with the solenoid and the C-terminal arm of Rpn1. Rpn2 is located above the N-ring and mounted atop the longest of the protruding coiled coils, formed by Rpt3 and Rpt6. These interactions strongly resemble those observed between Rpt1 and Rpn1 (Fig. 3a).

Localizing the ubiquitin receptors and DUBs within the regulatory particle is of particular interest. In addition to the DUB Rpn11 in the lid, we identified the positions of both intrinsic ubiquitin receptors, Rpn10 and Rpn13, and of the base-associated DUB Ubp6 by imaging proteasome particles from yeast deletion strains (Fig. 3b and Supplementary Fig. 10d–f). The ubiquitin receptor Rpn13 binds to Rpn2 as expected[23,24]. The globular VWA domain of the second receptor Rpn10 has been shown previously to stabilize the lid–base interaction[25,26]; however, we found that it does not contact the base directly. This domain bridges Rpn11 and Rpn9, which might increase the lid–base affinity indirectly by stabilizing Rpn11 in its Rpn2-bound conformation (see below). The flexibly attached ubiquitin interacting motif (UIM) of Rpn10 probably contacts the coiled coil formed by Rpt4 and Rpt5, stabilizing its position relative to other subunits and potentially communicating with the AAA+ motor. The DUB Ubp6 seems to be flexible and does not give rise to ordered density. Nonetheless, variance

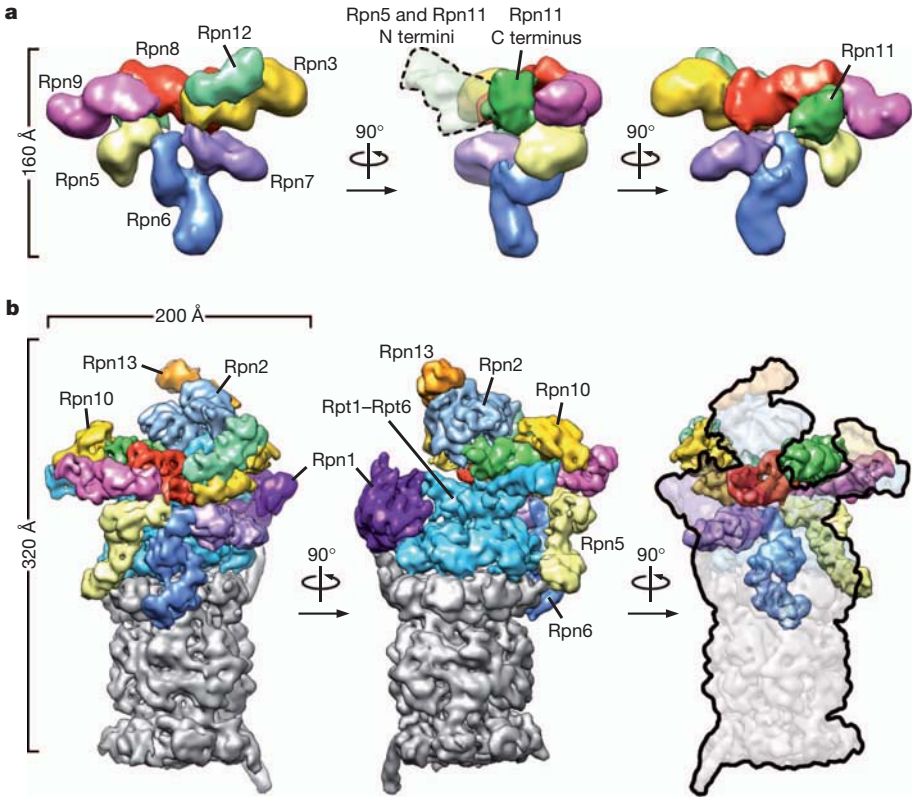


Figure 2 | Three-dimensional reconstructions of the recombinant lid subcomplex and the yeast 26S proteasome. **a**, Negative-stain reconstruction of the isolated lid subcomplex at 15-Å resolution, coloured by subunit and shown from the exterior (left), the side (middle) and the interior, base-facing side (right). A dotted line (middle) indicates the highly variable electron density for the flexible N-terminal domains of Rpn5 and Rpn11. **b**, Subnanometre cryoelectron microscopy reconstruction of the holoenzyme, shown in three views corresponding to the isolated lid and coloured as above, with the core particle in grey.

maps indicate that it interacts with the C-terminal arm of Rpn1, as suggested by immunoprecipitations[14].

Inter-subcomplex contacts

The complete localization of subunits within the holoenzyme revealed unexpected contacts between the lid and core subcomplexes. Rpn5 and Rpn6 form fingers that touch the C termini of the core subunits α1 and α2, respectively. We confirmed the interaction between Rpn6 and α2 by *in vitro* crosslinking, using an engineered cysteine in α2 and a 7-Å heterobifunctional crosslinker (Supplementary Fig. 11). These previously unknown direct interactions between lid and core may stabilize the entire holoenzyme assembly, and/or be part of an allosteric network that modulates the activities of either subcomplex.

Our holoenzyme structure shows that Rpn3, Rpn7, Rpn8 and Rpn11 make extensive contacts with the base. Compared to their positions in the isolated lid, Rpn8 and Rpn11 have undergone significant conformational changes in the holoenzyme (Fig. 4). The C terminus of Rpn8 is detached from Rpn3 to interact with the coiled coil of Rpt3/Rpt6, while the N-terminal MPN domain of Rpn11 extends towards the centre of the regulatory particle to bind the solenoid portion of Rpn2. Similarly, the N-terminal region of Rpn3 is more elongated than in the isolated lid and also contacts the Rpn2 solenoid, but from the opposite side. In turn, the extended C-terminal arm of Rpn2 interacts with Rpn3 and Rpn12, and thus forms a direct connection between the solenoid section of Rpn2, the coiled coil of Rpt3/Rpt6, and the lid (Fig. 3b).

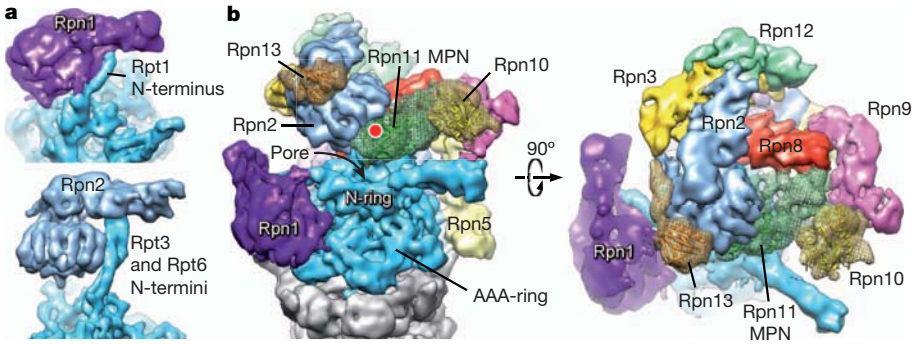


Figure 3 | Localization of Rpn1 and Rpn2, and ubiquitin-interacting subunits. **a**, Rpn1 (top) and Rpn2 (bottom) are oriented to emphasize similarities in their domain structure and solenoid attachment to the extended N-terminal helices of Rpt1 and Rpt3/Rpt6, respectively. **b**, Side and top views of the regulatory particle, showing the locations of the ubiquitin receptors Rpn10 and Rpn13, and the DUB Rpn11 relative to the central pore. Crystal structures for Rpn10 (PDB ID: 2X5N), Rpn13 (PDB ID: 2R2Y), and an MPN domain homologous to Rpn11 (AMSH-LP, PDB ID: 2ZNR) are shown docked into the electron microscopy density. The predicted active site of Rpn11 is indicated (red dot).

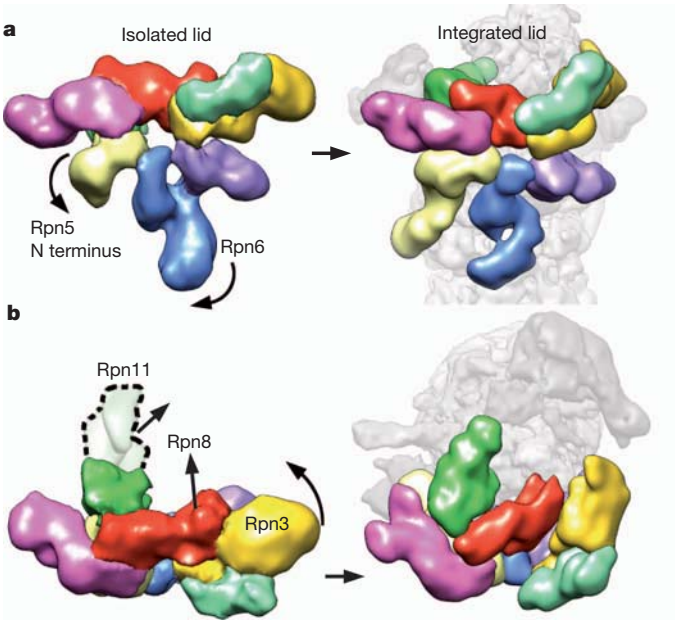


Figure 4 | Conformational rearrangements of the lid subcomplex upon integration into the holoenzyme. a, **b**, The lid complex in its isolated (left) and integrated (right) state is shown as viewed from the exterior (**a**) and top (**b**) of the regulatory particle. Major subunit rearrangements are depicted by arrows. The N terminus of Rpn5 (light yellow) interacts with Rpn11 in the isolated complex, and swings down to contact the core particle upon incorporation into the holoenzyme. The N-terminal domain of Rpn6 swings to the left to interact similarly with the core particle. Rpn3, Rpn8 and Rpn11 undergo notable rearrangements, in which they move towards the centre of the regulatory particle.

We speculate that Rpn2 stabilizes a lid conformation in which Rpn3, Rpn8 and the DUB Rpn11 extend towards the base (Fig. 4b). Together, the lid, Rpn2 and the coiled coils of the N-ring seem to function as a scaffold that positions the two intrinsic ubiquitin receptors Rpn10 and Rpn13, and the DUB Rpn11 for substrate binding, deubiquitination and transfer to the subjacent central pore of the AAA+ motor (Fig. 3b). Interestingly, several lid subunits interact directly with AAA+ domains of the Rpt subunits. Rpn7 contacts the AAA+ domains of Rpt2 and Rpt6, while Rpn6 and Rpn5 touch Rpt3. These interactions with contiguous motor domains are surprising, because current models for ATP-dependent unfoldases suggest significant conformational changes of individual subunits in the hexamer during ATP hydrolysis and substrate translocation[27–29]. The observed contacts between lid and the motor domains might form only transiently; alternatively, the AAA+ ring of the proteasome may be much more static than previously assumed.

Lid conformational changes may regulate DUB activity

Comparing the structures of the lid in isolation and when bound to holoenzyme revealed major conformational changes that suggest an allosteric mechanism for the regulation of Rpn11 DUB activity (Fig. 4). In the isolated lid, the N-terminal MPN domain of Rpn11 forms extensive interactions with Rpn9 and the curled up Rpn5 finger. Upon lid binding to the holoenzyme, this Rpn5 finger swings down to contact the α1 subunit of the 20S core and thereby releases Rpn11, which then extends towards the Rpn2 solenoid. Docking the MPN domain of a related DUB (PDB ID: 2ZNR) into the electron density of Rpn11 indicates the approximate location of the active site (Fig. 3b). The interactions of Rpn11 with Rpn9 and Rpn5 in the free lid probably restrict access to this active site, which would prevent futile substrate deubiquitination in the absence of base and 20S core, and explain previous observations that the lid subcomplex has DUB activity only within the holoenzyme[7] (and our unpublished data).

Functional asymmetry in the AAA+ unfoldase

Our subnanometre structure of the holoenzyme provides new insights into the architecture and potential mechanisms of the base AAA+ unfoldase. As suggested by previous electron microscopy studies[15,16], the ring of the base and the 20S core are slightly offset from a coaxial alignment, with the base shifted by approximately 10 Å towards the lid (Fig. 5a). Despite or perhaps because of this offset, the C-terminal tails of Rpt2, Rpt3 and Rpt5 are docked into their cognate 20S binding pockets at the interfaces of the subunits α3 and α4, α1 and α2, and α5 and α6, respectively. Those three Rpt tails contain the terminal HbYX motif, which is critical for triggering gate opening in the 20S core[5,10], and indeed our structure is consistent with an open-gate conformation. The tails of Rpt1, Rpt4 and Rpt6 lack this motif and were not observed to interact statically with 20S in our holoenzyme structure.

Current mechanistic models for AAA+ unfoldases predict that ATPase subunits in the hexamer are in different nucleotide states and undergo significant conformational changes driven by coordinated ATP hydrolysis[27,30,31]. Because we determined the structure of wild-type proteasome in the presence of saturating ATP, we expected that different complexes would have any given Rpt subunit in different conformations, leading to reduced electron density or low resolution when averaging thousands of these unsynchronized motors. However, our reconstruction shows highly ordered density throughout the AAA+ domains of all six Rpt subunits. Whereas the C-terminal 'small AAA+' subdomains (except for Rpt6) arrange in one plane above the 20S core, the 'large AAA+' subdomains of Rpt1–Rpt5 are oriented in a spiral staircase around the hexameric ring, with Rpt3 at the highest and Rpt2 at the lowest position (Fig. 5b and Supplementary Movie 1). The AAA+ domain of Rpt6 adopts a tilted orientation, bridging Rpt2 and Rpt3. Similar staircase arrangements have been observed previously for helicases of the AAA+ and RecA superfamilies[32,33]. It was suggested that during ATP hydrolysis, individual subunits progress through the different conformational stages of the staircase, thereby translocating substrate through the pore. The particular staircase orientation we observed identically for all proteasome particles may represent a low-energy state of the base, adopted under our experimental conditions. Alternatively, this staircase arrangement of Rpt1–Rpt6 may be static and reflect the functional state of the base, in which substrates are translocated by local motions of the pore loops while the relative positions of the motor subunits remain fixed. Future biochemical and structural studies will be required to distinguish between these two models.

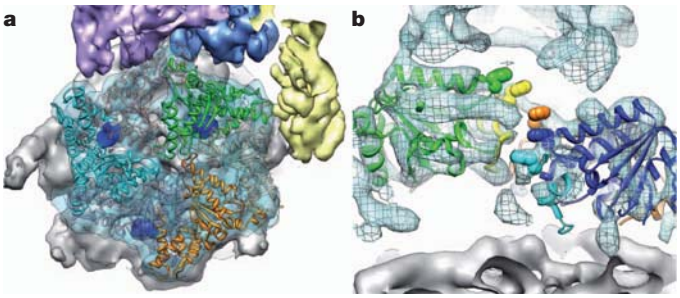


Figure 5 | Structural features of the base ATPase subunits. a, Positions of Rpt2 (cyan), Rpt3 (green) and Rpt5 (orange) within the base hexameric ring and relative to the 20S core (grey) are shown using fitted crystal structures of the homologous PAN AAA+ domain (PDB ID: 3H4M). The electron microscopy density contains the molecular envelope of the C-terminal tails (dark blue), docked into their cognate binding sites on the 20S core. Corresponding densities were not found for the tails of Rpt1, Rpt4 and Rpt6 (grey ribbon structure). **b**, Cutaway side view of the holoenzyme electron microscopy density with Rpt1–Rpt5 visible. Individually docked copies of the PAN crystal structure reveal a spiral staircase arrangement of the Rpt subunits, emphasized by space-filling representations of the PAN pore-1 loop residues (not resolved in the Rpt subunits).

Spatial arrangement of ubiquitin receptors and DUBs

Localizing all subunits of the regulatory particle enabled us to infer the requirements and potential mechanisms for the recognition and degradation of ubiquitin-tagged substrates (Fig. 6). After a substrate binds to an ubiquitin receptor, its polyubiquitin chain must be removed by Rpn11 cleavage at the proximal ubiquitin to permit subsequent fast degradation[6,7]. To allow cleavage without disengaging from the receptor, an ubiquitin chain must be long enough to span the distance between receptor and DUB. Based on our structure, both Rpn13 and the UIM of Rpn10 are located 70–80 Å from the predicted position of the Rpn11 MPN domain (Fig. 3b). The shuttle receptors Rad23, Ddi1 and Dsk2 are expected to reside ~80–120 Å away from Rpn11, depending on where they bind Rpn1 (ref. 13). For receptor interaction, at least part of the ubiquitin chain has to be in an extended conformation with the hydrophobic patches exposed[24,34,35]. Because a single ubiquitin moiety in an extended K48-linked chain contributes approximately 30 Å in length[36], it would take three ubiquitins to span the distance between Rpn10 or Rpn13 and Rpn11. Moreover, both Rpn10 and Rpn13 bind between two consecutive ubiquitin moieties[24,34], such that at least a tetra-ubiquitin chain would be required on a substrate to allow interaction with a receptor and simultaneous deubiquitination by Rpn11 (Fig. 6). This model agrees with *in-vitro* studies that indicate a minimum of four K48-linked ubiquitins is necessary for efficient substrate degradation[3], although this number may differ for other chain types[37]. Given the arrangement of Rpn10 and Rpn13, an ubiquitin chain would have to be significantly longer to interact with both receptors. However, knockout studies have shown that ubiquitin chains are not required to bind to multiple receptors simultaneously[38].

In contrast to Rpn11, Ubp6 is known to cleave within polyubiquitin chains or trim them from their distal end[39]. Of all the ubiquitin-interacting subunits in the regulatory particle, we found Ubp6 to be the furthest away from the entrance to the pore, which may allow it to clip extended or unnecessary ubiquitin chains from substrates. Because Ubp6 is located closer to Rad23, Dsk2 or Ddi1 than to Rpn10 or Rpn13, it may act preferentially on substrates delivered by these shuttle receptors.

To avoid dissociation upon deubiquitination, a substrate polypeptide must be engaged with the unfolding machinery of the base before or shortly after removal of its ubiquitin chain. Engagement by the base is known to depend on an unstructured initiation site or "tail" on the substrate[40], which needs to be long enough to reach through the narrow N-ring and into the AAA+ pore (Fig. 6). In addition, this tail would have to be sufficiently spaced from the attachment point of the polyubiquitin chain to allow concurrent substrate engagement by the pore and deubiquitination by Rpn11. The distance between the predicted active site of Rpn11 and the AAA+ pore below the N-ring is approximately 60 Å, which could easily be bridged by 40–45 unstructured residues or a shorter tail combined with a folded structure.

As an alternative to the above model for simultaneous receptor binding and deubiquitination, it has been proposed that commencing substrate translocation by the base might move the proximal ubiquitin from a receptor towards Rpn11 for cleavage[7]. Our structure suggests for this model that efficient substrate processing would only require a mono- or diubiquitin for receptor binding and a 50–60 Å longer spacing between the ubiquitin and the flexible tail to reach the AAA+ pore. This length dependence of engagement is consistent with recent *in vitro* degradation studies, using model substrates with different lengths and ubiquitin modifications[19]. Future experiments will be required to assess whether substrates get deubiquitinated in a translocation-dependent or -independent manner.

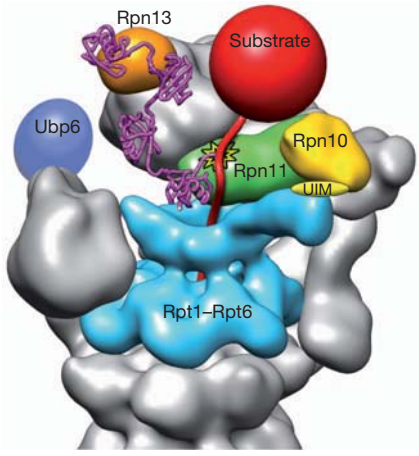


Figure 6 | Model for the recognition, deubiquitination and engagement of a polyubiquitinated substrate by the 26S proteasome. A K48-linked tetra-ubiquitin chain (magenta, PDB ID: 2KDE) is conjugated to the unstructured initiation region of a substrate (red) and bound to the ubiquitin receptor Rpn13 (orange). The substrate is poised for deubiquitination by Rpn11 (green, active site indicated by star), and its unstructured initiation region is engaged by the translocation machinery of the base (cyan). A polyubiquitin chain could alternatively bind to the UIM of Rpn10 (yellow) or interact with both receptors simultaneously. The DUB Ubp6 is localized further from the central pore, in a position to trim excess ubiquitin chains.

Concluding remarks

The work presented here defines the architecture of the entire proteasome regulatory particle and provides a much-needed structural framework for the mechanistic understanding of ubiquitin-dependent protein degradation. We localized Rpn11 directly above the entrance of the pore, surrounded by the ubiquitin receptors Rpn10 and Rpn13. This insight allows us to visualize the substrate's path towards degradation and will be critical in elucidating how the characteristics of ubiquitin modifications affect substrate recognition and processing. Moreover, our study significantly furthers the understanding of the heterohexameric AAA+ motor of the proteasome. Individual ATPase subunits were found in a spiral staircase arrangement and may operate with more limited dynamics than previously assumed for AAA+ protein unfoldases.

Unexpectedly, the lid is bound to the side of the holoenzyme and interacts with both the base and core particle. These interactions induce major conformational changes in lid subunits that may allosterically activate the DUB Rpn11, allowing critical removal of ubiquitin chains during substrate degradation in the holoenzyme, while preventing futile deubiquitination by the isolated lid. In addition, contacts between the subcomplexes could have unexplored roles in coordinating individual substrate processing steps, for instance ubiquitin binding, deubiquitination, and the onset of translocation. The intricate architecture of the proteasome highlights the complex requirements for this proteolytic machine, which must accommodate and specifically regulate a highly diverse set of substrates in the eukaryotic cell.

METHODS SUMMARY

Protein expression and purification. Endogenous holoenzyme, core particle[41] and lid subcomplex[42] were purified from *S. cerevisiae* essentially as described. The base subcomplex was purified according to protocols for the holoenzyme preparation, but with minor modifications as described in the Methods. Details of yeast strain construction are provided in Supplementary Table 1.

Yeast lid was recombinantly expressed from three plasmids in *E. coli* BL21-star (DE3), and purified on anti-Flag M2 resin and by size-exclusion chromatography (see Methods).

Electron microscopy and image analysis. All electron microscopy data were collected using the Leginon data collection software[43] and processed in the Appion electron microscopy processing environment[44]. Three-dimensional maps were calculated using libraries from the EMAN2 and SPARX software packages[45,46]. UCSF Chimera was used for volume segmentation, atomic coordinate docking and figure generation[47].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 October; accepted 12 December 2011.
Published online 11 January 2012.

1. Finley, D. Recognition and processing of ubiquitin-protein conjugates by the proteasome. *Annu. Rev. Biochem.* **78,** 477–513 (2009).
2. Glickman, M. H., Rubin, D. M., Fried, V. A. & Finley, D. The regulatory particle of the *Saccharomyces cerevisiae* proteasome. *Mol. Cell. Biol.* **18,** 3149–3162 (1998).
3. Thrower, J. S., Hoffman, L., Rechsteiner, M. & Pickart, C. M. Recognition of the polyubiquitin proteolytic signal. *EMBO J.* **19,** 94–102 (2000).
4. Groll, M. *et al.* A gated channel into the proteasome core particle. *Nature Struct. Biol.* **7,** 1062–1067 (2000).
5. Smith, D. M. *et al.* Docking of the proteasomal ATPases' carboxyl termini in the 20S proteasome's α ring opens the gate for substrate entry. *Mol. Cell* **27,** 731–744 (2007).
6. Yao, T. & Cohen, R. E. A cryptic protease couples deubiquitination and degradation by the proteasome. *Nature* **419,** 403–407 (2002).
7. Verma, R. *et al.* Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome. *Science* **298,** 611–615 (2002).
8. Tomko, R. J. Jr, Funakoshi, M., Schneider, K., Wang, J. & Hochstrasser, M. Heterohexameric ring arrangement of the eukaryotic proteasomal ATPases: implications for proteasome structure and assembly. *Mol. Cell* **38,** 393–403 (2010).
9. Rabl, J. *et al.* Mechanism of gate opening in the 20S proteasome by the proteasomal ATPases. *Mol. Cell* **30,** 360–368 (2008).
10. Gillette, T. G., Kumar, B., Thompson, D., Slaughter, C. A. & DeMartino, G. N. Differential roles of the COOH termini of AAA subunits of PA700 (19 S regulator) in asymmetric assembly and activation of the 26 S proteasome. *J. Biol. Chem.* **283,** 31813–31822 (2008).
11. da Fonseca, P. C. & Morris, E. P. Structure of the human 26S proteasome: subunit radial displacements open the gate into the proteolytic core. *J. Biol. Chem.* **283,** 23305–23314 (2008).
12. Elsasser, S. *et al.* Proteasome subunit Rpn1 binds ubiquitin-like protein domains. *Nature Cell Biol.* **4,** 725–730 (2002).
13. Gomez, T. A., Kolawa, N., Gee, M., Sweredoski, M. J. & Deshaies, R. J. Identification of a functional docking site in the Rpn1 LRR domain for the UBA-UBL domain protein Ddi1. *BMC Biol.* **9,** 33 (2011).
14. Leggett, D. S. *et al.* Multiple associated proteins regulate proteasome structure and function. *Mol. Cell* **10,** 495–507 (2002).
15. Bohn, S. *et al.* Structure of the 26S proteasome from *Schizosaccharomyces pombe* at subnanometer resolution. *Proc. Natl Acad. Sci. USA* **107,** 20992–20997 (2010).
16. Nickell, S. *et al.* Insights into the molecular architecture of the 26S proteasome. *Proc. Natl Acad. Sci. USA* **106,** 11943–11947 (2009).
17. Förster, F. *et al.* An atomic model AAA-ATPase/20S core particle sub-complex of the 26S proteasome. *Biochem. Biophys. Res. Commun.* **388,** 228–233 (2009).
18. Xu, P. *et al.* Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell* **137,** 133–145 (2009).
19. Inobe, T., Fishbain, S., Prakash, S. & Matouschek, A. Defining the geometry of the two-component proteasome degron. *Nature Chem. Biol.* **7,** 161–167 (2011).
20. Tian, G. *et al.* An asymmetric interface between the regulatory and core particles of the proteasome. *Nature Struct. Mol. Biol.* **18,** 1259–1267 (2011).
21. Effantin, G., Rosenzweig, R., Glickman, M. H. & Steven, A. C. Electron microscopic evidence in support of α-solenoid models of proteasomal subunits Rpn1 and Rpn2. *J. Mol. Biol.* **386,** 1204–1211 (2009).
22. Zhang, F. *et al.* Structural insights into the regulatory particle of the proteasome from *Methanocaldococcus jannaschii*. *Mol. Cell* **34,** 473–484 (2009).
23. Hamazaki, J. *et al.* A novel proteasome interacting protein recruits the deubiquitinating enzyme UCH37 to 26S proteasomes. *EMBO J.* **25,** 4524–4536 (2006).
24. Schreiner, P. *et al.* Ubiquitin docking at the proteasome through a novel pleckstrin-homology domain interaction. *Nature* **453,** 548–552 (2008).
25. Glickman, M. H. *et al.* A subcomplex of the proteasome regulatory particle required for ubiquitin-conjugate degradation and related to the COP9-signalosome and eIF3. *Cell* **94,** 615–623 (1998).
26. Verma, R., Oania, R., Graumann, J. & Deshaies, R. J. Multiubiquitin chain receptors define a layer of substrate selectivity in the ubiquitin-proteasome system. *Cell* **118,** 99–110 (2004).
27. Glynn, S. E., Martin, A., Nager, A. R., Baker, T. A. & Sauer, R. T. Structures of asymmetric ClpX hexamers reveal nucleotide-dependent motions in a AAA+ protein-unfolding machine. *Cell* **139,** 744–756 (2009).
28. Maillard, R. A. *et al.* ClpX(P) generates mechanical force to unfold and translocate its protein substrates. *Cell* **145,** 459–469 (2011).
29. Aubin-Tam, M. E., Olivares, A. O., Sauer, R. T., Baker, T. A. & Lang, M. J. Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine. *Cell* **145,** 257–267 (2011).
30. Martin, A., Baker, T. A. & Sauer, R. T. Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. *Nature* **437,** 1115–1120 (2005).
31. Hersch, G. L., Burton, R. E., Bolon, D. N., Baker, T. A. & Sauer, R. T. Asymmetric interactions of ATP with the AAA+ ClpX6 unfoldase: allosteric control of a protein machine. *Cell* **121,** 1017–1027 (2005).
32. Thomsen, N. D. & Berger, J. M. Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell* **139,** 523–534 (2009).
33. Enemark, E. J. & Joshua-Tor, L. Mechanism of DNA translocation in a replicative hexameric helicase. *Nature* **442,** 270–275 (2006).
34. Riedinger, C. *et al.* Structure of Rpn10 and its interactions with polyubiquitin chains and the proteasome subunit Rpn12. *J. Biol. Chem.* **285,** 33992–34003 (2010).
35. Eddins, M. J., Varadan, R., Fushman, D., Pickart, C. M. & Wolberger, C. Crystal structure and solution NMR studies of Lys48-linked tetraubiquitin at neutral pH. *J. Mol. Biol.* **367,** 204–211 (2007).
36. Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z. & Pickart, C. M. Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J. Biol. Chem.* **267,** 16467–16471 (1992).
37. Bremm, A., Freund, S. M. & Komander, D. Lys11-linked ubiquitin chains adopt compact conformations and are preferentially hydrolyzed by the deubiquitinase Cezanne. *Nature Struct. Mol. Biol.* **17,** 939–947 (2010).
38. Husnjak, K. *et al.* Proteasome subunit Rpn13 is a novel ubiquitin receptor. *Nature* **453,** 481–488 (2008).
39. Hanna, J. *et al.* Deubiquitinating enzyme Ubp6 functions noncatalytically to delay proteasomal degradation. *Cell* **127,** 99–111 (2006).
40. Prakash, S., Tian, L., Ratliff, K. S., Lehotzky, R. E. & Matouschek, A. An unstructured initiation site is required for efficient proteasome-mediated degradation. *Nature Struct. Mol. Biol.* **11,** 830–837 (2004).
41. Verma, R. *et al.* Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* **11,** 3425–3439 (2000).
42. Leggett, D. S., Glickman, M. H. & Finley, D. Purification of proteasomes, proteasome subcomplexes, and proteasome-associated proteins from budding yeast. *Methods Mol. Biol.* **301,** 57–70 (2005).
43. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
44. Lander, G. C. *et al.* Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166,** 95–102 (2009).
45. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
46. Hohn, M. *et al.* SPARX, a new environment for Cryo-EM image processing. *J. Struct. Biol.* **157,** 47–55 (2007).
47. Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157,** 281–287 (2007).

Acknowledgements We thank the members of the Martin and Nogales labs for helpful discussions, and G. Cardone for help with local resolution calculations. G.C.L. acknowledges support from Damon Runyon Cancer Research Foundation, M.E.M. acknowledges support by the American Cancer Society grant 121453-PF-11-178-01-TBE, C.B. acknowledges support from the NSF Graduate Research Fellowship. This research was funded in part by the Searle Scholars Program (A.M.), start-up funds from the UC Berkeley MCB Department (A.M.), the NIH grant R01-GM094497-01A1 (A.M.), the Lawrence Berkeley National Laboratory (G.C.L.), and the Howard Hughes Medical Institute (E.N.). Some of the work presented here was conducted at the National Resource for Automated Molecular Microscopy, which is supported by the NIH through the NCRR P41 program (RR017573).

Author contributions E.E., M.E.M. and C.B. designed, expressed and purified proteasome constructs, and performed biochemical experiments. G.C.L. performed the electron microscopy, processing and segmentation analysis. All authors contributed to experimental design, data analysis and manuscript preparation.

Author information The cryoelectron microscopy density map for the 26S proteasome can be found at the Electron Microscopy Data Bank under accession number EMD-1992. The negative stain reconstructions of the recombinantly expressed and yeast-purified lid have been assigned accession numbers EMD-1993 and EMD-1994, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M. (a.martin@berkeley.edu).

METHODS

Recombinant lid construction and purification. Yeast Rpn5, Rpn6, Rpn8, Rpn9 and Rpn11–6×His were cloned into pETDuet-1 (Novagen), yeast Rpn3, Flag–Rpn7 and Rpn12 were cloned into pCOLADuet-1 (Novagen), and yeast Sem1 and Hsp90 were cloned into pACYCDuet-1 (Novagen). A T7 promoter preceded each gene and each plasmid contained a T7 terminator following the multiple cloning site. Genes for select rare transfer RNAs were included in the pACYCDuet-1 plasmid to account for codon-usage differences between yeast and *E. coli*. To ensure full-length of Rpn6 in lid particles used for biochemical experiments and the negative stain reconstruction of recombinant lid, we used a construct with the Flag tag moved from Rpn7 to Rpn6. *E. coli* BL21-star (DE3) cells were co-transformed with the three plasmids mentioned above. Lid proteins and the chaperone Hsp90 were coexpressed overnight at 18 °C after inducing cells with 0.5 mM isopropyl-β-D-thiogalactopyranoside at $D_{600} = 0.7$. Cells were collected by centrifugation (4,000*g* for 30 min), resuspended in Flag buffer (50 mM HEPES, pH 7.6, 100 mM NaCl, 100 mM KCl and 5% glycerol) supplemented with protease inhibitors and 2 mg ml^{-1} lysozyme, and sonicated on ice for 2 min in 15-s bursts. The lysate was clarified by centrifugation (27,000*g* for 30 min), and the complex was affinity-purified on anti-Flag M2 resin (Sigma-Aldrich) using an N-terminal Flag-tag on Rpn6 or Rpn7. The protein was concentrated in a 30,000 MWCO concentrator (Amicon) for further purification on a Superose 6 size-exclusion column (GE Healthcare) equilibrated in Flag buffer. Intact, assembled lid particles eluted at 13.1 ml, similar to lid purified from yeast.

His_6-tagged yeast Rpn10 was expressed in *E. coli* and purified by Ni-NTA affinity and size-exclusion chromatography.

Yeast strain construction. Wild-type holoenzyme was purified from the strain YYS40 (*MAT***a** *ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1 RPN11::RPN11-3XFLAG (HIS3)*)[48]. To generate *RPN10*, *RPN13* and *UBP6* deletion strains, the kanMX6 sequence was integrated at the respective genomic locus, replacing the gene in YYS40. To generate the strains used to purify GST–Rpn2, GFP–Rpn5 and GFP–Rpn8 holoenzyme, sequences encoding the respective tags under the control of the P_{GAL1} promoter were integrated 5′ of the respective genes in YYS40. To generate the strain used to purify Rpn1–Flag holoenzyme, a sequence encoding the Flag-tag was integrated 3′ to *RPN1* in aW303 background strain (*MAT***a** *ade2-1 his3-11 leu2-3,112 trp1-1 ura3-1 can1-100 bar1*).

To generate the strains used to purify α2 mutant-containing core particle for the crosslinking experiments shown in Supplementary Fig. 11, pRS305 (*LEU2*) containing the mutant α2 and the genomic sequences found 500 nucleotides upstream and 100 nucleotides downstream of the gene was integrated at the *LEU2* locus of RJD1144 (*MAT***a**, *his3Δ200 leu2-3,112 lys2-801 trpΔ63 ura3-52 PRE1-FLAG-6xHIS::Ylplac211* (*URA3*))[41], and the chromosomal copy of α2 was deleted. To generate the strain used to purify lid with Rpn6 tagged with three haemagglutinin (HA) for crosslinking, the 3×HA sequence was integrated 3′ of *RPN6* in YYS40.

Expression and purification of yeast holoenzyme and subcomplexes. Endogenous holoenzyme, core particle[41] and lid subcomplex[42] were purified from *S. cerevisiae* essentially as described. Frozen yeast cells were lysed in a Spex SamplePrep 6870 Freezer/Mill. For holoenzyme purification, lysed cells of a strain containing a Flag-tag on Rpn11 were resuspended in lysis buffer containing 60 mM HEPES pH 7.6, 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 10% glycerol, 0.2% NP-40, and ATP regeneration mix (5 mM ATP, 0.03 mg ml^{-1} creatine kinase, 16 mM creatine phosphate). Holoenzyme was bound to anti-Flag M2 resin and washed with wash buffer (60 mM HEPES pH 7.6, 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 10% glycerol, 0.1% NP-40 and 500 μM ATP) before elution with 3×Flag peptide and separation over Superose-6 in gel-filtration buffer (60 mM HEPES pH 7.6, 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 10% glycerol and 500 μM ATP). Lid, base or core particle were purified similarly but from different yeast strains and including a salt wash to separate subcomplexes. Lid was purified from a yeast strain containing Rpn11–Flag using a 900 mM NaCl wash. Base was purified from a yeast strain containing a C-terminal Flag tag on Rpn2 and including a 500 mM NaCl wash, with 500 μM ATP present throughout the purification. Core particle was purified from a yeast strain containing a Flag–6×His tag on Pre1 and including a 500 mM NaCl wash. All subcomplexes were further purified by size-exclusion chromatography on Superose-6 in gel filtration buffer (see above).

GFP degradation assay. Proteasome holoenzyme was reconstituted from 20S core, base, Rpn10 and recombinant or endogenous yeast lid in the presence of ATP. A GFP–titin–cyclin fusion protein was modified with a K48-linked polyubiquitin chain[49] and degraded by reconstituted proteasome at 30 °C in Flag buffer with an ATP-regeneration system (5 mM ATP, 16 mM creatine phosphate, 6 μg ml^{-1} creatine phosphokinase). Degradation was monitored by the loss of fluorescence using a QuantaMaster spectrofluorimeter (PTI).

Protein crosslinking. Sulfo-MBS (Thermo Scientific) is a short (7.3 Å), heterobifunctional crosslinker, whose maleimide moiety reacts primarily with sulphydryls between pH 6.5 and 7.5, and whose NHS ester reacts with primary amines between pH 7 and 9. We purified core particle from yeast strains in which the only copy of the core α2 subunit was either wild type, a D245C mutant, or an A249C mutant. Other intrinsic cysteines of the core were found largely non-reactive towards sulphydryl-modifying agents (not shown). 10 μM reduced core particle purified from strains containing wild type, A249C and D245C α2 was incubated with 150 μM sulpho-MBS for 15 min at pH 6.5, allowing conjugation of the crosslinker to cysteines. Core particle was buffer-exchanged to remove excess crosslinker and increase the pH to 7.5, activating the amine-reactive functional group on sulpho-MBS. This core particle was added at a final concentration of 2 μM to a proteasome reconstitution mixture, containing 2 μM purified base, 10 μM purified Rpn10, 0.5 mM ATP, and 2 μM lid purified from a yeast strain in which Rpn6 was C-terminally tagged with a 3×HA tag. Crosslinking was allowed to proceed for 15 min before reactions were stopped by the addition of 0.5 mM glycine pH 7.5 and divided equally for separation by SDS–PAGE, followed by either Coomassie staining or anti-HA western blotting.

Electron microscopy. Sample preparation: negative-stain analysis of both the purified proteasome lid and holoenzyme complexes was performed using 400 mesh continuous carbon grids that had been plasma-cleaned in a 75% argon/25% oxygen atmosphere for 20 s using a Solarus plasma cleaner (Gatan). Due to the tendency for holoenzyme to adopt a preferential orientation on the carbon substrate, 5 μl of 0.1% poly L-lysine hydrobromide (Polysciences catalogue no. 09730) was placed onto the hydrophilized carbon grids and adsorbed for 90 s, washed twice with 5 μl drops of water, and allowed to dry completely. This polylysine step was skipped when preparing grids containing the lid samples, as the lid does not adopt a preferred orientation on the carbon substrate. The remaining steps were identical for both holoenzyme and lid. A 4-μl drop of sample at a concentration of 25 μM was placed onto the grid and allowed to adsorb for 1 min. The grid was blotted to near-dryness and a 4-μl drop of fresh 2% (w/v) uranyl formate was quickly placed onto the grid. To reduce the amount of glycerol remaining on the grids, they were subsequently floated on four successive 25-μl drops of the uranyl formate solution, waiting 10 s on each drop. The grids were then blotted to dryness.

Preservation of both lid and holoenzyme complexes in vitreous ice was performed in the same manner. 400-mesh C-flats containing 2 μm holes with a spacing of 2 μm (Protochips) were plasma cleaned in a 75% argon/25% oxygen atmosphere for 8 s using a Solarus plasma cleaner (Gatan). The purified sample, at a concentration of 5 μM in a buffer containing 5% glycerol, was first diluted 1:5 from 60 mM HEPES, pH 7.6, 50 mM NaCl, 50 mM KCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 10% glycerol, 1 mM DTT, 0.5 mM ATP into a buffer containing 20 mM HEPES, pH 7.6, 50 mM NaCl, 50 mM KCl, 1 mM ATP, 1 mM DTT and 0.05% NP40, and 4-μl aliquots were placed onto the grids. Grids were immediately loaded into a Vitrobot (FEI company) whose climate chamber had equilibrated to 4 °C and 100% humidity. The grids were blotted for 3 s at an offset of −1 mm, and plunged into liquid ethane. The frozen grids were transferred to a grid box and stored in liquid nitrogen until retrieved for data collection.

Data collection: negative-stain analysis of the lid and holoenzyme samples was performed using a Tecnai T12 Bio-TWIN and a Tecnai F20 TWIN transmission electron microscope operating at 120 keV. Lid samples were imaged at a nominal magnification of ×68,000 (1.57 Å per pixel at the specimen level) on the T12, and ×80,000 (1.45 Å per pixel) on the F20. Holoenzyme samples were imaged at a magnification of ×49,000 (2.18 Å per pixel) on the T12, and ×50,000 (2.16 Å per pixel) on the F20. T12 data were acquired on a F416 CMOS 4Kx4K camera (TVIPS), F20 data were acquired on a Gatan 4Kx4K camera, and all micrographs were collected using an electron dose of 20 e^- Å^{-2} with a randomly set focus ranging from −0.5 to −1.2 μm. The automatic rastering application of the Leginon data collection software was used for data acquisition. Between 300 and 500 micrographs were collected for each of the negatively stained data sets.

For cryoelectron microscopy, individual grids were loaded into a 626 single-tilt cryotransfer system (Gatan) and inserted into a Tecnai F20 TWIN transmission electron microscope operating at 120 keV. Data were acquired at a nominal magnification of ×100,000 (1.08 Å per pixel) using an electron dose of 20 e^- Å^{-2} with a randomly set focus ranging from −1.2 to −2.5 μm. A total of 9,153 micrographs were collected of the holoenzyme using the MSI-T application of the Leginon software. While the holoenzyme was remained intact during the freezing process, the isolated lid specimen became completely disassembled during the freezing process. In an attempt to overcome this, the isolated lid was also frozen using grids onto which a thin carbon film was floated. Due to the elevated background noise from the addition of a carbon substrate, the resulting images lacked the sufficient signal-to-noise ratio necessary to solve a cryoelectron

microscopy structure of the isolated lid to a better resolution than the negative-stain structure.

Image processing of negative-stain data. All image pre-processing and two-dimensional classification was performed in the Appion image processing environment[44]. Due to the large number of data sets acquired for both the negatively stained lid and holoenzyme complexes, a generalized schema was used for image analysis. This schema also minimized user bias during comparison of tagged and deletion constructs with their wild-type counterparts. The contrast transfer function (CTF) of each micrograph was estimated concurrently with data collection using the ACE2 and CTFFind programs[50,51], providing a quantitative measurement of the imaging quality. Particle selection was also performed automatically concurrent with data collection. Negatively stained lid particles were selected from the micrographs using a difference of Gaussians (DoG) transform-based automated picker[52], and holoenzyme particles were selected using a template-based particle picker. Micrograph phases were corrected using ACE2, and both lid and holoenzyme particles were extracted using a 288 × 288-pixel box size. The data were then binned by a factor of two for processing. Each particle was normalized to remove pixels whose values were above or below 4.5 σ of the mean pixel value using the XMIPP normalization program[53].

To remove aggregation, contamination or other non-particle selections, particle stacks were decimated by a factor of 2 and subjected to five rounds of iterative multivariate statistical analysis (MSA) and multi-reference alignment (MRA) using the IMAGIC software package[54]. Two-dimensional class averages depicting properly assembled complexes were manually selected, and the non-decimated particles contributing to these class averages were extracted to create a new stack for further processing. To include a larger range of holoenzyme views, particles contributing to doubly capped proteasome averages were removed. This stack of particles went through five rounds of MSA/MRA in IMAGIC[54], and a final correspondence analysis and classification based on Eigen images using the SPIDER software package[55] was performed to generate two-dimensional class averages of the complexes.

Initial models for reconstructions of both the holoenzyme and lid were determined using the established "C1 startup" routines in IMAGIC. Two-dimensional class averages were manually inspected to select three images representing orthogonal views of the complex, which were in turn used to assign Eulers in a stepwise fashion to the entire data set of reference-free class averages. The resulting low-resolution models of the lid and holoenzyme were low-pass filtered to 60-Å resolution, and these densities were used as starting points for refinement of the three-dimensional structure.

Three-dimensional reconstructions were all performed using an iterative projection-matching and back-projection approach using libraries from the EMAN2 and SPARX software packages[45,46]. Refinement of the starting models began using an angular increment of 25°, progressing down to 2° for the lid, and 1° for the holoenzyme. The refinement only continued to the subsequent angular increment once greater than 95% of the particles showed a pixel error of less than 1 pixel. The resolution was estimated by splitting the particle stack into two equally sized data sets, calculating the Fourier shell correlation (FSC) between the resulting back-projected volumes. The estimated resolutions for the final endogenous and recombinant lid structures based on their FSC curves at 0.5 were about 15 Å.

Image processing of cryoelectron microscopy holoenzyme. Processing of the holoenzyme cryo data set proceeded in a very similar fashion to that of the negatively stained particle data sets. Only ACE2 was used to estimate CTF of the images and measure image quality, and particles were extracted using a box size of 576 pixels. Reference-free two-dimensional classification was performed to remove particles that did not contribute to averages depicting a doubly capped proteasome. Three rounds of reference-free two-dimensional classification, and particles were removed after each round. From an initial data set of 312,483 automatically selected particles, 93,679 were kept for the three-dimensional reconstruction. C2 symmetry was applied to one of the previously determined asymmetric negative-stained reconstructions to serve as a starting model for structure refinement. The reconstruction began using an angular increment of 25°, and iterated down to 0.6°. C2 symmetry was imposed during the reconstruction. Low-resolution Fourier amplitudes of the final map were dampened to match those of an experimental GroEL SAXS curve using the SPIDER software package[55].

The estimated resolution based on the FSC of the half-volumes at 0.5 was approximately 9 Å, although a local resolution calculation using the "blocres" function in the Bsoft package[56] indicated a range of resolutions within the density. The majority of the core particle subunits and the AAA+ ATPases were resolved to between 7- and 8-Å resolution, whereas the non-ATPase subunits in the regulatory particle ranged from 8- to 12-Å resolution (Supplementary Fig. 7). Notably, Rpn1 and the ubiquitin receptors Rpn10 and Rpn13 were the lowest resolution features of the holoenzyme. To filter the low-resolution portions of the map properly, without destroying the details of the better-ordered features, a resolution-driven adaptive localized low-pass filter was applied to the final volume (G. Cardone, personal communication).

The segmentation analysis was manually performed using the "Volume Tracer" tool in the UCSF Chimera visualization software[47]. This software was additionally used to perform all rigid-body fitting of crystal structures into the holoenzyme cryoelectron microscopy density, as well as to generate all renderings for figure images.

48. Sone, T., Saeki, Y., Toh-e, A. & Yokosawa, H. Sem1p is a novel subunit of the 26 S proteasome from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **279,** 28807–28816 (2004).
49. Kim, H. C. & Huibregtse, J. M. Polyubiquitination by HECT E3s and the determinants of chain type specificity. *Mol. Cell. Biol.* **29,** 3307–3318 (2009).
50. Mallick, S. P., Carragher, B., Potter, C. S. & Kriegman, D. J. ACE: automated CTF estimation. *Ultramicroscopy* **104,** 8–29 (2005).
51. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
52. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166,** 205–213 (2009).
53. Sorzano, C. O. *et al.* XMIPP: a new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.* **148,** 194–204 (2004).
54. van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116,** 17–24 (1996).
55. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116,** 190–199 (1996).
56. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157,** 3–18 (2007).

ARTICLE

doi:10.1038/nature10777

A PGC1-α-dependent myokine that drives brown-fat-like development of white fat and thermogenesis

Pontus Boström[1], Jun Wu[1], Mark P. Jedrychowski[2], Anisha Korde[1], Li Ye[1], James C. Lo[1], Kyle A. Rasbach[1], Elisabeth Almer Boström[3], Jang Hyun Choi[1], Jonathan Z. Long[1], Shingo Kajimura[4], Maria Cristina Zingaretti[5], Birgitte F. Vind[6], Hua Tu[7], Saverio Cinti[5], Kurt Højlund[6], Steven P. Gygi[2] & Bruce M. Spiegelman[1]

Exercise benefits a variety of organ systems in mammals, and some of the best-recognized effects of exercise on muscle are mediated by the transcriptional co-activator PPAR-γ co-activator-1 α (PGC1-α). Here we show in mouse that PGC1-α expression in muscle stimulates an increase in expression of FNDC5, a membrane protein that is cleaved and secreted as a newly identified hormone, irisin. Irisin acts on white adipose cells in culture and *in vivo* to stimulate UCP1 expression and a broad program of brown-fat-like development. Irisin is induced with exercise in mice and humans, and mildly increased irisin levels in the blood cause an increase in energy expenditure in mice with no changes in movement or food intake. This results in improvements in obesity and glucose homeostasis. Irisin could be therapeutic for human metabolic disease and other disorders that are improved with exercise.

PGC1-α is a transcriptional co-activator that mediates many biological programs related to energy metabolism. Originally described as a co-activator of PPAR-γ that modulated expression of uncoupling protein 1 (UCP1) and thermogenesis in brown fat[1], it has also been shown to control mitochondrial biogenesis and oxidative metabolism in many cell types. PGC1-α is induced in muscle by exercise and stimulates many of the best-known beneficial effects of exercise in muscle: mitochondrial biogenesis, angiogenesis and fibre-type switching[2]. It also provides resistance to muscular dystrophy and denervation-linked muscular atrophy[3]. The health benefits of elevated muscle expression of PGC1-α may go beyond the muscle tissue itself. Transgenic mice with mildly elevated muscle PGC1-α are resistant to age-related obesity and diabetes and have a prolonged lifespan[4]. This suggests that PGC1-α stimulates the secretion of factors from skeletal muscle that affect the function of other tissues. Here we show that PGC1-α stimulates the expression of several muscle gene products that are potentially secreted, including FNDC5. The *Fndc5* gene encodes a type I membrane protein that is processed proteolytically to form a newly identified hormone secreted into the blood, termed irisin. Irisin is induced in exercise and activates profound changes in the subcutaneous adipose tissue, stimulating browning and UCP1 expression. Importantly, this causes a significant increase in total body energy expenditure and resistance to obesity-linked insulin resistance. Thus, irisin action recapitulates some of the most important benefits of exercise and muscle activity.

Muscle PGC1-α transgenics

Mice with transgenically increased PGC1-α in muscle are resistant to age-related obesity and diabetes[4], suggesting that these animals have a fundamental alteration in systemic energy balance. We therefore analysed the adipose tissue of the PGC1-α transgenic mice for expression of genes related to a thermogenic gene program and genes characteristic of brown fat development. There were no significant alterations in the expression of brown-fat-selective genes in the interscapular brown adipose tissue or in the visceral (epididymal) white adipose tissue (Fig. 1a). However, the subcutaneous fat layer (inguinal), a white adipose tissue that is particularly prone to 'browning' (that is, formation of multilocular, UCP1-positive adipocytes), had significantly increased levels of *Ucp1* and *Cidea* messenger RNAs (Fig. 1b). We also observed increased UCP1 protein levels and more UCP1-positive stained multilocular cells in transgenic mice compared to controls (Fig. 1c, d). There are recent reports that exercise causes a mild increase in the expression of a thermogenic gene program in the visceral adipose tissue, a depot that has minimal expression of these genes[5]. As it is the subcutaneous white adipose depot that has the greatest tendency to turn on a thermogenic gene program and alter the systemic energy balance of mice[6], we re-investigated this with regard to browning of the white adipose tissues in two types of exercise. Similar to what has been reported[5], a twofold increase in *Ucp1* mRNA expression was observed in the visceral, epididymal fat with 3 weeks of wheel running (Supplementary Fig. 1). However, a much larger change (approximately 25 fold) was seen in the same mice in the subcutaeneous inguinal fat depot. Similarly, a small increase in *Ucp1* mRNA expression was seen in the epididymal fat with repeated bouts of swimming in warm (32 °C) water (Supplementary Fig. 1); however a very large increase (65 fold) was observed in the inguinal white depot (Supplementary Fig. 1). Thus, muscle-specific expression of PGC1α drives browning of subcutaneous white adipose tissue, possibly recapitulating part of the exercise program.

Media from PGC1-α-expressing myocytes

The effect on browning of the adipose tissues from PGC1-α-expressing muscle could be due to direct muscle–fat signalling or to another,

[1]Dana-Farber Cancer Institute and Harvard Medical School, 3 Blackfan Circle, CLS Building, Floor 11, Boston, Massachusetts 02115, USA. [2]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Renal Division, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]UCSF Diabetes Center and Department of Cell and Tissue Biology, University of California, San Francisco, California 94143, USA. [5]Department of Experimental and Clinical Medicine, Università Politecnica delle Marche, Electron Microscopy Unit-Azienda Ospedali Riuniti, Ancona 60020, Italy. [6]Diabetes Research Center, Department of Endocrinology, Odense University Hospital, DK-5000, Odense, Denmark. [7]LakePharma, Inc., 530 Harbor Blvd, Belmont, California 94002, USA.

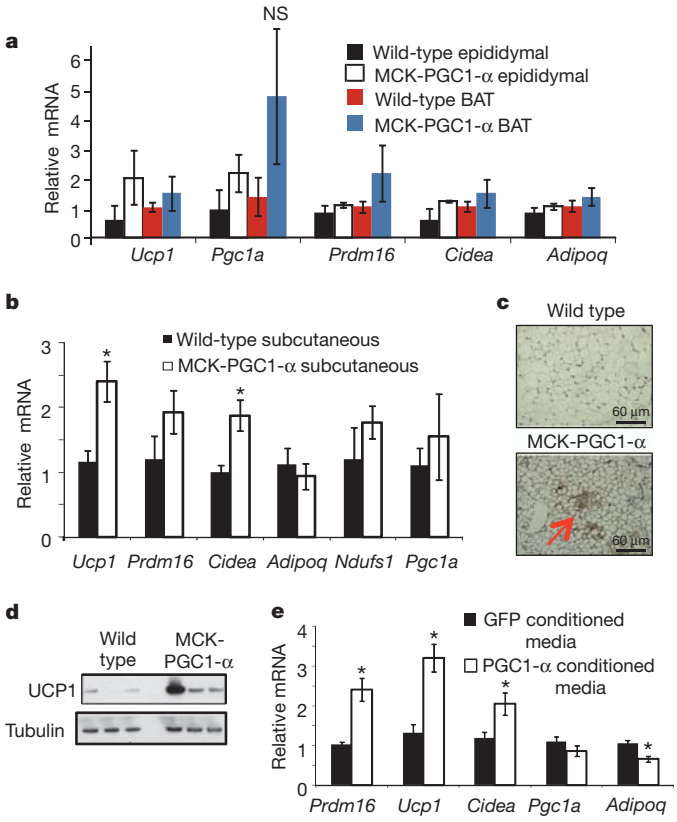


Figure 1 | Muscle-specific PGC1-α transgenic mice have increased brown/beige fat cells in the subcutaneous depot. **a**, **b**, Quantitative polymerase chain reaction (qPCR) against brown fat and thermogenic genes in epididymal fat, BAT (**a**) and subcutaneous, inguinal (**b**) fat depots in muscle creatine kinase promoter (MCK)-PGC1-α transgenic mice or littermate controls. $n = 7$ for each group, repeated in a separate cohort with similar results. **c**, Representative immunohistochemistry against UCP1 in the inguinal depot from indicated mice. **d**, Western blot against UCP1 in the inguinal fat depot ($n = 3$ and repeated in an independent cohort with similar results). **e**, qPCR against indicated genes in adipocytes differentiated for 6 days from stromo vascular fraction (SVF) cells. This was done in the presence of conditioned media from primary myocytes with forced expression of GFP or PGC1-α (representative for three independent experiments). Data are presented as mean ± s.e.m., and $^*P < 0.05$ compared to control group. Student's *t*-test was used for single comparisons. NS, not significant.

indirect mechanism. To investigate this, we treated cultured primary subcutaneous adipocytes with serum-free media conditioned by myocytes expressing PGC1-α or cells expressing green fluorescent protein (GFP). As shown in Fig. 1e, the media from cells expressing ectopic PGC1-α increased the mRNA levels of several brown-fat-specific genes (Fig. 1e). This suggested that PGC1-α causes the muscle cells to secrete a molecule(s) that can induce a thermogenic gene program in the cells.

Candidate, secreted PGC1-α-dependent proteins

We used a combination of Affymetrix-based gene expression arrays and an algorithm that predicts protein secretion to search for proteins that could mediate the browning of adipose tissues under the control of muscle PGC1-α (Methods). Proteins with mitochondrial targeting sequences were excluded, and all candidates were validated in gain-of-function systems for PGC1-α *in vivo* (Methods). Five proteins were identified as PGC1-α target genes in muscle and as likely to be secreted: IL-15, FNDC5, VEGF-β, LRG1 and TIMP4 (Fig. 2a). Conversely, expression of these genes was reduced in mice with muscle-specific deletion of PGC1-α (Supplementary Fig. 2). Furthermore, they were also found to be increased at the mRNA level in muscle from exercised mice (Fig. 2b). The expression of this same set of genes was also examined in muscle biopsies from human subjects before and after

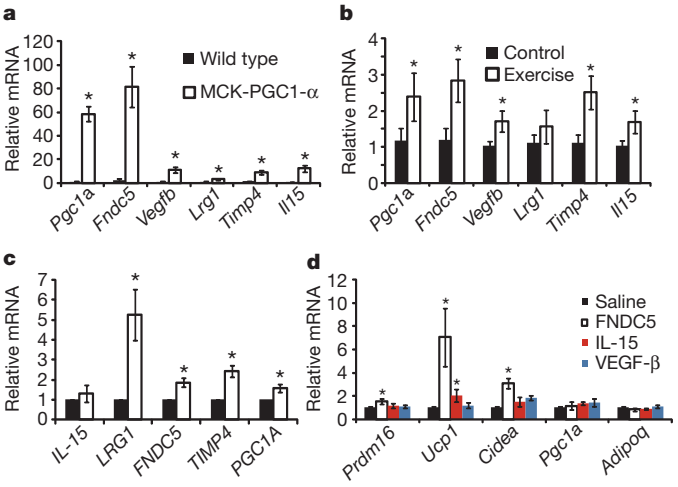


Figure 2 | FNDC5 is induced with forced PGC1-α expression or exercise, and turns on brown/beige fat gene expression. **a**, qPCR against indicated genes in skeletal muscle from MCK-PGC1-α transgenic mice or littermate controls ($n = 7$ from each group). **b**, qPCR against indicated genes in skeletal muscle from sedentary mice or mice given 3 weeks of free wheel running ($n = 10$ from each group). Mice were rested for 12 h before being killed. **c**, mRNA expression levels from human muscle biopsies before and after 10 weeks of endurance exercise training (8 subjects included). All data points are normalized to baseline levels. **d**, SVF from the inguinal fat depot, differentiated into adipocytes for 6 days in the presence of saline or recombinant FNDC5 (20 nM), IL-15 (10 μM) or VEGF-β (50 μM). The graph show normalized mRNA levels of indicated genes. This experiment was repeated three times with similar results. For **d**, we performed one-way ANOVA tests where $^*P < 0.05$ for the effect of FNDC5 on *Ucp1* and *Cidea* expression. All other statistics were performed using Student's *t*-test, and bar graphs are mean ± s.e.m.

a controlled period of endurance exercise[7] (Fig. 2c). *FNDC5*, *VEGFB* and *TIMP4* mRNAs were all significantly induced in humans with exercise. IL-15 has previously been reported as being secreted from muscle under the influence of exercise[8], while the regulation of FNDC5, VEGF-β, LRG1 and TIMP4 by exercise has not been described. *FNDC5* mRNA and brown fat markers in subcutaneous fat were not regulated by acute exercise, and *FNDC5* mRNA was not induced by exposure to cold (4 °C) for 6 h (Supplementary Fig. 2).

Fndc5 induces a browning *in vitro*

Several of the proteins encoded by these genes were commercially available, so they were applied directly to primary subcutaneous white adipocytes during differentiation. Factors such as IL-15 or VEGF-β had minimal effects on the expression of *Ucp1* and other brown fat genes at concentrations of 200 nM or higher. However, FNDC5 promoted a sevenfold induction of *Ucp1* mRNA at a concentration of 20 nM (Fig. 2d). The transcriptional changes induced by FNDC5 were addressed on a global scale using gene expression arrays (Supplementary Fig. 3). Notably, *Ucp1* and three other known brown fat genes, *Elovl3*, *Cox7a1* and *Otop1*, were among the eight most upregulated genes (Supplementary Fig. 4). Conversely, genes characteristic of white fat development were downregulated, such as leptin (Supplementary Fig. 3). These data indicate that the activation of browning and thermogenic genes by FNDC5 is a major part of the action of this polypeptide on these cells.

The effects of FNDC5 treatment were remarkably robust; *Ucp1* mRNA was increased 7–500 fold in more than ten experiments using FNDC5 at a concentration of 20 nM (Fig. 3a and Supplementary Fig. 5). Moreover, we could demonstrate a clear dose-dependence, with an effective range between 20–200 ng ml^{-1} (1.5–15 nM) (Fig. 3b). In contrast, BMP7, reported as a potent inducer of browning[9], had a much smaller effect (maximum of twofold) on the same cells at 3.3 μM (Fig. 3a).

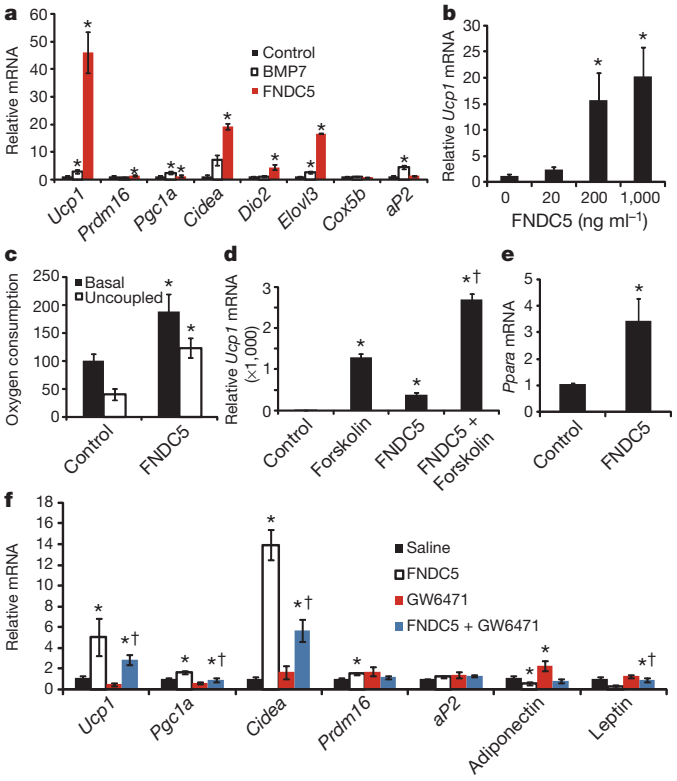


Figure 3 | FNDC5 is a potent inducer of the brown/beige fat gene program **a**, SVF from the inguinal fat depot was differentiated into adipocytes for 6 days in the presence of saline, recombinant FNDC5 (20 nM) or BMP7 (3.3 μM). The graph show normalized mRNA levels for indicated genes. Similar results were obtained in more than 10 experiments with the fold induction of *Ucp1* between 7–500 fold. **b**, mRNA levels of *Ucp1* from inguinal-derived SVF treated with FNDC5 for 6 days at indicated doses. **c**, Clark electrode measurements of oxygen consumption in SVF from the inguinal fat depot, differentiated into adipocytes for 6 days in the presence of saline or recombinant FNDC5 (20 nM). Data are representative for three independent experiments and normalized to total cellular protein. **d**, qPCR of *Ucp1* mRNA from SVF, differentiated into adipocytes, and treated with FNDC5 or saline for 6 days followed by addition of forskolin for 8 h. †$P < 0.05$ compared to forskolin treatment. **e**, qPCR of *Ppara* after 6 days of FNDC5 treatment (20 nM) during differentiation of primary SVF. **f**, SVF differentiated into adipocytes and treated with FNDC5 and/or GW6471 for 6 days. The graph shows qPCR of indicated genes. †$P < 0.05$ compared to FNDC5 treatment. For **d** and **f**, combined one- and two-way ANOVA was used. *$P < 0.05$. All other statistics were performed using Student's *t*-test, and bar graphs are mean ± s.e.m.

We also used immunohistochemistry to study cells treated with FNDC5 and observed a robust increase in UCP1-positive adipocytes with multilocular lipid droplets (Supplementary Fig. 4). Electron microscopic analysis of FNDC5-treated cells showed a higher density of mitochondria compared to control cells, consistent with a brown-fat-like phenotype and elevated mitochondrial gene expression (Supplementary Fig. 5). The sizes of mitochondria, however, were similar between groups (Supplementary Fig. 4). Lastly, measurements of oxygen consumption provided functional evidence of increased energy expenditure with FNDC5 exposure. Total oxygen consumption was greatly increased (100%) by 20 nM of FNDC5, and the majority of this respiration was uncoupled (Fig. 3c). Thus, FNDC5 potently induces thermogenesis and a brown-fat-like gene program in cultured adipocytes. In marked contrast, FNDC5 showed little or no effects on the classical brown fat cells isolated from the interscapular depot (Supplementary Fig. 4).

We sought to define the timeframe during the differentiation process when FNDC5 was effective. FNDC5 was applied to cells in 2-day windows from day 0–6, and this was compared to cells to which the protein was added during the entire 6-day differentiation process. As shown in Supplementary Fig. 5, treatment during days 3–6 was effective at inducing *Ucp1* mRNA, although not as effective as when FNDC5 was present throughout the differentiation process. Furthermore, treatment during the initial 2 days had no effect on *UCP1* levels, suggesting that FNDC5 acts mainly during the differentiation process of cells committed to the adipocyte lineage. Cyclic AMP (cAMP) is an important signalling pathway in thermogenesis, promoting the brown-fat gene program downstream of β-adrenergic stimulation. We therefore asked whether FNDC5 effects were additive or redundant with cAMP signalling. As shown in Fig. 3d, FNDC5-exposed cells increase UCP1 expression in an additive manner when exposed to forskolin, an adenyl cyclase activator. Two-way ANOVA tests demonstrated that there was a significant ($P < 0.01$) interaction between FNDC5 and forskolin treatment, indicating synergistic effects.

PPAR-α acts downstream of Fndc5

A key question is how FNDC5 is able to stimulate a thermogenic gene program. One potentially important transcription factor induced by FNDC5, identified using gene expression arrays, was PPAR-α. This nuclear receptor has been shown to drive *Ucp1* expression and several other genes involved in browning of adipose cells[10]. *Ppara* is increased threefold at the mRNA level by FNDC5 treatment (Fig. 3e). Importantly, the FNDC5-mediated increase in UCP1 was significantly reduced when cells were simultaneously subjected to the selective PPAR-α antagonist GW6471 (Fig. 3f). The functional interaction between the FNDC5 and GW6471 treatments on *UCP1* expression was confirmed using two-way ANOVA ($P < 0.05$). Conversely, the PPAR-α antagonist normalized the reduction seen in white adipose genes leptin and adiponectin after FNDC5 treatment. Together, these data indicate that FNDC5 acts to induce *Ucp1* gene expression, at least in part, via PPAR-α.

Irisin is a cleaved and secreted fragment of Fndc5

FNDC5 (also known as FRCP2 and PeP), was previously shown to have a signal peptide, two fibronectin domains and one hydrophobic domain that is likely to be membrane inserted[11,12] (Fig. 4a). Previous studies did not investigate whether part of this protein might be secreted[11,12]. With this structure in mind, we considered the possibility that FNDC5 might be synthesized as a type I membrane protein, followed by proteolytic cleavage and release of the amino (N)-terminal part of the protein into the extracellular space. Thus, any carboxy (C)-terminal or N-terminal tags would be lost during processing of the mature protein or interfere with localization. Indeed, expression of a C-terminally Flag-tagged FNDC5 (Fig. 4a) did not result in any Flag immunoreactivity in the medium from cells expressing this construct (Fig. 4b). However, when we immunoblotted the same samples with an antibody that recognizes the endogenous FNDC5 protein, we could easily detect substantial amounts of FNDC5 in the media at approximately 32 kilodalton (kDa): this is slightly larger than the cellular FNDC5 (Fig. 4b). These data indicate that FNDC5 is C-terminally cleaved, secreted and possibly further modified.

Western blot of media fractions with antibodies against wild-type FNDC5 showed multiple bands, suggestive of glycosylation, a common feature of secreted proteins. Treatment of supernatants from FNDC5-expressing cells with peptide *N*-glycosidase F (PNGase F) resulted in a significant size decrease as detected by SDS–polyacrylamide gel electrophoresis (SDS–PAGE), from 32 kDa to 20 kDa (Fig. 4c).

Mass spectrometry was used to determine the sequence of the FNDC5-derived polypeptide found in the media (Methods). To do this, we fused the N terminus of FNDC5 (without the signal peptide) to the C terminus of the crystallisable fragment (Fc) domain of immunoglobulin G (IgG). After purification and enzymatic deglycosylation of the secreted material, mass spectrometry analyses indicated that secreted FNDC5 was truncated at glutamic acid 112 (not including the signal sequence), as shown in Fig. 4e. The secreted portion of FNDC5 has remarkable conservation between species, with 100% identity between

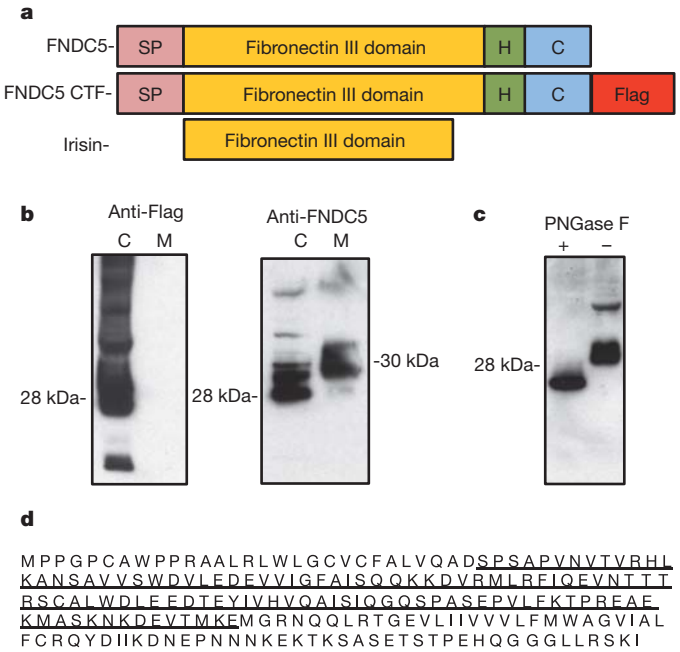


Figure 4 | FNDC5 is proteolytically cleaved and secreted from cells. **a**, Schematic representation of the FNDC5 protein structure (top), Flag-tagged FNDC5 protein (middle) and irisin (bottom). C, C-terminal domain; H, hydrophobic domain; SP, signal peptide. **b**, HEK 293 cells transfected with a vector expressing the C-terminal Flag-tagged FNDC5 (CTF-F5, third panel from **a**), followed by isolation of cell and culture media protein. Samples were adjusted for protein content and western blots were performed against the Flag antigen (left) or FNDC5 (right). This was repeated in several experiments with similar results. Adjusting for volume (instead of protein content) also gave similar results. C, cell fraction; M, media fraction. Arrows indicate molecular weight. **c**, HEK 293 cells transfected with a vector expressing FNDC5–CTF, followed by isolation of cell and media protein. Respective protein fraction was then treated with PNGase F followed by western blot against FNDC5 after SDS–PAGE. **d**, Representation of the full-length FNDC5 and the irisin fragment mapped by mass spectrometry (underlined).

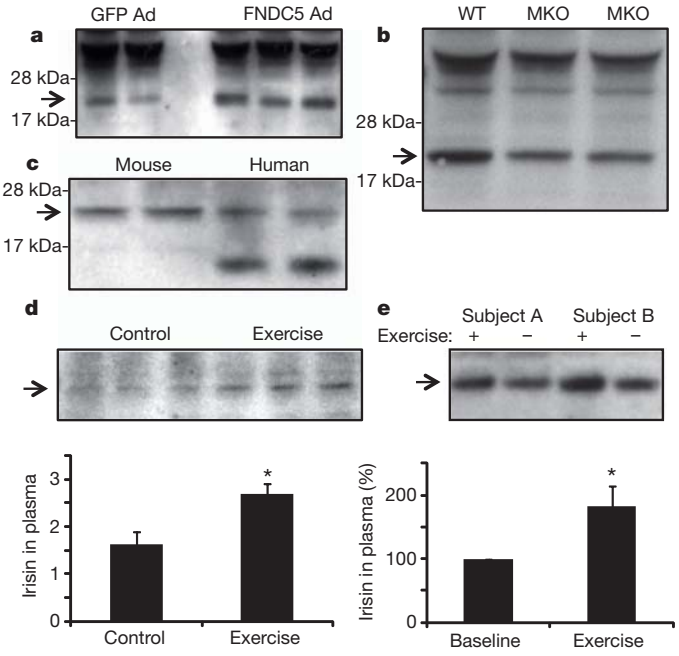


Figure 5 | Detection of irisin in mouse and human plasma. **a**, Plasma from mice injected intravenously with adenoviral vectors (Ad) expressing FNDC5 or GFP was subjected to western blot against FNDC5. **b**, Western blot against irisin in plasma from muscle-specific PGC1-α knockout (MKO) mice or control littermates (flox/flox). WT, wild type. **c**, Western blot against irisin in plasma from wild-type mice or two healthy human subjects (representative for 8 subjects analysed identically). **d**, Western blot against irisin in serum from control or 3 weeks exercised mice, followed by 12 h rest. Bottom panel shows quantification of the bands. **e**, Western blot analysis of irisin in plasma from human subjects before and after 10 weeks of endurance exercise. Eight subjects in total were analysed; quantification after internal normalization is displayed in bottom panel. For all plasma analyses, samples were depleted for albumin/IgG, and deglycosylated using PNGase F. Arrow indicated irisin band. Data are presented as mean ± s.e.m., and $*P < 0.05$ compared to control group. Student's *t*-test was used for single comparisons.

mice and humans (Supplementary Fig. 6). Because this distinct, secreted polypeptide has not been previously described and signals from muscle to other tissues we named it irisin, after Iris, the Greek messenger goddess.

The ability of the anti-FNDC5 antibodies to react with irisin allowed us to investigate the contribution made by irisin to the browning activity caused by muscle cells expressing PGC1-α. Media conditioned by muscle cells that had forced expression of PGC1-α were incubated with control or anti-FNDC5 antibodies before they were applied to the fat cell cultures. As shown in Supplementary Fig. 6, the FNDC5 antibody caused a marked reduction in the ability of the PGC1-α conditioned media to induce *Ucp1* and *Cidea* mRNA in the primary inguinal cells. This suggests that irisin accounts for a significant fraction of this activity found in secreted media from muscle cells with forced PGC1-α expression. We cannot, however, exclude the possibility that other factors might also contribute to this response.

Irisin is present in mouse and human plasma

We next analysed levels of irisin in plasma from wild-type mice, using intravenous adenoviral delivery of full-length FNDC5 as a positive control. This method results in strong forced expression from the liver and potential secretion to the plasma, where we detected irisin using western blot after albumin/IgG-depletion and deglycosylation. As seen in Fig. 5a, FNDC5-expressing virus resulted in a clear increase in an immunoreactive band at 22 kDa. Importantly, this was the only band altered on these blots. Using western blots of purified FNDC5 protein as a quantitative standard, irisin is present in the plasma of control mice at a concentration of approximately 40 nM. We also analysed plasma of PGC1-α muscle-specific knockout mice as a negative control, and the irisin band at 22 kDa was decreased by 72% in these animals (Fig. 5b). Furthermore, an immunoreactive band of identical electrophoretic mobility was found in plasma from healthy human subjects (Fig. 5c). This band was greatly diminished when the anti-FNDC5 antibody was neutralized with an excess of antigen (Supplementary Fig. 7).

We examined blood levels of irisin after exercise in mice and human subjects. Mice had significantly elevated (65%) plasma concentrations of irisin after they were subjected to 3 weeks of free wheel running (Fig. 5d). Similar analyses in healthy adult humans subjected to supervised endurance exercise training for 10 weeks revealed a twofold increase in the circulating irisin levels compared to the non-exercised state (Fig. 5e). Thus, irisin is present in mouse and human plasma, and is increased with exercise. The increase in circulating protein in both species is roughly proportional to the increases observed at the mRNA level in muscle (Fig. 2c).

Irisin reduces obesity and insulin resistance

We used adenoviral vectors to express full-length FNDC5 (or a control GFP) and examined its biological and therapeutic effects. This method resulted in a 15-fold increase in liver *Fndc5* mRNA, although the liver expresses very low endogenous levels of this mRNA. Plasma levels of irisin were increased 3–4 fold (Fig. 5a). The mice did not display any adverse reaction, and upon gross pathological examination, there was no apparent toxicity in any major organ system. There was also no increase in plasma AST levels, and inflammatory genes were not significantly altered in the liver when the two groups were compared

(Supplementary Fig. 8). Ten days after injection, *Ucp1* mRNA was increased by 13-fold in the subcutaneous depot relative to the same depot in mice receiving the virus expressing GFP (Fig. 6a, b); *Cidea* was also significantly upregulated (Fig. 6a). There were no changes in expression of UCP1 in the interscapular classical brown fat (BAT), but we did observe a minor elevation in *Prdm16* and *Pgc1a* mRNA (Supplementary Fig. 8). The changes in gene expression in the subcutaneous adipose tissues were accompanied by a clear increase in the number of UCP1-positive, multilocular adipocytes (Fig. 6c). We did not, however, detect any change in body weight in the GFP versus FNDC5 groups of animals. We observed similar results in young C57BL/6 mice (Supplementary Fig. 9). Thus, moderate increases in circulating irisin can induce browning of white adipose tissues *in vivo*, including increased expression of UCP1.

As activation of the classical brown fat or browning of white fat has been shown to improve obesity and glucose homeostasis *in vivo*[6,13], we delivered FNDC5-expressing adenovirus to mice rendered obese and insulin-resistance by feeding a high fat diet. We chose C57BL/6 mice for these experiments because they are highly prone to diet-induced obesity and diabetes. The expression of irisin increased *Ucp1* gene expression to the same degree as in lean mice (Supplementary Fig. 9).

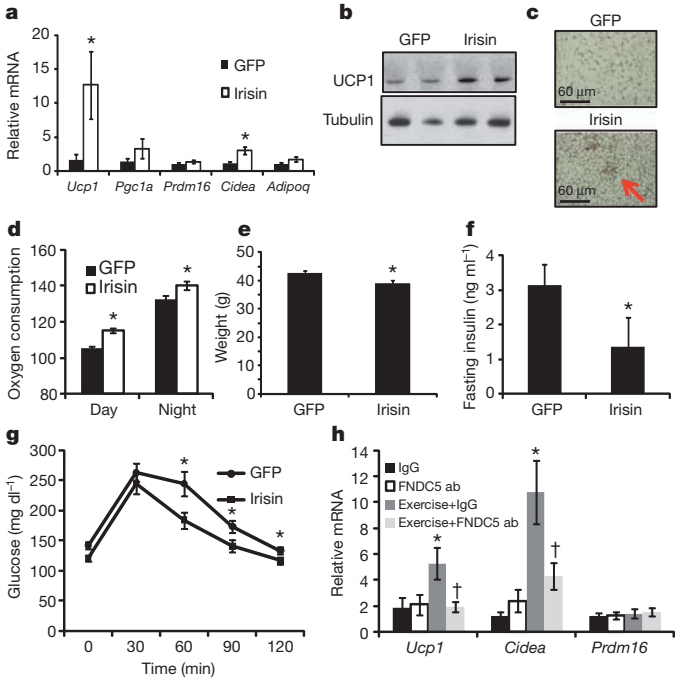


Figure 6 | Irisin induces browning of white adipose tissues *in vivo* and protects against diet-induced obesity and diabetes. **a**–**c**, Wild-type BALB/c mice were injected with 10^{10} GFP- or FNDC5-expressing adenoviral particles intravenously ($n = 7$ for each group). **a**, **b**, Animals were killed after 10 days and inguinal/subcutaneous fat pads were collected and analysed using qPCR analysis of indicated mRNAs (**a**) and western blot against UCP1 (**b**). **c**, Representative images from immunohistochemistry against UCP1 in these mice. All results in **a**–**c** were repeated two times with similar results. **d**–**g**, C57BL/6 mice fed a 60% kcal high-fat diet for 20 weeks were intravenously injected with GFP- or FNDC5-expressing adenovirus and all analyses were done 10 days thereafter ($n = 7$ for both groups). **d**, Oxygen consumption at day and night. **e**, Body weights of mice 10 days after injection with indicated adenovirus. **f**, Fasting plasma insulin measured using enzyme-linked immunosorbent assay (ELISA). **g**, Intraperitoneal glucose tolerance test. **h**, Mice were injected intraperitoneally with 50 μg of rabbit IgG or a rabbit anti-FNDC5 antibody (ab) and were either subjected to swimming for 7 days or kept sedentary ($n = 10$ for all groups). Data show mRNA expression levels from inguinal white adipose tissue. All data in **d**–**j** were performed at least twice in a separate mouse cohort with similar results. †$P < 0.05$ compared to exercise and IgG. One-way ANOVA was used for statistics in **h**. All other statistics were performed using Student's *t*-test, and bar graphs are mean ± s.e.m.

There was also an elevation in expression of several mitochondrial genes (Supplementary Fig. 8). Notably, these changes occurred with only moderately increased irisin blood levels, threefold compared to the GFP-expressing mice. This effect was accompanied with a large increase in oxygen consumption (Fig. 6d), consistent with the gene expression data, and body weights of the irisin-expressing mice were slightly reduced after 10 days compared to GFP-expressing controls (Fig. 6e). These effects of irisin on mitochondrial gene expression in the fat were not seen in skeletal muscle *in vivo* or in cultured myocytes (Supplementary Fig. 10). Irisin expression in the mice fed a high fat diet caused a significant improvement in glucose tolerance when compared to the control mice expressing GFP. In addition, fasting insulin was also reduced (Fig. 6f, g). These data illustrate that even moderately increased levels of circulating irisin potently increase energy expenditure, reduce body weight and improve diet-induced insulin resistance.

Finally, we asked whether irisin is required for the exercise-induced effects on the subcutaneous white fat. Injection of anti-FNDC5 antibodies into mice before 10 days of swim training dramatically reduced the effect of this exercise on *Ucp1* and *Cidea* gene expression, compared to injection of control antibodies (Fig. 6h). In contrast, *Prdm16* mRNA levels were not increased with exercise and were also not affected by the anti-FNDC5 antibodies. Thus, irisin is required for a substantial part of the effect of exercise on these gene expression events in the browning of white fat.

Discussion

Exercise has the capacity to improve metabolic status in obesity and type 2 diabetes, but the mechanisms are poorly understood. Importantly, exercise increases whole body energy expenditure beyond the calories used in the actual work performed[14]. However, the relative contribution of the adipose tissues to this phenomenon has not been clarified. Because transgenic mice expressing PGC1-α selectively in muscle showed a remarkable resistance to age-related obesity and diabetes[4], we sought factors secreted from muscle under the control of this co-activator that might increase whole body energy expenditure. We describe a new polypeptide hormone, irisin, which is regulated by PGC1-α, secreted from muscle into blood and activates thermogenic function in adipose tissues. Irisin is remarkable in several respects. First, it has powerful effects on the browning of certain white adipose tissues, both in culture and *in vivo*. Nanomolar levels of this protein increase UCP1 in cultures of primary white fat cells by 50 fold or more, resulting in increased respiration. Perhaps more remarkable, viral delivery of irisin that causes only a moderate increase (~3 fold) in circulating levels stimulates a 10–20 fold increase in UCP1, increased energy expenditure and an improvement in the glucose tolerance of mice fed a high fat diet. As this is in the range of increases seen with exercise in mouse and man, it is likely that irisin is responsible for at least some of the beneficial effects of exercise on the browning of adipose tissues and increases in energy expenditure. It is important to note that the evidence provided here does not exclude a role for other tissues besides muscle in the biological regulation and secretion of irisin.

Second, the cleaved and secreted portion of FNDC5, the hormone irisin, is highly conserved in all mammalian species sequenced. Mouse and human irisin are 100% identical, compared to 85% identity for insulin, 90% for glucagon and 83% identity for leptin. This certainly implies a highly conserved function that is likely to be mediated by a cell surface receptor. The identity of such a receptor is not yet known.

On the basis of the gene structure of *Fndc5*, with a signal peptide that was evidently missed in previous studies[12], we considered that FNDC5 might be a secreted protein. Indeed, we have observed that the signal peptide is removed, and the mature protein is further proteolytically cleaved and glycosylated, to release the 112-amino-acid polypeptide irisin. The cleavage and secretion of irisin is similar to the release/shedding of other transmembrane polypeptide hormones and hormone-like molecules such as epidermal growth factor (EGF) and transforming growth factor-α (TGF-α).

As the conservation of calories would probably provide an overall survival advantage for mammals, it seems paradoxical that exercise would stimulate the secretion of a polypeptide hormone that increases thermogenesis and energy expenditure. One explanation for increased irisin expression with exercise in mouse and man may be that it evolved as a consequence of muscle contraction during shivering. Muscle secretion of a hormone that activates adipose thermogenesis during this process might provide a broader, more robust defence against hypothermia.

The therapeutic potential of irisin is obvious. Exogenously administered irisin induces the browning of subcutaneous fat and thermogenesis, and it presumably could be prepared and delivered as an injectable polypeptide. Increased formation of brown or beige/brite fat has been shown to have anti-obesity, antidiabetic effects in multiple murine models[6], and adult humans have significant deposits of UCP1-positive brown fat[15]. Data presented here show that even relatively short treatments of obese mice with irisin improves glucose homeostasis and causes a small weight loss. Whether longer treatments with irisin and/or higher doses would cause more weight loss remains to be determined. The worldwide, explosive increase in obesity and diabetes renders attractive the therapeutic potential of irisin in these and related disorders.

Another potentially important aspect of this work relates to other beneficial effects of exercise, especially in some diseases for which no effective treatments exist. The clinical data linking exercise with health benefits in many other diseases suggests that irisin could also have significant effects in these disorders.

METHODS SUMMARY

Primary mouse stromal vascular fractions from adipose tissues were differentiated as described[6]. FNDC5/irisin were purchased from ABNOVA (GST fused), or produced from Syd Laboratories (GST fused) or LakePharma (Fc fusions). Hydrodynamic injections[16], electron microscopy[17] and treadmil running[18] were performed as previously described. Unless otherwise stated, bar graph data are presented as mean ± s.e.m., and $*P < 0.05$ compared to control group. Student's *t*-test was used for single comparisons and one-way ANOVA for multiple.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 August; accepted 13 December 2011.
Published online 11 January 2012.

1. Puigserver, P. *et al.* A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **92,** 829–839 (1998).
2. Handschin, C. & Spiegelman, B. M. The role of exercise and PGC1α in inflammation and chronic disease. *Nature* **454,** 463–469 (2008).
3. Sandri, M. *et al.* PGC-1α protects skeletal muscle from atrophy by suppressing FoxO3 action and atrophy-specific gene transcription. *Proc. Natl Acad. Sci. USA* **103,** 16260–16265 (2006).
4. Wenz, T., Rossi, S. G., Rotundo, R. L., Spiegelman, B. M. & Moraes, C. T. Increased muscle PGC-1α expression protects from sarcopenia and metabolic disease during aging. *Proc. Natl Acad. Sci. USA* **106,** 20405–20410 (2009).
5. Xu, X. *et al.* Exercise ameliorates high-fat diet-induced metabolic and vascular dysfunction, and increases adipocyte progenitor cell population in brown adipose tissue. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **300,** R1115–R1125 (2011).
6. Seale, P. *et al.* Prdm16 determines the thermogenic program of subcutaneous white adipose tissue in mice. *J. Clin. Invest.* **121,** 96–105 (2011).
7. Vind, B. F. *et al.* Impaired insulin-induced site-specific phosphorylation of TBC1 domain family, member 4 (TBC1D4) in skeletal muscle of type 2 diabetes patients is restored by endurance exercise-training. *Diabetologia* **54,** 157–167 (2011).
8. Nielsen, A. R. & Pedersen, B. K. The biological roles of exercise-induced cytokines: IL-6, IL-8, and IL-15. *Appl. Physiol. Nutr. Metab.* **32,** 833–839 (2007).
9. Tseng, Y. H. *et al.* New role of bone morphogenetic protein 7 in brown adipogenesis and energy expenditure. *Nature* **454,** 1000–1004 (2008).
10. Komatsu, M. *et al.* Multiple roles of PPARα in brown adipose tissue under constitutive and cold conditions. *Genes Cells* **15,** 91–100 (2010).
11. Teufel, A., Malik, N., Mukhopadhyay, M. & Westphal, H. *Frcp1* and *Frcp2*, two novel fibronectin type III repeat containing genes. *Gene* **297,** 79–83 (2002).
12. Ferrer-Martínez, A., Ruiz-Lozano, P. & Chien, K. R. Mouse PeP: a novel peroxisomal protein linked to myoblast differentiation and development. *Dev. Dyn.* **224,** 154–167 (2002).
13. Cederberg, A. *et al. FOXC2* is a winged helix gene that counteracts obesity, hypertriglyceridemia, and diet-induced insulin resistance. *Cell* **106,** 563–573 (2001).
14. Speakman, J. R. & Selman, C. Physical activity and resting metabolic rate. *Proc. Nutr. Soc.* **62,** 621–634 (2003).
15. Enerbäck, S. Human brown adipose tissue. *Cell Metab.* **11,** 248–252 (2010).
16. Bell, J. B., Aronovich, E. L., Schreifels, J. M., Beadnell, T. C. & Hackett, P. B. Duration of expression and activity of *Sleeping Beauty* transposase in mouse liver following hydrodynamic DNA delivery. *Mol. Ther.* **18,** 1796–1802 (2010).
17. Cinti, S., Zingaretti, M. C., Cancello, R., Ceresi, E. & Ferrara, P. Morphologic techniques for the study of brown adipose tissue and white adipose tissue. *Methods Mol. Biol.* **155,** 21–51 (2001).
18. Wu, J. *et al.* The unfolded protein response mediates adaptation to exercise in skeletal muscle through a PGC-1α/ATF6α complex. *Cell Metab.* **13,** 160–169 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was supported by National Institutes of Health grants DK54477, DK31405, DK61562 to B.M.S. P.B. and E.A.B. were supported by the Wenner-Gren Foundation, Swedish Heart and Lung Foundation and the 'Svenska Sällskapet för Medicinsk Forskning'. J.W. was supported by a postdoctoral fellowship from the American Heart Association (Founders Affiliate #09POST2010078). The animal procedures were in accordance with Institutional Animal Use and Care Committee protocols 110-2008 and 056-2009. The authors thank S. Loffredo and M. Kirschner for discussions and suggestions on the manuscript.

Author Contributions P.B. and B.M.S. planned the majority of experiments and wrote the paper, and P.B. executed most of the experiments. J.W. performed a subset of cultured cell experiments and contributed valuable materials. M.P.J. and S.P.G. performed the peptide fingerprinting identification of irisin cleavage. A.K. contributed with technical assistance and L.Y. and S.K. performed the CLARK electrode experiments. E.A.B. assisted with the hydrodynamic injections. J.C.L. assisted with intravenous injections and K.A.R. with bioinformatics. J.Z.L. and J.H.C. performed *in vitro* experiments. P.B. and H.T. and LakePharma designed and provided Fc fusion proteins. K.H. and B.F.V. performed the human cohort study, and M.C.Z. and S.C. performed the electron microscopy studies.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.M.S. (bruce_spiegelman@dfci.harvard.edu).

METHODS

Materials. Antibodies against UCP1, tubulin and FNDC5 were from Abcam. Forskolin, insulin, dexamethasone, rosliglitazone, GW6471 and antibody against Flag were from Sigma. Primers for all qPCR experiments are listed in Supplementary Table 1. Recombinant FNDC5, LRG1, IL-15, VEGF-β and TIMP4 were from ABNOVA (Taiwan). Coomassie staining kit and Lipofectamine 2000 were from Invitrogen.

Identification of PGC1-α-dependent secreted proteins. All PGC1-α-induced genes as judged from gene expression analysis in MCK-PGC1-α muscle with a fold change of at least 2 and $P < 0.05$ were subjected to the following analysis. The protein sequence of the longest transcript were analysed in the SignalP-software[19] (http://www.cbs.dtu.dk/services/SignalP/). Sequences with positive *S*, *C*, *Y* and *D* scores were considered positive for a signal sequence. All positive proteins were then screened for mitochondrial target sequences (http://www.cbs.dtu.dk/services/TargetP/) whereas positive sequences were removed. All remaining hit proteins were then analysed using qPCR in muscle from MCK-PGC1-α mice and myocytes overexpressing PGC1-α.

Primary cell cultures and recombinant protein treatments. The SVF from inguinal fat depots of 8–12-week-old BALB/C mice were prepared and differentiated for 6 days as previously described[20]. Rosiglitazone was used for the two first days of differentiation. For all experiments, unless otherwise indicated, recombinant FNDC5 was added to the culture media at a concentration of 1 μg ml^{-1} for the last 4 days of differentiation. Primary myoblasts were cultured and differentiated as described previously[21].

Preparation of protein fractions from cells and media. 293HEK or primary myocytes were transfected by standard methods or transduced with adenovirus at a multiplicity of infection (m.o.i.) of 20 as indicated. Twenty-four hours after transfection, media was removed, and cells were washed in large volumes of PBS five times, followed by incubation in Freestyle serum-free media (GIBCO) for 24 h. The cells and media were then collected separately, and media were centrifuged three times at 2,000*g* to pellet debris. Thereafter, a quarter volume of ice-cold TCA was added and precipitated protein was pelleted at 6,000*g* and washed three times in acetone. Pellet was then dried and resuspended in SDS-containing lysis buffer. Protein concentration was measured in both cell and media fraction and adjusted either by protein or volume as indicated.

RT–PCR. qPCR was carried out after Trizol-based RNA extraction using RNAeasy (Invitrogen) and thereafter SYBR green. All data were normalized to TBP, 18S or indicated in-house genes and quantitative measures were obtained using the $\Delta\Delta C_T$ method.

Western blot and quantification. Protein amounts from all samples were assessed using the BCA-kit (Thermo Scientific) followed by protein concentration normalization before all western blot experiments. Western blot was carried out following standard procedure and final band intensity (QL-BG) was quantified using BioPix iQ[22]. All data were normalized to background and loading controls.

Additional methods. CLARK electrode measurements, energy expenditure *in vivo*, interaperitoneal glucose tolerance test (IGTT) and immunohistochemistry against UCP1 were performed as described previously[6], with the exception that CLARK output was normalized to total cell protein. Fc-fusion construction and protein purification was performed by LakePharma.

Comprehensive laboratory animal monitoring system. C57/Bl6J mice were fed a high fat (60% kcal) diet (D12492, Research Diets) for 20 weeks, starting at 6 weeks of age. Mice were then injected with indicated doses of adenovirus expressing GFP or FNDC5, and comprehensive laboratory animal monitoring system (CLAMS; Columbia Instruments)-cage analysis was performed as described previously[6]. Briefly, mice were acclimated in metabolic chambers for 2 days before analysis in order to minimize stress. CO_2 and O_2 levels were then collected every 36 min for a period of 3 days. Data on activity, heat generation and food intake were measured at more frequent intervals. Circadian rhythm was observed for most parameters. Data were not normalized to body weights unless otherwise stated.

Mass spectrometry and peptide fingerprinting. Gel bands were digested with sequencing grade trypsin (Promega) or ASP-N (Sigma-Aldrich) as per manufacturer's instructions. Extracted in-gel protein digests were resuspended in 8 μl 5% formic acid/5% acetonitrile, and 4 μl were analysed by microcapillary liquid chromatography electrospray ionization tandem mass spectrometry (LC-MS/MS). Analyses were done on a LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) equipped with a Thermo Fisher Scientific nanospray source, an Agilent 1100 Series binary HPLC pump and a Famos autosampler. Peptides were separated on a 100 m × 16 cm fused silica microcapillary column with an in-house made needle tip. The column was packed with MagicC18AQ C_{18} reversed-phase resin (particle size, 5 μm; pore size, 200 Å; Michrom Bioresources). Separation was achieved through applying a 30 min gradient from 0–28% acetonitrile in 0.125% formic acid. The mass spectrometer was operated in a data-dependent mode essentially as described previously[23] with a full MS scan acquired with the Orbitrap, followed by up to ten LTQ MS/MS spectra on the most abundant ions detected in the MS scan. Mass spectrometer settings were: full MS (automated gain control, 1×10^6; resolution, 6×10^4; *m*/*z* range, 375–1,500; maximum ion time, 1,000 ms); MS/MS (AGC, 5×10^3; maximum ion time, 120 ms; minimum signal threshold, 4×10^3; isolation width, 2 Da; dynamic exclusion time setting, 30 s). After MS data acquisition, RAW files were converted into mzXML format and processed using a suite of software tools developed in-house for analysis. All precursors selected for MS/MS fragmentation were confirmed using algorithms to detect and correct errors in monoisotopic peak assignment and refine precursor ion mass measurements. All MS/MS spectra were then exported as individual DTA files and searched with no enzyme using the Sequest algorithm. These spectra were then searched against a database containing sequence of mouse FNDC5 in both forward and reversed orientations. The following parameters were selected to identify FNDC5: 10 p.p.m. precursor mass tolerance, 0.8 Da product ion mass tolerance, fully tryptic or ASP-N digestion, and up to two missed cleavages. Variable modifications were set for methionine (+15.994915). In addition, a fixed modification for the carbamidomethylation for cysteine (+57.021464) was used as well. The C-terminal fragment for FNDC5 was identified (KDEVTMKE) by trypsin digestion and reconfirmed by a separate ASP-N digestion.

Preparation of plasma samples for western blot. Thirty-five microlitres of mouse or human plasma were precleared for albumin/IgG using the ProteoExtract-kit (CalBiochem) as recommended by the manufacturer. Samples were then concentrated to approximately 100 μl and >8 μg μl^{-1}, followed by deglycosylation of 150 μg using PNGase F (New England Biolabs). Totally, 80 μl were then prepared containing 1× sample buffer with reducing agent and 1.7 μg μl^{-1} protein, sonicated, boiled and analysed using western blot against FNDC5 or indicated antibody.

Construction of adenoviral vectors. The FNDC5 expression vector was purchased with a C-terminal Flag-tag from OriGene. The QuickChange Multi Site XL Directed Mutagenesis Kit (Aligent Technologies) was used to introduce a Flag tag downstream of the signal sequence and to mutate the C-terminal Flag tag, thus resulting in the N-terminal Flag (NTF)–FNDC5 construct (Fig. 5a). The NTF and CTF FNDC5 constructs were then subcloned into the pENTR1a vector (Invitrogen) and recombined into the pAd-CMV-DEST-V5 vector (Invitrogen) and adenovirus was produced using the virapower system (Invitrogen), including three rounds of amplification. Thereafter, virus was concentrated using the Vivapure adenopack 100 (Sartorius Stedim Biotech) and buffer exchanges to saline reaching a concentration of 9–10 i.f.u. μl^{-1}. A GFP-containing adenovirus previously used was prepared in parallel.

Transgenic mice. The MCK-PGC1-α transgenic and muscle-specific PGC1-α knockout mice have been described previously[24].

Exercise protocols. Twelve-week-old B6 mice were exercised either using swimming[25] or using free wheel running, as described previously[26].

Human material and exercise training program. Blood samples and skeletal muscle biopsies were obtained from eight male non-diabetic individuals before and after 10 weeks of aerobic training as described previously[7]. In brief, the exercise-training program consisted of cycling on stationary bikes with 4–5 sessions of 20–35 min per week at an average intensity of ~65% of maximal oxygen consumption. Informed consent was obtained from all volunteers before participation. The study was approved by the Local Ethics Committee and was performed in accordance with the Helsinki Declaration.

19. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2,** 953–971 (2007).
20. Kajimura, S. *et al.* Initiation of myoblast to brown fat switch by a PRDM16-C/EBP-β transcriptional complex. *Nature* **460,** 1154–1158 (2009).
21. Rasbach, K. A. *et al.* PGC-1α regulates a HIF2α-dependent switch in skeletal muscle fiber types. *Proc. Natl Acad. Sci. USA* **107,** 21866–21871 (2010).
22. Bostrom, P. *et al.* The SNARE protein SNAP23 and the SNARE-interacting protein Munc18c in human skeletal muscle are implicated in insulin resistance/type 2 diabetes. *Diabetes* **59,** 1870–1878 (2010).
23. Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature Protocols* **3,** 1630–1638 (2008).
24. Handschin, C. *et al.* Skeletal muscle fiber-type switching, exercise intolerance, and myopathy in PGC-1α muscle-specific knock-out animals. *J. Biol. Chem.* **282,** 30014–30021 (2007).
25. Boström, P. *et al.* C/EBPβ controls exercise-induced cardiac growth and protects against pathological cardiac remodeling. *Cell* **143,** 1072–1083 (2010).
26. Chinsomboon, J. *et al.* The transcriptional coactivator PGC-1α mediates exercise-induced angiogenesis in skeletal muscle. *Proc. Natl Acad. Sci. USA* **106,** 21401–21406 (2009).

LETTER

doi:10.1038/nature10781

Genetic contributions to stability and change in intelligence from childhood to old age

Ian J. Deary[1,2]*, Jian Yang[3]*, Gail Davies[1,2], Sarah E. Harris[2,4], Albert Tenesa[4,5], David Liewald[1,2], Michelle Luciano[1,2], Lorna M. Lopez[1,2], Alan J. Gow[1,2], Janie Corley[1], Paul Redmond[1], Helen C. Fox[6], Suzanne J. Rowe[5], Paul Haggarty[7], Geraldine McNeill[6], Michael E. Goddard[8], David J. Porteous[2,4], Lawrence J. Whalley[6], John M. Starr[2,9] & Peter M. Visscher[2,3,10,11]*

Understanding the determinants of healthy mental ageing is a priority for society today[1,2]. So far, we know that intelligence differences show high stability from childhood to old age[3,4] and there are estimates of the genetic contribution to intelligence at different ages[5,6]. However, attempts to discover whether genetic causes contribute to differences in cognitive ageing have been relatively uninformative[7–10]. Here we provide an estimate of the genetic and environmental contributions to stability and change in intelligence across most of the human lifetime. We used genome-wide single nucleotide polymorphism (SNP) data from 1,940 unrelated individuals whose intelligence was measured in childhood (age 11 years) and again in old age (age 65, 70 or 79 years)[11,12]. We use a statistical method that allows genetic (co)variance to be estimated from SNP data on unrelated individuals[13–17]. We estimate that causal genetic variants in linkage disequilibrium with common SNPs account for 0.24 of the variation in cognitive ability change from childhood to old age. Using bivariate analysis, we estimate a genetic correlation between intelligence at age 11 years and in old age of 0.62. These estimates, derived from rarely available data on lifetime cognitive measures, warrant the search for genetic causes of cognitive stability and change.

General cognitive ability (also known as general intelligence, or *g*[18]) is an important human trait. It shows consistent and strong associations with important life outcomes such as educational and occupational success, social mobility, health, illness and survival[18]. Maintaining good general cognitive ability in old age is associated with better physical health and the ability to carry out everyday tasks[19,20]. Intelligence differences are highly heritable from adolescence, and through adulthood to old age[5,6]. Long-term follow-up studies have shown that about half of the phenotypic variance in general intelligence in old age is accounted for by its measure in childhood[3,4]. The corollary of this is that there are systematic changes through the life course in the rank order of intelligence between people; that is, some people's intelligence ages better than others. The determinants of stability and change in intelligence across the human life course are being sought, and candidate determinants include a wide range of genetic and environmental factors[1,5,7,19,21,22]. There have been longitudinal studies within childhood/adolescence, middle adulthood and old age, but none that stretches from childhood to old age with the same individuals (to our knowledge). Until now, the proportion of the variance in lifetime cognitive stability and change explained by genetic and environmental causes has been almost unknown. Apart from a small contribution from variation in the *APOE* gene, suggested individual genetic contributions to stability and change in intelligence across the life course are largely unreplicated[22]. Therefore, an important novel contribution would be to partition the covariance between intelligence scores at either end of the human life course into genetic and environmental causes. To address this, the present study applies a new analytical method[13–17] to genome-wide association data from human participants with general cognitive ability test scores in childhood and again in old age.

Participants were members of the Aberdeen Birth Cohort 1936 (ABC1936) and the Lothian Birth Cohorts of 1921 and 1936 (LBC1921, LBC1936)[11,12,17]. They are community-dwelling, surviving members of the Scottish Mental Surveys of 1932 (the 1921-born individuals) and 1947 (the 1936-born individuals), in which they took a well-validated test of general intelligence (Moray House Test) at a mean age of 11 years. They were traced and re-tested again in old age on a large number of medical and psychosocial factors for studies of healthy mental and physical ageing. Here, we use cognitive ability test data from childhood and from the first occasion of testing in old age for each subject. For all three cohorts, cognitive ability in old age was measured using the first unrotated principal component from a number of diverse cognitive tests. Additionally, the LBC1921 and LBC1936 cohorts re-took the Moray House Test in old age. Thus, the present study partitions into genetic and environmental causes the variance in stability and change in general intelligence over a period of between 54 and 68 years. Testing for 599,011 SNPs was performed on the Illumina610-Quadv1 chip (Illumina); the genotyping of the samples in this study was described previously[17] and quality control is described in Methods Summary.

To estimate additive genetic and environmental contributions to variation in cognitive ageing we used genotype information from 536,295 genome-wide autosomal SNPs. The method used here is a multivariate extension of our recently developed method, which allows the estimation of distant relationships between conventionally unrelated individuals from the SNP data and correlates genome-wide SNP similarity with phenotypic similarity[13,15]. A detailed description of the overall approach and statistical methods is given in Supplementary Fig. 1 and the Supplementary Note. We used a linear mixed model to estimate variance components. The methodology for the estimation of genetic variation from population samples was described previously and has been applied to continuous traits, including height, body-mass index and cognitive ability[13,15–17], and to disease[23]. The method is analogous to a pedigree analysis, with the important difference that we estimate distant relatedness from SNP markers. Because the relationships are estimated from common SNP markers, phenotypic variance explained by such estimated relationships is due to linkage

[1]Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. [2]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. [3]Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia. [4]Medical Genetics Section, Molecular Medicine Centre, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK. [5]The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Edinburgh EH25 9RG, UK. [6]Institute of Applied Health Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [7]Nutrition and Epigenetics Group, Rowett Institute of Nutrition and Health, University of Aberdeen, Greenburn Road, Bucksburn, Aberdeen AB21 9SB, UK. [8]Department of Food and Agricultural Systems, University of Melbourne, Parkville, Victoria 3011, Australia and Victorian Department of Primary Industries, Bundoora, Victoria 3083, Australia. [9]Alzheimer Scotland Dementia Research Centre, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. [10]University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia. [11]The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia.
*These authors contributed equally to this work.

disequilibrium between the genotyped markers and unknown causal variants[13,14,21]. The method estimates genetic variation from SNPs that are in linkage disequilibrium with unknown causal variants, and so provides a lower limit of the total narrow sense heritability because additive variation due to variants that are not in linkage disequilibrium with the genotyped SNPs is not captured.

We first performed a univariate analysis of cognitive ageing (Supplementary Note), which we had defined previously as intelligence scores in old age phenotypically adjusted for intelligence at childhood, by fitting the Moray House Test of intelligence at age 11 as a linear covariate[24]. We estimated that 0.24 (standard error 0.20) of phenotypic variance in cognitive ageing was accounted for by the SNP-based similarity matrix. We next conducted a bivariate genetic analysis of intelligence scores early and later in life, to partition the observed phenotypic covariance in intelligence measured in childhood and old age into genetic and environmental sources of variation. Information on the environmental correlation comes from the comparison of the two phenotypes within individuals whereas the genetic correlation is inferred from between-individual comparisons of the two phenotypes (Supplementary Note). That is, the analysis can inform us about genetic and environmental contributions to stability and change in intelligence across the life course. The phenotypic correlation between Moray House Test intelligence at age 11 and the general intelligence component in old age was 0.63 (standard error 0.02) (Table 1). The bivariate analysis resulted in estimates of the proportion of phenotypic variation explained by all SNPs for cognition, as follows: 0.48 (standard error 0.18) at age 11; and 0.28 (standard error 0.18) at age 65, 70 or 79 (referred to hereafter as 65–79). The genetic correlation between these two traits was 0.62 (standard error 0.22), and the environmental correlation was 0.65 (standard error 0.12). From the results of the bivariate analyses we can make a prediction about the proportion of phenotypic variance explained by the SNPs for cognition at 65–79 years given the phenotype at age 11 years. This provided a prediction of 0.21 (standard error 0.20), which is consistent with the actual estimate of 0.24 (standard error 0.20) from the univariate analysis (Supplementary Table 1), suggesting that the bivariate normal distribution assumption underlying the bivariate analysis is reasonable. Hence, the results from the bivariate analysis contain the full description of the genetic and environmental relationships between cognition at childhood, cognition at old age, and cognitive change. We re-ran this model with different cut-offs for relatedness (Supplementary Table 2). The estimates are very similar but with, as expected, larger standard errors for more stringent cut-offs, which result in a smaller sample size. This shows that the results are not driven by unusually high correlations for a few close relatives.

In the present analyses we did not adopt the usual procedure of dividing the parameter estimates by the standard errors to obtain test statistics and accompanying *P* values, because the standard errors were derived from a first-order Taylor series of the logarithm of the likelihood about the parameter estimates[25] and these can be biased for modest sample sizes. A more appropriate procedure is to use the likelihood-ratio test statistic to test the hypotheses that the genetic correlation coefficient is zero (no genetic correlation) or 1 (perfect genetic correlation). When using a likelihood-ratio test, the estimated genetic correlation coefficient of 0.62 has a borderline significant difference from zero (likelihood-ratio test statistic = 2.56, $P = 0.055$, one-sided test) (Supplementary Fig. 2), and does not differ significantly from 1. This was tested by fitting a repeatability model (which implies a genetic correlation of 1.0 and the same heritability of repeat observations) that has three fewer parameters than the full bivariate model. It resulted in a very similar value of the maximum log-likelihood value; the likelihood-ratio test statistic was 5.6 ($P = 0.133$, 3 degrees of freedom) (Supplementary Table 3).

LBC1921 and LBC1936 had the same Moray House Test administered at age 11 and again in old age. The bivariate analyses were repeated, therefore, using the same test of intelligence in childhood and old age in this subsample of the cohorts. The phenotypic correlation between Moray House Test intelligence at age 11 and in old age was 0.68 (standard error 0.01) (Table 1). The bivariate analysis resulted in estimates of the proportion of phenotypic variation explained by all SNPs for the Moray House Test, as follows: 0.30 (standard error 0.23) at age 11; and 0.29 (standard error 0.22) at age 70–79. The genetic correlation between these two traits was 0.80 (standard error 0.27). When using a likelihood-ratio test, the estimated genetic correlation coefficient of 0.80 is not significantly different from zero (likelihood-ratio test statistic = 1.51, $P = 0.11$). The environmental correlation between these two traits was 0.63 (standard error 0.13). From the results of the bivariate analyses we can make a prediction of the proportion of phenotypic variance explained by the SNPs for the Moray House Test at 70–79 years conditional on the phenotype at age 11 years. This results in an estimate of 0.074 (standard error 0.24) (Supplementary Table 4). Although the standard errors of the estimates are larger because a smaller data set was used, the results are similar to those using the full data and it appears that the choice of phenotype at old age (Moray House Test or a linear combination of a number of tests) has not led to a bias in inference. The estimates suggest that cognition early and late in life are similar traits, with possibly some genetic variation for cognitive change.

Using population-based genetic analyses, we have quantified, for the first time, the genetic and environmental contribution to stability and change in intelligence differences for most of the human lifespan. Genetic factors seem to contribute much to the stability of intelligence differences across the majority of the human lifespan. We provide a lower limit of the narrow sense heritability of lifetime cognitive ageing. The point estimate using a general cognitive ability component in old age is 0.24, albeit with a large standard error (0.20). We describe the estimate as a lower limit because the methods used in the present study allow us only to estimate the proportion of the genetic variation contributing to cognitive ageing that is captured by genetic variants in linkage disequilibrium with common SNPs; this will be lower than the total narrow sense heritability. We do not have a good estimate of the total amount of additive genetic variation for cognitive ageing, and so we cannot easily quantify any heritability that is missing from our estimate. Some of the possible genetic contribution we have found to cognitive change might be attributable to developmental change between age 11 and young adulthood. However, the large phenotypic correlation between age 11 and old-age intelligence, and the fact that heritability estimates of general intelligence by age 11 are at about adult levels[5], lead us to posit that most of the genetic variation we have found is a contribution to ageing-related cognitive changes. The estimate of the genetic contribution to lifetime cognitive change was lower when, for a subsample, the same test was used in childhood and old age.

The bivariate analysis conducted here quantifies how differences in intelligence early and late in life are attributable to environmental

Table 1 | Bivariate analysis of intelligence at age 11 and at age 65–79

	Using general intelligence component in old age		Using Moray House Test in old age	
	Estimate	Standard error*	Estimate	Standard error*
h_1^2	0.478	0.177	0.298	0.229
h_2^2	0.280	0.177	0.289	0.221
r_G	0.623	0.218	0.798	0.266
r_e	0.652	0.125	0.630	0.132
r_P	0.627	0.015	0.680	0.014

Where h_1^2 and h_2^2 are variance explained by all SNPs for intelligence at age 11 and old age, respectively; r_G is genetic correlation; r_e is residual correlation; r_P is phenotypic correlation. A total of 1,940 unrelated individuals were included with the general intelligence component phenotype data at childhood (1,830) or old age (1,839) (1,729 individuals had both phenotypes). Of the 1,515 LBC1921 and LBC1936 individuals, there were 1,391 with genetic information and Moray House Test scores both at age 11 and in old age.

* The standard errors are estimated from a first-order Taylor series expansion about the estimated maximum likelihood values and may be biased downwards[25]. For testing hypotheses we have used the likelihood-ratio test statistic, which is more accurate.

or genetic factors. A genetic correlation of zero would imply that intelligence early and late in life are entirely separate traits genetically, and that variation in the change in intelligence from childhood to old age is partly genetic and a function of the heritability of intelligence early and late in life. At the other extreme, a genetic correlation of one implies that the two traits have the same genetic determinants, so that any variation in the change in intelligence between the two stages in life is purely environmental. At conventional levels of significance we could not rule out either a genetic correlation of zero or one; however, our estimates suggest that genetics and environment could each contribute substantially to the covariance between intelligence at age 11 and old age, and that genetic factors might have a role in cognitive change between the two stages of the life course.

The samples studied here comprise the birth cohorts' survivors, those healthy enough to take part in the studies, and people with less cognitive decline. Therefore, we considered whether our estimate of genetic variation at older ages may be biased downwards because of censoring. From life tables officially published by the Scottish Government based on census data, we estimate that the individuals in our oldest sample who were born in 1921 and alive at age 11 are among the ~50% that were still alive at the time of sample collection. We know that lower childhood cognitive ability per se is associated with premature mortality[26], which, of course, our analyses adjust for, as specified in the models. However, because there is a paucity of data about genetic influences on lifetime cognitive change, we have limited information with regard to how these might affect life expectancy. The only way to know across the lifespan would have been if all children (that is, the ones who survived to older ages—whom we know about—and the ones who did not) had been genotyped in 1947. For non-normative (that is, pathological) cognitive change, there are genetic risk factors associated with younger-onset Alzheimer's disease that result in premature mortality, but such strongly heritable disease is rare and the genes do not seem to affect normative cognitive ageing in those aged 70 years and over[22]. Hence, this is not a concern with regard to our analyses. *APOE* ε4 is a well-known risk factor for non-normative cognitive decline, but any differential effect on survival occurs later in life, and is thus unlikely to have resulted in attrition in our cohort. Moreover, *APOE* is in Hardy–Weinberg equilibrium in even our oldest samples[24], supporting this inference. Other known genetic risk factors for Alzheimer's disease have a very small effect on the risk of disease[27]. Hence, a priori, we have nothing to suggest anything but a largely neutral effect of genes that influence cognitive ageing on survival. However, if there is an effect, the example of cognition[26] (by contrast with cognitive change) would suggest that this would be negative, which would somewhat reduce genetic variation in cognitive change across the lifespan among the survivors.

Until now, studies aimed at finding genetic contributions to cognitive ageing have offered little information. They use too-short follow-up periods, thereby providing too small an amount of cognitive change[7,22]. Cognitive assessments tend to be made only within old age, even though cognitive ageing occurs from young adulthood onwards. They are largely based on behavioural data in twin samples rather than information on DNA variation. The present study is unusual and valuable in capturing over half a century of cognitive stability and change and examining its causes. The results here provide estimates for the genetic and environmental contributions to cognitive stability and change across most of the human lifespan. Even with almost 2,000 individuals, the study's power was insufficient to achieve conventional levels of significance for the estimates. Our emphasis here has not been on the traditional significance thresholds for *P* values per se, but in trying to partition variance in cognitive ability into environmental and genetic causes. The phenotypes available here are rare, and so these point estimates are useful to guide future research. The present findings render attractive a search for genetic mechanisms of cognitive change across the life course. They also suggest the importance of environmental contributions to lifetime cognitive change.

METHODS SUMMARY

Subjects. Recruitment, phenotyping and genotyping of the samples were described previously[11,12,17]. The mental test at age 11 was a Moray House Test[11,12]. In old age, general intelligence was derived using principal components analysis of a number of mental tests and saving scores on the first unrotated principal component (Supplementary Note). In old age, the assessments of general intelligence were made at ages as follows: ABC1936, 64.6 years (standard deviation 0.9); LBC1936, 69.5 (standard deviation 0.8); LBC1921, 79.1 (standard deviation 0.6). The LBC1921 and LBC1936 samples, but not the ABC1936, had repeat testing of the Moray House Test (already taken at age 11 years) at 79.1 and 69.5 years, respectively. After applying the genome-wide complex trait analysis method[13,15], the distribution of inferred relationships in the samples was as shown in Supplementary Fig. 3. We removed one of each pair of individuals whose estimated genetic relatedness was >0.2. We retained 1,940 individuals with childhood or old-age phenotype data (1,729 individuals had both): ABC1936, 425; LBC1921, 512; and LBC1936, 1,003. Of the 1,515 LBC1921 and LBC1936 individuals, there were 1,391 with genetic information and Moray House Test scores at age 11 and in old age.

Genotyping quality control. Quality control procedures were performed per SNP and per sample. Individuals were excluded from further analysis if genetic and reported gender did not agree. Samples with a call rate ≤ 0.95, and those showing evidence of non-European descent by multidimensional scaling, were removed[17]. SNPs were included in the analyses if they met the following conditions: call rate ≥ 0.98, minor allele frequency ≥ 0.01, and Hardy–Weinberg equilibrium test with $P \geq 0.001$. After these quality control stages, 1,948 samples remained (ABC1936, $N = 426$; LBC1921, $N = 517$; LBC1936, $N = 1{,}005$), and 536,295 autosomal SNPs were included in the analysis.

Received 5 September; accepted 12 December 2011.
Published online 18 January 2012.

1. Plassman, B. L., Williams, J. W., Burke, J. R., Holsinger, T. & Benjamin, S. Systematic review: factors associated with risk for and possible prevention of decline in later life. *Ann. Intern. Med.* **153,** 182–193 (2010).
2. Brayne, C. The elephant in the room—healthy brains in later life, epidemiology and public health. *Nature Rev. Neurosci.* **8,** 233–239 (2007).
3. Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R. & Starr, J. M. The stability of individual differences in mental ability from childhood to old age: follow-up of the 1932 Scottish Mental Survey. *Intelligence* **28,** 49–55 (2000).
4. Gow, A. J. *et al.* Stability and change in intelligence from age 11 to ages 70, 79, and 87: the Lothian Birth Cohorts of 1921 and 1936. *Psychol. Aging* **26,** 232–240 (2011).
5. Deary, I. J., Johnson, W. & Houlihan, L. M. Genetic foundations of human intelligence. *Hum. Genet.* **126,** 215–232 (2009).
6. Deary, I. J., Penke, L. & Johnson, W. The neuroscience of human intelligence differences. *Nature Rev. Neurosci.* **11,** 201–211 (2010).
7. Lee, T., Henry, J. D., Trollor, J. N. & Sachdev, P. S. Genetic influences on cognitive functions in the elderly: a selective review of twin studies. *Brain Res. Rev.* **64,** 1–13 (2010).
8. Reynolds, C. A. *et al.* Quantitative genetic analysis of latent growth curve models of cognitive abilities in adulthood. *Dev. Psychol.* **41,** 3–16 (2005).
9. Finkel, D., Reynolds, C. A., McArdle, J. J., Hamagami, F. & Pedersen, N. L. Genetic variance in processing speed drives variation in aging of spatial and memory abilities. *Dev. Psychol.* **45,** 820–834 (2009).
10. McGue, M. & Christensen, K. Social activity and healthy aging: a study of aging Danish twins. *Twin Res. Hum. Genet.* **10,** 255–265 (2007).
11. Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. & Fox, H. C. The impact of childhood intelligence in later life: following up the Scottish Mental Surveys of 1932 and 1947. *J. Pers. Soc. Psychol.* **86,** 130–147 (2004).
12. Deary, I. J. *et al.* The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr.* **7,** 28 (2007).
13. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42,** 565–569 (2010).
14. Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res. Hum. Genet.* **13,** 517–524 (2010).
15. Yang, J., Lee, H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
16. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genet.* **43,** 519–525 (2011).
17. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16,** 996–1005 (2011).
18. Deary, I. J. Intelligence. *Annu. Rev. Psychol.* **63,** 453–482 (2012).
19. Deary, I. J. *et al.* Age-associated cognitive decline. *Br. Med. Bull.* **92,** 135–152 (2009).
20. Tucker-Drob, E. M. Neurocognitive functions and everyday functions change together in old age. *Neuropsychology* **25,** 368–377 (2011).
21. Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Rev. Genet.* **11,** 800–805 (2010).

22. Harris, S. E. & Deary, I. J. The genetics of cognitive ability and cognitive ageing in healthy older people. *Trends Cogn. Sci.* **15,** 388–394 (2011).
23. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 294–305 (2011).
24. Deary, I. J. *et al.* Cognitive change and the *APOE* ε4 allele. *Nature* **418,** 932 (2002).
25. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51,** 1440–1450 (1995).
26. Calvin, C. M. *et al.* Intelligence in youth and all-cause mortality: systematic review with meta-analysis. *Int. J. Epidemiol.* **40,** 626–644 (2011).
27. Hollingworth, P. *et al.* Common variants at *ABCA7, MS4A6A/MS4A4E, EPHA1, CD33,* and *CD2AP* are associated with Alzheimer's disease. *Nature Genet.* **43,** 429–435 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the cohort participants who contributed to these studies. Genotyping of the ABC1936, LBC1921 and LBC1936 cohorts and the analyses conducted here were supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC). Phenotype collection in the LBC1921 was supported by the BBSRC, The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the LBC1936 was supported by Research Into Ageing (continues as part of Age UK's The Disconnected Mind project). Phenotype collection in the ABC1936 was supported by the BBSRC, the Wellcome Trust and the Alzheimer's Research Trust. The Australian-based researchers acknowledge support from the Australian Research Council and the National Health and Medical Research Council. M.L. is a Royal Society of Edinburgh/Lloyds TSB Foundation for Scotland Personal Research Fellow. The work was undertaken in The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698), for which funding from the BBSRC, EPSRC, ESRC and MRC is gratefully acknowledged.

Author Contributions I.J.D. and P.M.V. designed the study. J.Y. and P.M.V. performed statistical analyses, with I.J.D., M.E.G., A.T. and S.J.R. contributing to discussions regarding analyses. G.D., S.E.H., D.L., A.T., M.L. and L.M.L. performed quality control analyses and prepared data. S.E.H., M.L., L.M.L., A.J.G., J.C., P.R., H.C.F., S.J.R., P.H., L.J.W., G.M., D.J.P., J.M.S. and I.J.D. contributed genotype and phenotype data. I.J.D., P.M.V. and J.Y. contributed to writing the paper and Supplementary Information. All authors contributed to revising the paper and Supplementary Information.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to I.J.D. (i.deary@ed.ac.uk) or P.M.V. (peter.visscher@uq.edu.au).

LETTER

doi:10.1038/nature10791

A novel sensor to map auxin response and distribution at high spatio–temporal resolution

Géraldine Brunoud[1], Darren M. Wells[2]*, Marina Oliva[1]*, Antoine Larrieu[2,3]*, Vincent Mirabet[1], Amy H. Burrow[4], Tom Beeckman[3], Stefan Kepinski[4], Jan Traas[1], Malcolm J. Bennett[2] & Teva Vernoux[1]

Auxin is a key plant morphogenetic signal[1] but tools to analyse dynamically its distribution and signalling during development are still limited. Auxin perception directly triggers the degradation of Aux/IAA repressor proteins[2–6]. Here we describe a novel Aux/IAA-based auxin signalling sensor termed DII-VENUS that was engineered in the model plant *Arabidopsis thaliana*. The VENUS fast maturing form of yellow fluorescent protein[7] was fused in-frame to the Aux/IAA auxin-interaction domain (termed domain II; DII)[5] and expressed under a constitutive promoter. We initially show that DII-VENUS abundance is dependent on auxin, its TIR1/AFBs co-receptors[4–6,8] and proteasome activities. Next, we demonstrate that DII-VENUS provides a map of relative auxin distribution at cellular resolution in different tissues. DII-VENUS is also rapidly degraded in response to auxin and we used it to visualize dynamic changes in cellular auxin distribution successfully during two developmental responses, the root gravitropic response and lateral organ production at the shoot apex. Our results illustrate the value of developing response input sensors such as DII-VENUS to provide high-resolution spatio-temporal information about hormone distribution and response during plant growth and development.

Central to auxin signalling is the ubiquitin- and proteasome-dependent degradation of Aux/IAA catalysed by the SCF-type E3 ubiquitin-ligase complexes $SCF^{TIR1/AFB1-5}$ (refs 2–6, 8). Aux/IAA repressors form heterodimers with transcription factors termed auxin response factors (ARFs)[9,10]. Auxin directly promotes the interaction between TIR1/AFBs auxin co-receptors and Aux/IAAs[5], thus recruiting Aux/IAAs to the SCF complex[3,6] and derepressing ARF-bound loci. This allows the transcription of target genes including most *Aux/IAA* genes, hence providing a negative feedback loop (Fig. 1a)[2,10]. The most widely used tools to monitor auxin distribution *in planta* are *DR5*-based auxin-inducible reporters whose promoter contains several ARF binding sites[11,12]. However, as an output of the auxin response pathway (Fig. 1a), reporter activity does not directly relate to endogenous auxin abundance but also reflects the contribution of a complex signalling pathway[2].

Monitoring the degradation of an Aux/IAA-based green fluorescent protein (GFP) reporter would provide a better target for an auxin sensor as its signal can be related more directly to hormone abundance (Fig. 1a)[3,5,6,13]. This has proved very challenging because Aux/IAA half-lives are often shorter than GFP maturation time[7,14–17]. To overcome this technical limitation, we fused the VENUS fast maturing yellow fluorescent protein (YFP)[7] to the auxin-interaction domain (termed domain II; DII)[5] from several Aux/IAA proteins and expressed these fusion proteins under the constitutive 35S promoter (Fig. 1b and Supplementary Fig. 1a, b). Confocal imaging of transgenic root apical tissues revealed similar fluorescence patterns but the strongest signal

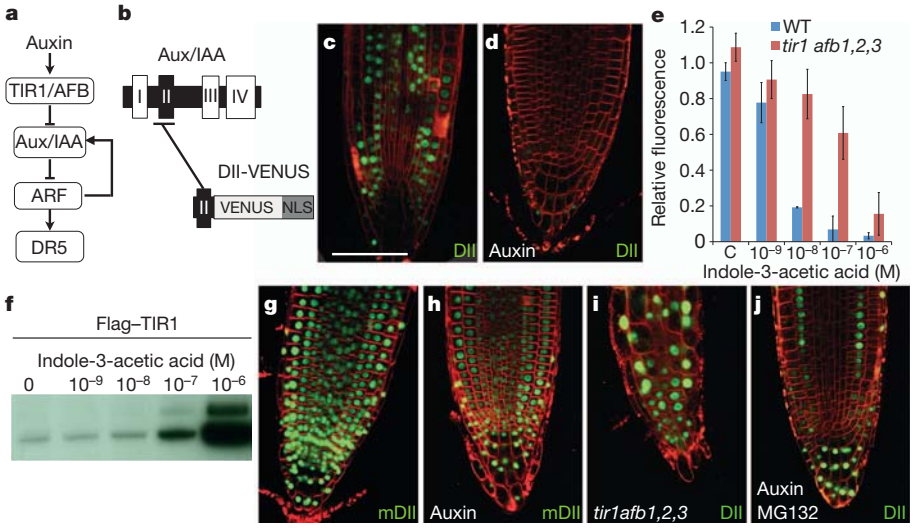


Figure 1 | DII-VENUS degradation is dependent on auxin, TIR1/AFBs and proteasome activity. **a**, Schematic representation of auxin signalling. **b**, Design of DII-VENUS; NLS, nuclear localization signal. **c**, **d**, DII-VENUS fluorescence in absence (**c**) or presence (**d**) of 1 μM indole-3-acetic acid. **e**, Dose-dependent degradation of DII-VENUS in wild-type and mutant roots. **f**, IAA28 domain II peptide pull-down assay with Flag–TIR1 in the presence of indole-3-acetic acid. **g**, **h**, mDII-VENUS fluorescence in the absence (**g**) or presence (**h**) of 1 μM indole-3-acetic acid. **i**, DII-VENUS signal in co-receptor quadruple mutant. **j**, DII-VENUS signal upon 1 μM indole-3-acetic acid and proteasome inhibitor (MG132) co-treatment. Green channel, VENUS; red channel, FM4-64. Scale bar, 50 μm. Error bars, s.d. (WT, $n = 2$; mutant, $n = 3$).

[1]Laboratoire de Reproduction et Développement des Plantes, CNRS, INRA, ENS Lyon, UCBL, Université de Lyon, 69364 Lyon, France. [2]Centre for Plant Integrative Biology, University of Nottingham, Sutton Bonington LE12 5RD, UK. [3]Department of Plant Systems Biology, VIB, Ghent University, B-9052 Gent, Belgium. [4]Centre for Plant Science, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK.
*These authors contributed equally to this work.

was obtained using domain II of the most stable Aux/IAA used, IAA28 (Fig. 1c and Supplementary Fig. 1c–f). We thus focused our analyses on this form of the sensor, henceforth called DII-VENUS.

We tested on root tissues the relationship between auxin, its response components and DII-VENUS in several independent ways. First, the DII-VENUS signal was sensitive in a dose-dependent fashion to exogenous auxin treatment (Fig. 1c–e). Second, the DII peptide interacted with its co-receptors TIR1 (Fig. 1f), AFB1 and AFB5 (ref. 18) in an auxin-dependent manner. Third, introducing a mutation in the domain II sequence of DII-VENUS (mDII-VENUS), that disrupts the interaction between Aux/IAA, auxin and the TIR1/AFBs[5], reduced the differential distribution of fluorescence (Fig. 1g; see below for description of pattern) and blocked its auxin-induced degradation (Fig. 1h). Fourth, DII-VENUS fluorescence was ubiquitously distributed in roots of the most strongly affected *tir1 afb1 afb2 afb3* quadruple mutant[4] (Fig. 1i) and the mutants were significantly less sensitive to auxin treatment (Fig. 1e). Fifth, disruption of ubiquitin-dependent breakdown of Aux/IAA proteins using proteasome inhibitors stabilized DII-VENUS and blocked its auxin-induced degradation (Fig. 1j and Supplementary Fig. 2). We conclude that DII-VENUS abundance is regulated by auxin via its receptors, consistent with the model for Aux/IAA degradation (Fig. 1a)[2]. We also demonstrated that DII-VENUS does not disrupt the activity of the auxin response machinery (Supplementary Fig. 3 and Supplementary Information). Hence, DII-VENUS directly reports, but does not interfere with, the input into the auxin signalling pathway.

We next took advantage of the simple cellular organization of the root apex to quantify the distribution of DII-VENUS fluorescence with cellular definition (Fig. 2a, b and Supplementary Fig. 4). Because the TIR1/AFB1-3 co-receptor distribution shows only limited variations in the root meristem region (except in the root cap; Supplementary Fig. 5a–d)[19], this will confer a homogeneous perception capacity. Hence, the spatial distribution of DII-VENUS fluorescence is likely to represent an inverted auxin distribution map in the root tip. This conclusion is further supported by the more homogeneous fluorescence distribution of mDII-VENUS (Fig. 2c and Supplementary Fig. 6) and by the complementary patterns of DII-VENUS and *DR5::VENUS* expression in the quiescent centre, columella and differentiating xylem cells (Fig. 2d, e)[11,20,21]. However, mDII-VENUS fluorescence distribution suggests a higher 35S promoter activity in the epidermis and cortex in the elongation zone and in the most external root cap cells (Fig. 2c and Supplementary Fig. 6). Lower expression of the TIR/AFBs is also expected to confer a lower sensitivity to auxin in the root cap (Supplementary Fig. 5a–d). In both cases, this will lead to an underestimation of auxin levels by DII-VENUS. Analyses of TIR/AFBs co-receptor distribution and of 35S promoter activity are thus essential to interpret the DII-VENUS pattern.

Even considering these biases (Supplementary Information), DII-VENUS quantification indicates that auxin levels are reproducibly higher in the first two tiers of columella cells and initials, the quiescent centre, the stele initials and early daughters and the differentiating xylem cells (Fig. 2b and Supplementary Fig. 4)[22]. The other cells in the root meristem have lower levels of auxin, with minima observed in the epidermis and cortex, but auxin levels significantly increased close to the start of the elongation zone. This increase occurs closer to the root tip in the epidermis and vasculature compared to the cortex (Fig. 2b). The DII-VENUS fluorescence map thus confirms the local maximum of auxin at the quiescent centre and in the columella cells[11] but also allows visualizing the distribution of auxin in the entire root tip. It also reveals a previously unsuspected auxin accumulation starting at the transition zone between the meristem and the elongation zone. These results are in partial agreement with measurements of auxin concentrations in root tissues obtained after cell sorting[23], the differences being possibly due to the higher resolution achieved using DII-VENUS.

We also detected differential distributions of DII-VENUS fluorescence in the vegetative shoot apical meristem (SAM), in the vascular tissues of the hypocotyl (Fig. 2f, g) and later during development in the inflorescence SAM and young floral meristems (Supplementary Fig. 7a–d)[18]. As in the root, reduced differential expression with mDII-VENUS, partly complementary *DR5::VENUS* patterns and distribution of TIR1/AFB1–3 co-receptors (Fig. 2h, i and Supplementary Fig. 5e–h) indicate that the distribution of DII-VENUS fluorescence is primarily controlled by auxin levels in the shoot apex (Supplementary Information)[18]. DII-VENUS is therefore able to report relative auxin distribution at high spatial resolution in various tissues and developmental stages. In addition, in both root and shoot tissues, DII-VENUS is degraded not only in cells where *DR5* is expressed but also in cells that do not express *DR5* (Figs 1c, 2a–f and Supplementary Fig. 7 and Supplementary Information)[18]. This observation demonstrates that the Aux/IAA-ARF signalling pathway contributes significantly to the definition of the *DR5* expression pattern.

To analyse the temporal resolution of the DII-VENUS sensor, we compared dynamic changes in DII-VENUS and *DR5::VENUS* signals

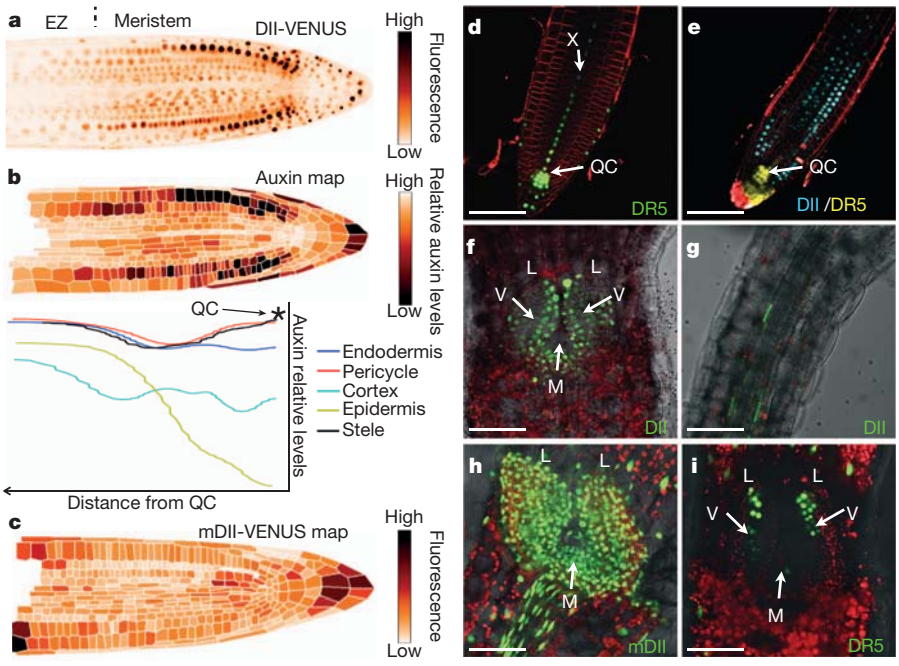


Figure 2 | DII-VENUS provides a sensor to map auxin distribution in plant tissues. **a**–**d**, Root meristematic tissues. **a**, **b**, DII-VENUS fluorescence (**a**) and corresponding map of relative auxin distribution (**b**; top) using the same look-up table. EZ, elongation zone. Tissue-specific changes in auxin levels along the root axis are shown (**b**; bottom). QC, quiescent centre. **c**, Fluorescence map of mDII-VENUS control. **d**, *DR5::VENUS*. X, xylem axis. **e**, DII-VENUS (blue) and *DR5::GFP* (yellow). **f**–**i**, DII-VENUS (**f**, **g**), mDII-VENUS (**h**) and *DR5::VENUS* (**i**) in the vegetative shoot apex (**f**, **h**, **i**) and hypocotyl (**g**). M, meristem; L, leaves; V, vasculature. Green or blue channel, VENUS; yellow channel in **e**, GFP; red channel, FM4-64 (**d**, **e**) or autofluorescence (**f**–**i**). In **f**–**i** the transmission channel has been added. Scale bars, 50 μm.

in roots following exogenous auxin treatment (Fig. 3a, b and Supplementary Movie 1). Time-lapse confocal imaging revealed that the DII-VENUS signal was rapidly lost from all root tissues, whereas the signal in untreated roots remained stable (Fig. 3a and Supplementary Movie 1). Quantification of VENUS fluorescence in the root tip showed that a reduction in the DII-VENUS signal was detected minutes after auxin addition and the signal was abolished within 60 min (Fig. 3b). In contrast, an increase in the *DR5::VENUS* signal was first detected only after 120 min (Fig. 3b). This delay is due to post-transcriptional processes, because quantitative reverse transcription-PCR (qRT–PCR) detected VENUS messenger RNA minutes after auxin treatment (Fig. 3b). We could further show that the DII-VENUS degradation kinetics upon auxin treatment is very similar in different root tissues, but is slower in the root cap (Supplementary Fig. 8). These results indicate similar auxin sensitivity throughout the root except for a lower sensitivity in the root cap, as already suggested by the distribution of the TIR/AFBs (Supplementary Fig. 5a–d). We also observed that the global dynamics of degradation of DII-VENUS was similar in the vegetative and inflorescence SAM, with a minimal fluorescence reached after 1 h (Fig. 3c and Supplementary Fig. 9). We conclude that DII-VENUS responds almost immediately and similarly to exogenous auxin application in various tissues. This observation strengthens our conclusion that DII-VENUS fluorescence is directly related to auxin levels in both shoot and root meristematic tissues but that co-receptor distribution needs to be considered.

Finally, we used DII-VENUS to follow changes in auxin distribution during developmental processes. Roots have been proposed to bend in response to gravity by accumulating auxin on the lower side of root apical tissues[21,24,25]. Consistent with this model, induction of the *DR5* reporter occurs after 1.5–2 h in the lateral root cap (LRC) and epidermis on the lower side of the root (Supplementary Fig. 10)[21]. By contrast, within 30 min of a 90° gravity stimulus the DII-VENUS signal was entirely lost in these tissues on the lower side, whereas fluorescence was stable on the upper side (Fig. 4a). A decrease in DII-VENUS fluorescence was also observed in the cortex and endodermis on both sides of the root and to a lesser extent in vascular tissues. Hence, DII-VENUS

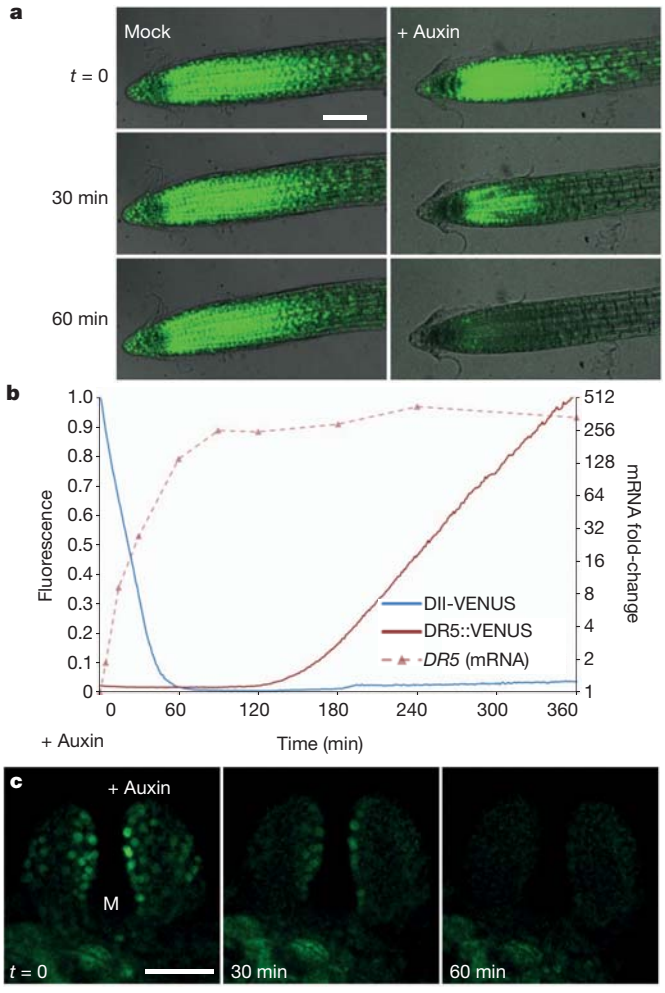


Figure 3 | DII-VENUS monitors changes in auxin response and distribution at high temporal resolution. **a**, Time-course of DII-VENUS fluorescence following either a mock or an auxin treatment (1 μM 1-naphthaleneacetic acid; NAA). **b**, Quantification of DII-VENUS and DR5::VENUS fluorescence and of *DR5::VENUS* mRNA levels in root apices treated with 1 μM NAA. **c**, Time-course of DII-VENUS fluorescence in the shoot apex upon 1 μM indole-3-acetic acid treatment; images are projections of 10 confocal serial sections. Green channel, VENUS. In **a** the transmission channel has been added. Scale bars, 50 μm.

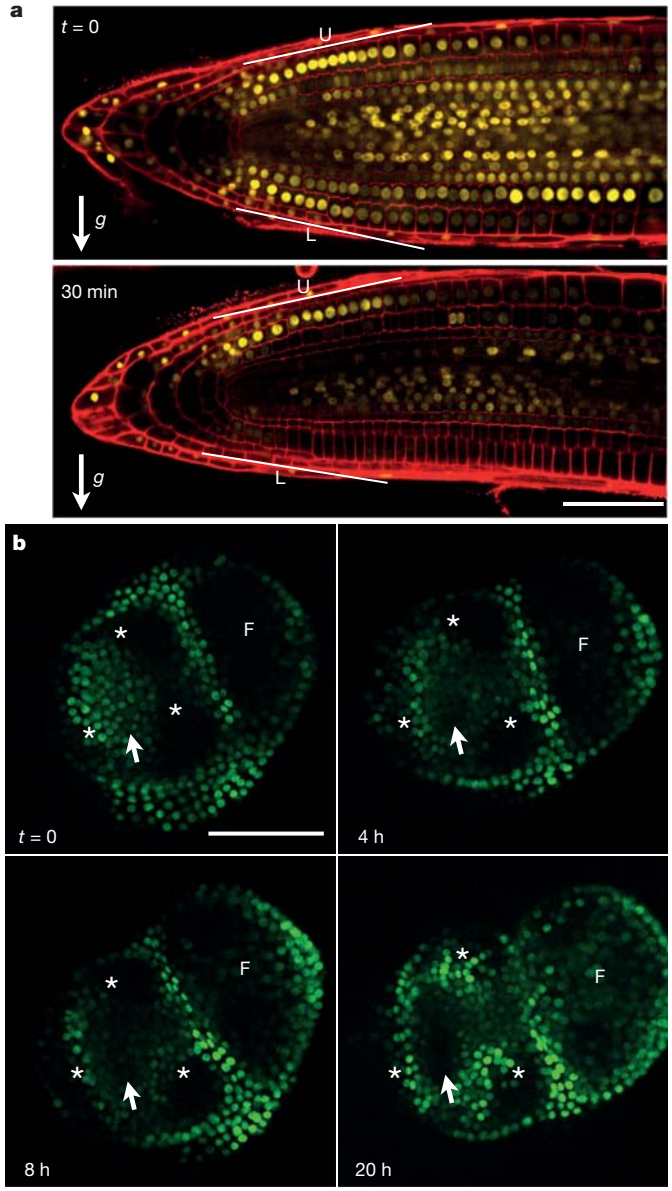


Figure 4 | DII-VENUS allows visualization of changes in auxin distribution during development. **a**, Changes in DII-VENUS fluorescence during a root gravitropic response; *g*, gravity vector; a line was drawn near equivalent cells in the LRC and epidermis on the lower (L) and upper (U) side. **b**, Auxin build-up during organ initiation at the shoot apex visualized using DII-VENUS; arrow, site of new organ initiation; asterisks, groups of nuclei showing notable changes in fluorescence; F, flower. Yellow or green channel, VENUS; red channel, propidium iodide. Scale bars, 50 μm.

indicates that the changes in auxin concentration are not restricted to the LRC and epidermis during the gravitropic response. We next used DII-VENUS to follow auxin-dependent organ initiation at the SAM (Fig. 4b)[26,27]. DII-VENUS allowed the visualization of the progressive build-up of auxin triggering the formation of a new organ (Fig. 4b). It also demonstrated significant redistribution of DII-VENUS fluorescence throughout the SAM, most probably reflecting the dynamics of auxin transport in the tissue[27]. Taken together, our observations demonstrate that, during both shoot and root development, DII-VENUS detects dynamic changes in endogenous auxin distribution and responses that are more complex than previously thought. By demonstrating that DII-VENUS and *DR5* fluorescence patterns are only partly complementary, we also provide evidence that the auxin signalling pathway has a key role in the spatial control of transcription in response to auxin during developmental processes. Finally, several other plant hormones have been shown to signal through degradation of key signalling regulators[28–30]. Our work provides the foundation for building synthetic signalling sensors for different hormones to explore their role during development.

METHODS SUMMARY

Generation and characterization of DII-VENUS reporter plants. Aux/IAA degron sequences[16] were fused to VENUS-NLS under the control of the 35S promoter (Supplementary Fig. 1) and recombined into gateway binary vector pH7m34GW (http://gateway.psb.ugent.be/). Plasmids were transformed into Col-0 plants by floral dipping.

Microscopy, live imaging and chemical treatments. All images were obtained using laser-scanning confocal microscope. Tissue organization was visualized using FM4-64, propidium iodide, transmission or chlorophyll autofluorescence. For live imaging of shoot meristems, plants were grown on the auxin transport inhibitor 1-*N*-naphthylphthalamic acid (NPA) to produce naked apices before transferring to a new medium without NPA. To generate maps of relative auxin distribution in the root meristem, a cellular grid was generated from a median optical section and for each cells fluorescence was extracted from the optical section cutting the nucleus at its most median part.

For analysis of chemically treated roots, 5-day-old DII-VENUS seedlings were transferred to media containing the chemicals at the stated concentration. For root live imaging, immediately following transfer, the seedlings were scanned every 2 to 5 min for 2 h. The fluorescence intensity over identical scanned portion of the root (corresponding approximately to the first 200 μM from the root tip) was extracted at each time point. For tissue-specific kinetics, fluorescence was extracted from groups of nuclei with nearly identical fluorescence in the different tissues. For the vegetative SAM, 5-day-old DII-VENUS seedlings were mounted in water after removing one cotyledon to allow observation of the vegetative shoot apex before treating with the stated concentration of auxin.

Pull-down assays. Pull-down assays were performed using a 35S:Flag-TIR1 line and biotinylated IAA28 peptide. The immunodetection of TIR1/AFB–Flag was performed with a 1:5,000 dilution of anti-Flag 2-peroxidase (HRP) antibody followed by chemiluminescent detection with ECL plus reagents.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 July; accepted 19 December 2011.
Published online 15 January 2012.

1. Friml, J. Auxin transport — shaping the plant. *Curr. Opin. Plant Biol.* **6,** 7–12 (2003).
2. Chapman, E. J. & Estelle, M. Mechanism of auxin-regulated gene expression in plants. *Annu. Rev. Genet.* **43,** 265–285 (2009).
3. Dharmasiri, N., Dharmasiri, S. & Estelle, M. The F-box protein TIR1 is an auxin receptor. *Nature* **435,** 441–445 (2005).
4. Dharmasiri, N. *et al.* Plant development is regulated by a family of auxin receptor F box proteins. *Dev. Cell* **9,** 109–119 (2005).
5. Tan, X. *et al.* Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature* **446,** 640–645 (2007).
6. Kepinski, S. & Leyser, O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature* **435,** 446–451 (2005).
7. Shaner, N. C., Steinbach, P. A. & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nature Methods* **2,** 905–909 (2005).
8. Greenham, K. *et al.* The AFB4 auxin receptor is a negative regulator of auxin signaling in seedlings. *Curr. Biol.* **21,** 520–525 (2011).
9. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. ARF1, a transcription factor that binds to auxin response elements. *Science* **276,** 1865–1868 (1997).
10. Guilfoyle, T. J. & Hagen, G. Auxin response factors. *Curr. Opin. Plant Biol.* **10,** 453–460 (2007).
11. Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99,** 463–472 (1999).
12. Ulmasov, T., Murfett, J., Hagen, G. & Guilfoyle, T. J. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9,** 1963–1971 (1997).
13. Gray, W. M., Kepinski, S., Rouse, D., Leyser, O. & Estelle, M. Auxin regulates SCF^{TIR1}-dependent degradation of AUX/IAA proteins. *Nature* **414,** 271–276 (2001).
14. Ramos, J. A., Zenser, N., Leyser, O. & Callis, J. Rapid degradation of auxin/indoleacetic acid proteins requires conserved amino acids of domain II and is proteasome dependent. *Plant Cell* **13,** 2349–2360 (2001).
15. Zenser, N., Ellsmore, A., Leasure, C. & Callis, J. Auxin modulates the degradation rate of Aux/IAA proteins. *Proc. Natl Acad. Sci. USA* **98,** 11795–11800 (2001).
16. Dreher, K. A., Brown, J., Saw, R. E. & Callis, J. The *Arabidopsis* Aux/IAA protein family has diversified in degradation and auxin responsiveness. *Plant Cell* **18,** 699–714 (2006).
17. Abel, S., Oeller, P. W. & Theologis, A. Early auxin-induced genes encode short-lived nuclear proteins. *Proc. Natl Acad. Sci. USA* **91,** 326–330 (1994).
18. Vernoux, T. *et al.* The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Mol. Syst. Biol.* **7,** 508 (2011).
19. Parry, G. *et al.* Complex regulation of the TIR1/AFB family of auxin receptors. *Proc. Natl Acad. Sci. USA* **106,** 22540–22545 (2009).
20. Benková, E. *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115,** 591–602 (2003).
21. Ottenschläger, I. *et al.* Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl Acad. Sci. USA* **100,** 2987–2991 (2003).
22. Santuari, L. *et al.* Positional information by differential endocytosis splits auxin response to drive *Arabidopsis* root meristem growth. *Curr. Biol.* **21,** 1918–1923 (2011).
23. Petersson, S. V. *et al.* An auxin gradient and maximum in the *Arabidopsis* root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis. *Plant Cell* **21,** 1659–1668 (2009).
24. Boonsirichai, K., Guan, C., Chen, R. & Masson, P. H. Root gravitropism: an experimental tool to investigate basic cellular and molecular processes underlying mechanosensing and signal transmission in plants. *Annu. Rev. Plant Biol.* **53,** 421–447 (2002).
25. Boonsirichai, K., Sedbrook, J. C., Chen, R., Gilroy, S. & Masson, P. H. ALTERED RESPONSE TO GRAVITY is a peripheral membrane protein that modulates gravity-induced cytoplasmic alkalinization and lateral auxin transport in plant statocytes. *Plant Cell* **15,** 2612–2625 (2003).
26. Reinhardt, D. *et al.* Regulation of phyllotaxis by polar auxin transport. *Nature* **426,** 255–260 (2003).
27. Heisler, M. G. *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem. *Curr. Biol.* **15,** 1899–1911 (2005).
28. Silverstone, A. L. *et al.* Repressing a repressor: gibberellin-induced rapid reduction of the RGA protein in *Arabidopsis*. *Plant Cell* **13,** 1555–1566 (2001).
29. Fu, X. & Harberd, N. P. Auxin promotes *Arabidopsis* root growth by modulating gibberellin response. *Nature* **421,** 740–743 (2003).
30. Santner, A. & Estelle, M. Recent advances and emerging trends in plant hormone signalling. *Nature* **459,** 1071–1078 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Erktan and C. Cellier for help with marker expression analysis; J. Neve, A. Miyawaki, M. Heisler and M. Estelle for providing the *35S::Flag-TIR1* line, *VENUS* complementary DNA, DR5::VENUS plasmids and TIR/AFB GUS lines, respectively; the PLATIM for access to confocal microscopes; F. Parcy, O. Hamant, A. Boudaoud and P. Das for discussions. T.V. was supported by the Human Frontier Science Program Organization (CDA 0047/2007 HFSPO) and the Agence National de la Recherche (ANR-07-JCJC-0115 and EraSysBio+ iSAM). D.M.W., A.L. and M.J.B. acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC) and Engineering and Physical Sciences Research Council (EPSRC) funding to the Centre for Plant Integrative Biology (CPIB); BBSRC grants BB/F013981/1 and BB/F007418/1 to S.K.; BBSRC Professorial Research Fellowship funding to D.M.W. & M.J.B.; and Belgian Scientific policy (BELSPO contract BARN) to A.L., T.B. and M.J.B.

Author Contributions T.V. designed the DII-VENUS tool. G.B., M.O. and T.V. engineered and characterized DII-VENUS transgenic lines. D.M.W., G.B., A.L. and V.M. quantified the spatial and temporal dynamics of DII-VENUS. A.H.B. did the pull-down assay. T.V. and M.J.B. designed the experiments with the help of T.B., S.K. and J.T. T.V. and M.J.B. analysed the data and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Seed for the lines described in this study have been deposited at the Nottingham *Arabidopsis* Stock Centre. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.V. (teva.vernoux@ens-lyon.fr).

METHODS

Plant material, growth conditions and plant treatments. All transgenic plants were generated in the Columbia ecotype (Col-0). The *tir1 afb1 afb2 afb3* quadruple mutant, DR5::GFP line and TIR1/AFB1–AFB3 GUS translational fusions have been described[4,19,31]. *DR5::VENUS* transgenic plants were generated by transforming a *DR5::VENUS* plasmid[27] by floral dipping[32]. Plants were cultivated *in vitro* on MS medium supplemented with 1% sucrose at 22 °C and under long-day conditions (16 h light/8 h darkness). For analysis on roots, the chemical treatments were done on 5-day-old plants by transferring them to liquid MS supplemented with the chemicals or on an MS agar supplemented with the chemicals for root live imaging. Indole-3-acetic acid (Sigma) or 1-naphthaleneacetic acid (NAA; Sigma) was dissolved in ethanol and used at the indicated concentration. MG132 and clasto-lactacystin-β-lactone (lactacystin; Sigma) were dissolved in dimethylsulphoxide (DMSO) and used at the final concentration of 50 μM for 2.5 h or 20 μM for 8 h respectively. For MG132/indole-3-acetic acid co-treatments, plants were pretreated with MG132 for 1.5 h before adding indole-3-acetic acid. For analysis on the vegetative shoot apex, 5-day-old seedlings were used after removing one cotyledon to allow observation of the vegetative shoot apex. Seedlings were mounted in water and treatments were done by replacing by capillarity the water with indole-3-acetic acid at the indicated concentration. For the inflorescence apex, the plants were transferred to water containing indole-3-acetic acid at the indicated concentration.

Generation of DII-VENUS transgenic plants. The DII-VENUS binary vectors were generated using Gateway technology and following the Multisite Gateway three-fragment vector construction kit protocol (Invitrogen). To generate the different versions of the DII-VENUS sensor (Supplementary Fig. 1), we used the region of IAA8, IAA9 and IAA28 starting from the conserved lysine up to the end of domain II (IAA8, amino acid positions 107–178; IAA9, 120–195; IAA28, 28–61; Supplementary Fig. 1). IAA8, IAA9 and IAA28 were chosen because their basal half-lives were potentially long enough[16] (ranging from 15–20 min for IAA8 and IAA9 to 80 min for IAA28) to allow for the maturation of the fast-maturing YFP variant VENUS and thus for visualization of its fluorescence. We cloned the *IAA8*, *IAA9* and *IAA28* cDNAs by standard RT–PCR from inflorescence mRNA. The different wild-type sequences were then amplified by PCR (see Supplementary Table 1 for primers) and cloned in pDONR 221 by recombination. We then mutated the conserved lysine (K to R mutation) by introducing this mutation in the forward primers (sequence in bold replaced by AGA: see Supplementary Table 1). To generate mDII-VENUS, site-directed mutagenesis (using standard inverted PCR procedures) was used to introduce the P53L mutation in the wild-type IAA28 sequence in pDONR 221 (Supplementary Fig. 1)[33]. The sequence of VENUS fused to the N7 nuclear localization signal[34] was amplified by PCR (see Supplementary Table 1 for primers) from a VENUS-N7 sequence cloned in pBG36 and cloned into pDONR P2R-P3 by recombination. Finally Aux/IAA-derived sequences were fused in-frame to VENUS-N7 (Supplementary Fig. 1) and put under the control of the strong constitutive CaMV 35S promoter, using a 35S promoter cloned in pDONR P4-P1R and recombination into the gateway-compatible pH7m34GW binary vector[35] (hygromycin resistance). The different plasmids were then introduced in plants by floral dipping[32].

Confocal microscopy, live imaging and quantification of fluorescence. Imaging was performed either on either a LSM-510 laser-scanning confocal microscope (Zeiss), a SP5 spectral detection confocal microscope (Leica) or an Eclipse Ti 2000 laser-scanning confocal microscope (Nikon). For visualization of the root organization the roots were stained either with FM4-64 (Invitrogen) as previously described[36] or propidium iodide (Sigma). To quantify fluorescence with cellular resolution in DII-VENUS and mDII-VENUS root meristems and generate maps of relative auxin using DII-VENUS, serial optical sections were obtained. A cellular grid was generated from the propidium iodide channel of the most median optical section using Merrysim[37]. For each cell defined in the grid, we then selected the optical section passing through the centre of each nuclei. Fluorescence was then summed inside the corresponding cell from that section. The loss of fluorescence due to tissue absorbance was also estimated using the spatial distribution of the propidium iodide channel and used to correct the fluorescence values. To obtain the changes in auxin levels along the root axis in the different tissues, fluorescence distribution was extracted and curvatures were smoothed using a Gaussian kernel with a sigma value of 5.

For root live imaging, immediately after the beginning of the treatment, the seedlings were scanned every 2 min for 2 h to follow the evolution of the DII-VENUS signal. To quantify fluorescence in the root tip, the average fluorescence intensity over identical scanned portion of the root (corresponding approximately to the first 200 μM from the root tip) was extracted using EZ-C1 software (v3.9, Nikon) and the values analysed using Microsoft Excel. For tissue specific kinetics, the seedlings were scanned every 5 min and fluorescence was extracted from groups of nuclei with nearly identical fluorescence (variations < 30%) in the different tissues. The fluorescence intensity of nuclei was extracted from the different tissues using the ROI tool of Fiji software (http://fiji.sc/wiki/index.php/Fiji) and the values analysed using Microsoft Excel. For dose-dependent quantification of DII-VENUS signal upon auxin treatment, fluorescence was measured 1 h after treatment using two and three roots for wild-type and *tir1 afb1 afb2 afb3*, respectively.

For the vegetative shoot apex, seedlings were mounted into water in between slide and cover slip. For live imaging of vegetative shoot apex, five plants were treated after observation at $t = 0$ and followed over 90 min. For the inflorescence apex, observation was performed as described[38]. Live imaging of the inflorescence apex was performed on plants grown on the auxin transport inhibitor NPA then transferred to a new medium without NPA as previously described[39]. Initiation of a new organ was confirmed a posteriori by visual inspection of the apex.

Root growth analysis, gravitropic assays and flower production rate. For analysis of root growth, plants were grown near-vertically and root length was measured at the indicated time. For gravitropic assays, plants were grown as detailed previously[40] and imaged at 30 min intervals following a 90° gravi-stimulus. Root tip angle was measured using modified RootTrace software[41] (http://www.cpib.ac.uk/tools-resources/roottrace). Estimation of flower production rate was done as described[42].

IAA28–Flag–TIR1 pull-down assays. To generate the *35S::Flag-TIR1* transgenic *Arabidopsis* line a plant expression vector containing a 3×Flag was first created by annealing complementary 101-base-pairs oligonucleotides including the 3×Flag coding sequences (see Supplementary Table 1) and cloning this fragment into XbaI and SalI sites of the vector pFP101. The Gateway C1 cassette (Invitrogen Gateway vector conversion reagent system) was then introduced into this plasmid by blunt-end ligation following SalI digestion and end-filling by Klenow reaction to create the destination vector pFP3FLAGSII. The full-length coding sequence for TIR1 was amplified from an *Arabidopsis* cDNA library using Gateway-compatible primers (see Supplementary Table 1) and incorporated into the Gateway donor vector pDONR207 by BP reaction (Invitrogen). The TIR1 coding sequence was then incorporated into pFP3FLAGSII via a Gateway vector. LR reaction was used to form plasmid pFP3FLAGSII-TIR1. Wild-type *Arabidopsis* plants were subsequently transformed using the floral dip method[32] and homozygous lines with single-site were selected from the T_3 generation. Extracts of 10-day-old *35S::Flag-TIR1* seedlings were made as described previously[6] and used in pull-down assays by combining 2.5 mg of crude extract with 5 μg of biotinylated IAA28 domain II peptide (biotinyl-NH-EVAPVVGWPPVRSSRRN-COOH, synthesized by Thermo Scientific), and 65 μl 50% streptavidin-agarose suspension. The assays were incubated for 1 h at 4 °C with mixing then washed three times for 5 min in extraction buffer (0.15 M NaCl, 0.5% Nonidet P40, 0.1 M Tris-HCl pH 7.5, containing 1 mM phenylmethylsulphonyl fluoride, 1 μM dithiothreitol, 10 μM MG132) containing the appropriate auxin treatment. The final processing of the pull-down assays including electrophoresis and western transfer were performed as described previously[6]. The immunodetection of TIR1/AFB–Flag was performed with a 1:5,000 dilution of anti-Flag 2-peroxidase (HRP) antibody (Sigma) followed by chemiluminescent detection with ECL plus reagents (Amersham).

Transcript profiling. Total RNA was extracted from roots using an RNeasy Plant Micro Kit (Qiagen), including on-column DNase digestion to eliminate genomic DNA from the samples. A 500-ng aliquot of RNA was reverse-transcribed using a Transcriptor First Strand cDNA synthesis kit (Roche) and anchored-oligo (dT)18. Real-time qPCRs were performed on a Roche Light Cycler 480 system using the prevalidated single hydrolysis probes, Sensimix probe master mix (Quantace), and gene-specific primers (see Supplementary Table 1).

31. Friml, J. *et al.* Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature* **426,** 147–153 (2003).
32. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743 (1998).
33. Rogg, L. E., Lasswell, J. & Bartel, B. A gain-of-function mutation in IAA28 suppresses lateral root development. *Plant Cell* **13,** 465–480 (2001).
34. Cutler, S. R., Ehrhardt, D. W., Griffitts, J. S. & Somerville, C. R. Random GFP:cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at a high frequency. *Proc. Natl Acad. Sci. USA* **97,** 3718–3723 (2000).
35. Karimi, M., De Meyer, B. & Hilson, P. Modular cloning in plant cells. *Trends Plant Sci.* **10,** 103–105 (2005).
36. Levesque, M. P. *et al.* Whole-genome analysis of the SHORT-ROOT developmental pathway in *Arabidopsis*. *PLoS Biol.* **4,** e143 (2006).
37. Barbier de Reuille, P., Bohn-Courseau, I., Godin, C. & Traas, J. A protocol to analyse cellular dynamics during plant development. *Plant J.* **44,** 1045–1053 (2005).
38. Fernandez, R. *et al.* Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution. *Nature Methods* **7,** 547–553 (2010).
39. Grandjean, O. *et al.* *In vivo* analysis of cell division, cell growth, and differentiation at the shoot apical meristem in *Arabidopsis*. *Plant Cell* **16,** 74–87 (2004).
40. Holman, T. J. *et al.* Statistical evaluation of transcriptomic data generated using the Affymetrix one-cycle, two-cycle and IVT-Express RNA labelling protocols with the *Arabidopsis* ATH1 microarray. *Plant Methods* **6,** 9 (2010).

41. French, A., Ubeda-Tomas, S., Holman, T. J., Bennett, M. J. & Pridmore, T. High-throughput quantification of root growth using a novel image-analysis tool. *Plant Physiol.* **150,** 1784–1795 (2009).

42. Lohmann, D. *et al.* SLOW MOTION is required for within-plant auxin homeostasis and normal timing of lateral organ initiation at the shoot meristem in *Arabidopsis*. *Plant Cell* **22,** 335–348 (2010).

MATERIALS SCIENCE

A fresh twist for self-assembly

Molecular helicity affects many of the bulk properties of materials. A study finds that helicity also controls the self-assembly of colloidal particles, opening the door to a new generation of functional materials. SEE LETTER P.348

VOLKER SCHALLER
& ANDREAS R. BAUSCH

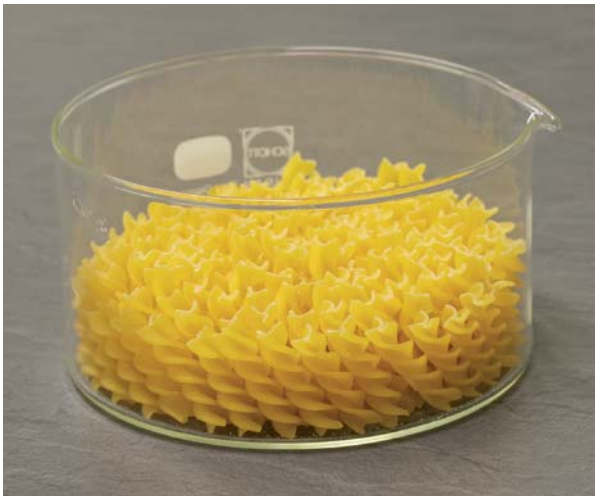
The next time you go to the supermarket, take a look at the pasta. You'll probably find everything from long, thin spaghetti to butterfly-shaped farfalle and twisted fusilli. On closer inspection, you'll see that the strands of spaghetti readily align and pack closely together, whereas the packing of the fusilli is considerably more complex. This complexity is due to the fusilli's chirality — its helical geometry. On page 348 of this issue, Gibaud *et al.*[1] report that such complexity of packing can be exploited to control the self-assembly of nanometre-scale particles, allowing the reversible formation of various architectures*.

To closely pack two individual pieces of fusilli, the pasta pieces have to twist with respect to each other so that their long axes are not parallel — this is an effect of their chirality. But when many fusilli are crammed together, this twist hinders the ability of the pasta to align in one direction, as required for efficient packing. The resulting competition between chirality and packing can have astonishing consequences for geometrically constrained chiral objects: if you stack fusilli upright in a beaker, the axes of the fusilli in the centre are perfectly vertical, whereas those at the edges of the beaker twist away from this alignment (Fig. 1).

This balance between geometric constraints and chiral interactions is at the heart of the unique properties of chiral liquid crystals. In 'blue' liquid-crystal phases, for instance, chiral molecules self-assemble into cylinders of twisted molecules that stack in a cubic lattice, giving rise to vividly coloured and reflective materials[2]. And in smectic liquid crystals, chiral interactions lead to a twisted, layered structure that is also manifested in the optical and mechanical properties of the materials[3].

Gibaud *et al.*[1] examined the balance between chirality and geometric constraints in the self-assembly of colloidal particles. Their model system was an ensemble of rod-like fd viruses, 7 nanometres in diameter and 880 nm in length, suspended in water. The chirality of these viruses depends on the ambient temperature: the rods are achiral at high temperatures, but chiral at low temperatures. The authors observed that, as for any other suspension of colloidal particles[4], the addition of an appropriate polymer that does not adsorb to the rods drove a phase separation of the mixture into a polymer-rich and a colloid-rich phase. At high temperatures, at which the rods are achiral, this led to the self-assembly of a circular colloidal membrane (a monolayer of vertically aligned rods[5]).

The authors observed that the rods at the centre of the membrane readily align parallel to each other, but at the perimeter of the membrane they twist gradually away from the vertical. This twist at the margin minimizes the interface between the polymer-rich and the rod-rich phases, and so reduces the interfacial tension that arises from the imbalance of forces between particles in this region. But twisting rods out of alignment with the other rods in the membrane has its own energetic cost, known as elastic energy. The formation of circular membranes with twisted margins is therefore the result of a trade-off between minimizing the phase interface and minimizing the elastic energy.



V. SCHALLER

Figure 1 | Pasta packing. When constrained in a circular container, fusilli pasta pieces mostly pack together so that their long axes are vertically aligned. But at the edges of the container, the pasta twists away from this alignment. This packing arrangement is a consequence of the pasta's helicity (chirality). Gibaud *et al.*[1] report that the chirality of colloidal particles affects the self-assembly of those particles.

But what happens if the rods are chiral, so that twisted packing is preferred — just as it is for closely packed fusilli? Gibaud *et al.* addressed this question by performing experiments at lower temperatures, thereby 'switching on' the chirality of the viruses. They observed that increases in chirality — that is, increases in the contribution of chiral interactions to the energy balance of the system — reduce the elastic energy, thus lowering the energetic cost of creating a twist at the membrane's margin. This destabilizes the edges of the circular membrane and triggers the formation of ribbon-like structures that splay out from the circular membrane (see Fig. 3a of the paper[1]). The effect of increasing the chirality of the viruses in the colloidal membrane is therefore similar to that of adding surfactants to oil–water mixtures: it decreases the interfacial tension and so allows more interfaces to form.

The beauty of the authors' experimental set-up is that it allows all the parameters involved in the self-assembly of chiral colloids to be readily teased out and measured. Another advantage of the system is that the colloidal building blocks are much bigger than the molecules typically used in studies of chiral materials. This allowed Gibaud *et al.* to precisely quantify the colloidal self-assembly process at all scales, from the microscopic movements of individual viruses to the macroscopic properties of the resulting membrane. It also allowed them to exert well-defined local forces on the system, to affect the balance between twist and interfacial energy, and so to induce structural transitions — for example, by pulling on the membrane, they converted it into ribbon structures.

By including chiral interactions in their system, Gibaud *et al.* have added a new degree of freedom to the self-assembly of colloidal particles. Transferring their results to other

*This article and the paper[1] under discussion were published online on 4 January 2012.

materials opens the door to the hierarchical assembly of functional materials that could react sensitively to changes in ambient conditions or to mechanical stimuli. One way to extend the functionality of their system could be to make colloidal particles that are not only chiral, but also have surfaces patterned with microscopic domains that attract or repel each other. Such patterned colloidal particles have recently been shown to self-assemble into useful functional structures such as 'kagome' lattices[6].

Chiral interactions between particles occur in all kinds of materials, from liquid crystals[2,3] to cytoskeletal filaments in cells (for which the hierarchical assembly of bundle-like structures is dependent on, and sensitive to, the helical twist of the filaments[7,8]). Most of the chiral materials studied so far are passive or in thermal equilibrium. This means that their structural assembly is governed only by diffusion and by local interactions between the constituent particles. But most naturally occurring materials are far from passive. Instead, they constantly consume energy so that their particles self-organize into higher-order structures — the cytoskeletons of cells are prime examples of this[9,10].

The next step, therefore, is to apply the principles identified by Gibaud and colleagues[1] to active systems. Imagine replacing the viruses used in this study with self-propelling chiral bacteria — would they still pack together in a two-dimensional membrane, and, if so, would the membrane move, or rotate, across mesoscopic or macroscopic distances? Could actively beating ribbons form, or would the noise stemming from active movement of the bacteria prevent the formation of such structures? All we can say for certain is that chiral interactions will add a new twist not only to the self-assembly of colloids, but also to the self-organization of active materials. ■

Volker Schaller *and* **Andreas R. Bausch** *are at the Lehrstuhl für Biophysik-E27, Technische Universität München, Garching 85748, Germany.*
e-mail: abausch@mytum.de

1. Gibaud, T. *et al. Nature* **481,** 348–351 (2012).
2. Coles, H. J. & Pivnenko, M. N. *Nature* **436,** 997–1000 (2005).
3. Hough, L. E. *et al. Science* **325,** 456–460 (2009).
4. Aarts, D. G. A. L., Schmidt, M. & Lekkerkerker, H. N. W. *Science* **304,** 847–850 (2004).
5. Barry, E. & Dogic, Z. *Proc. Natl Acad. Sci. USA* **107,** 10348–10353 (2010).
6. Chen, Q., Bae, S. C. & Granick, S. *Nature* **469,** 381–384 (2011).
7. Claessens, M. M. A. E., Semmrich, C., Ramos, L. & Bausch, A. R. *Proc. Natl Acad. Sci. USA* **105,** 8819–8822 (2008).
8. Grason, G. M. & Bruinsma, R. F. *Phys. Rev. Lett.* **99,** 098101 (2007).
9. Kasza, K. E. *et al. Curr. Opin. Cell Biol.* **19,** 101–107 (2007).
10. Köhler, S., Schaller, V. & Bausch, A. R. *Nature Mater.* **10,** 462–468 (2011).

GENOMICS

The path to retinoblastoma

Genomic analyses of tumours of the childhood cancer retinoblastoma reveal a low mutation rate, challenging the view that genomic instability is crucial for its progression. The work also identifies a new therapeutic target. SEE ARTICLE P.329

JULIEN SAGE & MICHAEL L. CLEARY

Retinoblastoma is a rare tumour that affects retinal cells in the eyes of children. Analyses of familial and sporadic cases of this cancer, backed by studies in genetically engineered mice, have shown that loss of function of the tumour-suppressor protein RB1 (also known as RB) is required for the development of most, if not all, tumours of this type. However, it is not clear how RB1-deficient retinal cells progress to malignant tumour cells[1]. In addition, emerging evidence that loss of RB1 function can induce genomic instability[2] has raised the tantalizing possibility that RB1-deficient retinal cells might be predisposed to accumulating many additional mutations, further complicating the identification of mutations that contribute to the development and maintenance of retinoblastoma. On page 329 of this issue, however, Zhang *et al.*[3] demonstrate that retinoblastoma genomes have very few recurrent mutations in genes other than *RB1*. Instead, the expression of cancer-related genes is affected by epigenetic modifications on chromosomes, which do not affect DNA sequence but are inherited after cell division.

To identify mutations that could cooperate with loss of RB1 function in tumour development, Zhang and colleagues[3] sequenced and compared the genomes of normal tissue and retinoblastoma tumours from four patients. The researchers found that *RB1* was the only known cancer-related gene consistently mutated, and that the retinoblastomas had 15-fold fewer total mutations than other types of solid tumour whose genomes have been sequenced[4].

Next, the authors searched for epigenetic alterations and for abnormal gene expression in retinoblastoma cells. They identified the gene that encodes the protein kinase enzyme SYK as a potential oncogene whose expression is consistently higher in retinoblastoma cells than in normal immature retinal cells. Moreover, the activity of SYK was essential for the growth of retinoblastoma cells. The authors also show that certain small molecules that selectively inhibit SYK activity induce the death of retinoblastoma cells in a mouse model of the disease.

These findings[3] indicate that SYK may be a promising target for treating patients with retinoblastoma. *SYK* was not identified in previous searches for genes with a role in this cancer because it is not mutated or structurally rearranged in retinoblastoma. Therefore, Zhang and colleagues' study emphasizes the importance of high-throughput approaches that integrate genome sequencing with gene-expression analysis and epigenomics to identify cancer genes.

Low mutation frequency has also been observed[5] in medulloblastoma (a type of brain tumour that affects children), suggesting that it could be a general feature of childhood cancers. A possible explanation for this difference between the cancers of children and adults is that childhood tumours arise in cells that are naturally undergoing rapid developmental growth, with fewer brakes on their proliferation than cells in adults. An alternative explanation is that, in children, these cells are negotiating crucial developmental checkpoints that are susceptible to corruption, leading to incomplete or abnormal maturation. In both cases, only a few mutations would be needed to trigger the cellular changes associated with cancer.

Furthermore, epigenetic changes in children and excessive mutations in adults may have similar roles in cancer development. Another childhood cancer, Wilms' tumour, also has a relatively stable genome and displays an increased variation in DNA-methylation patterns compared with normal cells[6]. RB1-deficient retinal cells may be particularly susceptible to this tumour-formation mechanism, because RB1 interacts with the machinery that controls the epigenetic status of cells, including enzymes that remodel chromatin (DNA–protein complexes) and other enzymes that add methyl groups to DNA[1]. Thus, childhood cancers such as retinoblastoma may carry epigenetic abnormalities that change gene expression and are stably propagated through subsequent cell divisions, helping to maintain tumour-associated features.

If this model is correct, it is possible that RB1-deficient tumours in adults — such as small-cell lung cancer and some breast,

ARTICLE

doi:10.1038/nature10799

Genome-wide structure and organization of eukaryotic pre-initiation complexes

Ho Sung Rhee[1] & B. Franklin Pugh[1]

Transcription and regulation of genes originate from transcription pre-initiation complexes (PICs). Their structural and positional organization across eukaryotic genomes is unknown. Here we applied lambda exonuclease to chromatin immunoprecipitates (termed ChIP-exo) to examine the precise location of 6,045 PICs in *Saccharomyces*. PICs, including RNA polymerase II and protein complexes TFIIA, TFIIB, TFIID (or TBP), TFIIE, TFIIF, TFIIH and TFIIK were positioned within promoters and excluded from coding regions. Exonuclease patterns were in agreement with crystallographic models of the PIC, and were sufficiently precise to identify TATA-like elements at so-called TATA-less promoters. These PICs and their transcription start sites were positionally constrained at TFIID-engaged downstream +1 nucleosomes. At TATA-box-containing promoters, which are depleted of TFIID, a +1 nucleosome was positioned to be in competition with the PIC, which may allow greater latitude in start-site selection. Our genomic localization of messenger RNA and non-coding RNA PICs reveals that two PICs, in inverted orientation, may occupy the flanking borders of nucleosome-free regions. Their unambiguous detection may help distinguish bona fide genes from transcriptional noise.

Assembly of the PIC and its post-assembly control are critical early steps in the transcription of eukaryotic genes. TBP (TATA-binding protein) arrives at most promoters as part of the multi-subunit TFIID complex that includes TBP-associated factors (TAFs)[1]. Together these proteins help recruit RNA polymerase (Pol) II and its entourage of general transcription factors (GTFs) to the transcription start sites (TSSs) of genes[2–4]. These PICs assemble in nucleosome-free promoter regions (NFRs) that are flanked by an upstream −1 nucleosome and a downstream +1 nucleosome[5]. PICs have largely been defined biochemically using purified GTFs at a few model genes[2–4], but little is known about their assembly and organization *in vivo*, particularly at near-base-pair resolution on a genome-wide scale.

An oddity of TBP is that when it is part of the TFIID complex, it tends to bind promoters that lack the TATA box consensus TATAWAWR (W indicates A/T; R indicates A/G)[6]. Approximately 80–90% of all *Saccharomyces* genes are thus designated as 'TATA-less', and have a predominant PIC assembly mechanism and chromatin architecture that differs substantially from those in the 'TATA box' class of genes[6–9]. So far, no TBP-binding motif has been identified at TATA-less promoters, and so the origins of TFIID-promoter specificity have been rather enigmatic[10]. When TBP is not part of the TFIID complex, the SAGA complex directs TBP to TATA-box-containing Pol II promoters[11–13].

TFIIA and TFIIB clamp TBP to DNA, and make DNA contacts immediately upstream and downstream of the TATA box. TFIIB is a linchpin between TBP and Pol II[14,15]. Its intimate contact with Pol II directs how far downstream Pol II productively initiates transcription[16,17]. TFIIF enhances the interaction of Pol II with TFIIB, assists in recruiting TFIIE, and promotes downstream elongation events[3,18]. TFIIE then stimulates DNA strand separation by Pol II at the TSS, and enhances the activity of TFIIH. TFIIH holoenzyme is multi-functional, having ATP-dependent helicase (Ssl2 and Rad3) and kinase (Kin28) activities that reside on biochemically separable sub-complexes (TFIIH and TFIIK, respectively), both of which are key to efficient open complex formation and transcription initiation[19–21].

We examined the structural organization of PICs and their specificity on a genomic scale using ChIP-exo[22]. This novel strategy substantially improved mapping resolution and eliminated many false positives. The exonuclease processively degrades a DNA strand in the 5′–3′ direction until a crosslinking point is encountered (Supplementary Fig. 1a). The crosslinking inefficiency inherent to ChIP allows multiple crosslinking points to be detected in a population by deep sequencing. When applied to the GTFs on a genomic scale, we obtained detailed and comprehensive information on PIC structure and genomic organization.

Genome-wide PIC structure

We applied ChIP-exo genome-wide to Pol II and each GTF (Fig. 1a), and verified binding for TFIIB by locus-specific PCR using a series of tiled primers (Supplementary Fig. 1b). When exonuclease stop sites were mapped over all annotated mRNA promoters that contained a TATA box consensus, a distinctive pattern was observed (Fig. 1b). Importantly, each GTF displayed a strand-specific composite pattern of exonuclease stop sites that occurred only when TATA boxes, but not TSSs (Supplementary Fig. 2a and data not shown), were aligned, indicating that PICs are positioned with respect to the TATA box.

For TFIIB, we detected four DNA crosslinking points (pairs of exonuclease stops), designated B1–B4 (Fig. 1c). The TATA box was precisely centred between B1 and B2, which were separated by 20 ± 3 base pairs (bp). Crosslinking point B3 and the diffuse B4 region indicate that TFIIB further crosslinked over a broad region downstream of the TATA box towards the TSS. We compared the four TFIIB crosslinking points to crystallographic-based models of open and closed TBP–TFIIB–Pol-II–promoter complexes (Fig. 1c)[14,15]. An important caveat of the crystallographic models is they were built from multiple independent structures of truncated TFIIB–TBP–TATA, TFIIB–Pol II and Pol II elongation complexes. Thus, the combined structures represent a hypothetical organization.

Within the modelled closed and open structures, crosslinking site B1 precisely (± 3 bp) mapped to where the TFIIB carboxy-terminal core straddles the upstream DNA-binding stirrup of TBP (11 bp

[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

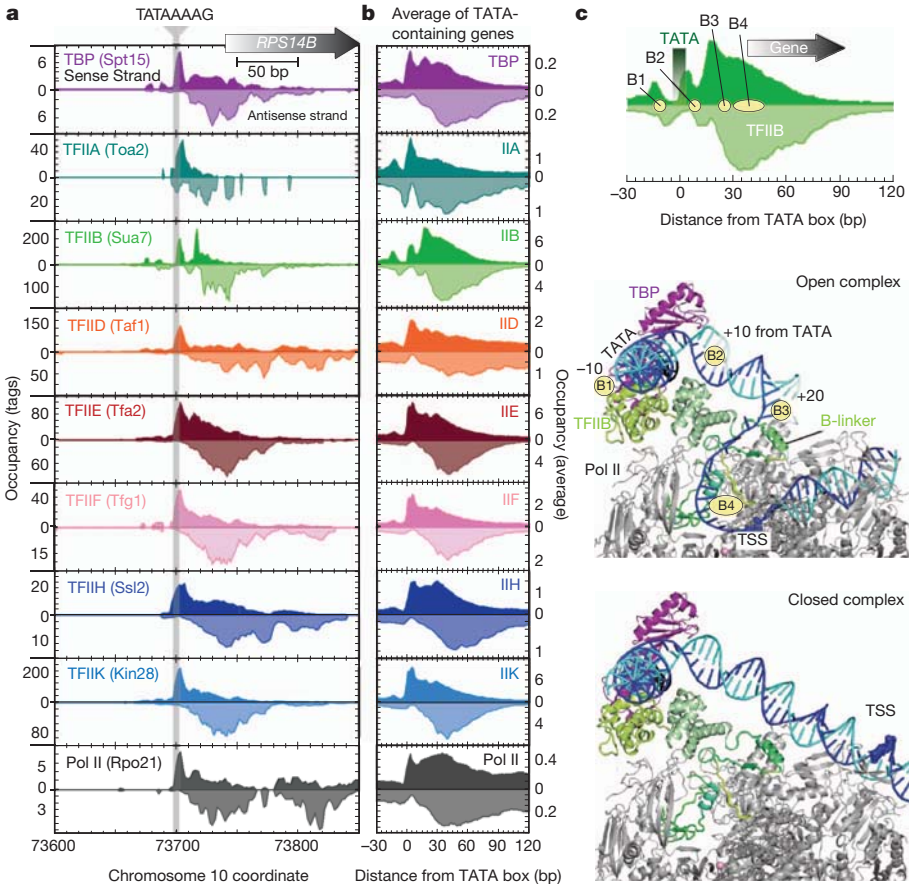


Figure 1 | Genome-wide structural organization of PICs. **a**, Raw ChIP-exo tag distribution for GTFs and Pol II around the *RPS14B* gene. Filled plots represent unfiltered 5′ ends of sequencing tags on the sense (darker fill) and antisense strand (lighter fill). **b**, Average GTF and Pol II occupancy around the TATA box of 676 annotated mRNA genes. Plots are as in panel **a**. **c**, Relationship of four TFIIB crosslinking points to crystallographic-based models of the PIC[14]. The top panel is expanded from panel **b** for TFIIB. The middle and bottom panels show modelled open and closed TBP–TFIIB–Pol-II–promoter DNA complexes, respectively.

upstream of the TATA box midpoint). B2 mapped precisely (± 3 bp) to where the TFIIB core amino-terminal cyclin fold encounters DNA just downstream of TBP's other stirrup (9 bp downstream of the TATA box midpoint). B3 mapped to where the TFIIB linker region is closest to the DNA, which was 19 bp downstream of the TATA box midpoint. B4 corresponded to a broad region defined by close proximity of the TFIIB reader (or finger) domain to single-stranded DNA within the modelled open complex, but was not evident within the closed complex. Similar broad regions of crosslinking were observed with the other GTFs (Fig. 1b), and may reflect indirect crosslinking. Support that these PICs represent open complexes is provided by permanganate reactivity studies of the *GAL1*, *GAL10* and *HSP82* loci[23,24]. Taken together, the entirety of the genomic crosslinking sites observed with the GTFs and Pol II fits remarkably well with the crystallographic models of the PIC open complex[14,15], and with many aspects of *in vitro* chemical crosslinking of these proteins[19,25–27].

TATA-like elements at TATA-less genes

An apparent paradox of so-called TATA-less promoters is their utilization of TFIID, which has long been described as the TATA-box-binding complex[4]. However, it is unclear whether the TBP subunit of TFIID recognizes specific DNA sequences at TATA-less promoters. Inasmuch as TBP is expected to be found at all TATA-less promoters, and motif searching algorithms failed to identify candidate TBP-binding sites, we instead opted to search for sequence elements with up to two mismatches to the TATAWAWR consensus. We also limited our search to measured PIC locations. Remarkably, 99% of the PICs at TATA-less promoters contained a sequence having two or less mismatches to the TATA box consensus (Fig. 2). We refer to these mismatched elements as

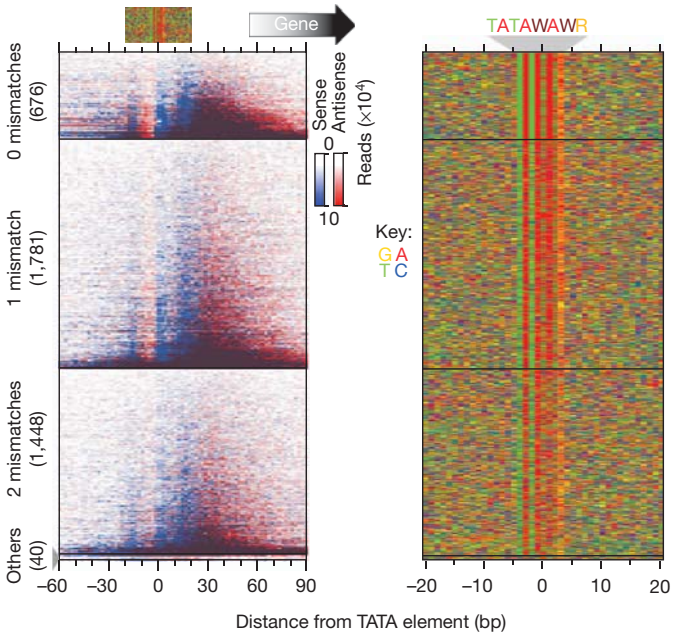


Figure 2 | Identification of TATA-like elements at TATA-less genes. Left, TFIIB occupancy around individual TFIIB-enriched TATA elements of mRNA genes (rows, $n = 3{,}945$), sorted by occupancy level. Occupancy on the sense (blue) and antisense (red) strands is shown with respect to TSS orientation. Right, a colour chart representation of the DNA sequence located ± 20 bp from the TATA element midpoint and ordered as shown in the left panel.

'TATA-like', as they did not form a consensus, whereas those conforming to the consensus retain the 'TATA box' designation. We refer to the two elements together as 'TATA elements'.

To assess whether TFIIB was positioned around these TATA-like elements in a canonical manner as seen at bona fide TATA boxes, strand-specific ChIP-exo tags were plotted around each element, separated into panels by 0, 1 or 2 mismatches to the TATA box consensus (Fig. 2). Notably, regardless of its occupancy level, the distribution of TFIIB crosslinking and thus its positioning relative to these TATA-like elements was quite similar to the positioning observed at bona-fide TATA boxes. When the other GTFs were examined, their patterns relative to TATA-like elements were also similar to those found at TATA boxes (Supplementary Fig. 2b), although some downstream differences were observed (addressed later). Thus, as previously seen at the three yeast TATA-box-containing genes *GAL1*, *GAL10* and *HSP82* (refs 23, 24), and in mammalian *in vitro* systems[28], at least the upstream portion of most PICs are positioned with respect to resident TATA elements nearly identically, regardless of their Pol II promoter classification as TATA-box containing or TATA-less. Although the 'TATA-less' designation may be a misnomer, this class of genes is not simply a slight variation of the TATA class, but instead has predominant regulation by TFIID versus SAGA, positive versus negative regulation by chromatin, and lower plasticity of expression[6–9].

TATA-less TSS positioning by nucleosomes

Permanganate reactivity experiments detect open complex formation upstream of the TSS at the *GAL1*, *GAL10* and *HSP82* TATA-box-containing promoters[23,24]. These and other studies[17,29] have led to the notion that Pol II scans downstream from the open complex to find the TSS. In agreement with this being a general mechanism, we find that PICs of TATA-box-containing genes generally reside upstream of the TSS (Supplementary Fig. 2a).

TAFs are largely depleted at TATA-box-containing promoters, although they are not entirely absent (Fig. 3a). The low level of Taf1 that was present tended to be positioned similarly to TBP and other GTFs (Fig. 1b and Supplementary Fig. 2a, b). In contrast, Taf1-enriched/TATA-less promoters (which are related, as shown in Supplementary Fig. 3) showed additional interactions downstream of the TSS that exactly coincided with the size and location of the +1 nucleosome (Fig. 3a). Indeed, Taf1 displayed a more uniform positioning pattern in relation to the TSS and +1 nucleosome than to TATA elements, which suggests that at least part of the TFIID TAF complex engages and is positioned by the +1 nucleosome at TATA-less promoters. Consistent with this, Bdf1, which is considered to be a missing piece of Taf1 (ref. 30), binds to the +1 nucleosome[31]. Furthermore, Bdf1 showed a nearly identical ChIP-exo pattern to that of Taf1 (Supplementary Fig. 4). TFIID–nucleosomal interactions have also been reported in mammalian systems, although the details may differ[32].

If TFIID binds simultaneously to both the +1 nucleosome and a TATA element, then the intervening Pol II would seem to be fenced in by TFIID, thereby limiting its ability to scan DNA. This model predicts that the TSS would reside closer to the TATA element and be positionally restricted relative to the +1 nucleosome, compared to Taf1-depleted/TATA-box-containing promoters. Indeed, the TSS at TATA-less promoters resided ~10–20 bp closer to the TATA element than at TATA-box-containing promoters (Supplementary Fig. 2a).

We also compared the position of TATA elements and TSSs in relation to the +1 nucleosome. We separately examined individual Taf1-depleted/TATA-box-containing and Taf1-enriched/TATA-less promoters (Fig. 3b). Strikingly, at the Taf1-enriched promoters, the TSS was tightly positioned at the border between the 5′ NFR and the +1 nucleosome in comparison to Taf1-depleted/TATA-box-containing promoters. The latter had TSSs distributed across the adjacent nucleosome position, and these nucleosomes were relatively depleted compared to the Taf1-enriched class (Fig. 3a).

Taken together, we interpret these observations as reflecting distinct functions of the +1 nucleosome at the two classes of genes. Nucleosomes and PICs of the TFIID-enriched/TATA-less class might cooperatively assembly, in which case the nucleosome may be instructive for TSS selection by impeding Pol II scanning (Fig. 3c). In contrast, nucleosomes and PICs of the Taf1-depleted/TATA-box-containing class may competitively assemble. This may allow for the greater stochasticity or plasticity of expression that is characteristic of this class[9], in which nucleosome loss would prime the gene for a high level of transcription. A nucleosome competition mechanism removes an impediment to Pol II scanning. Pol II could scan further, thereby allowing productive initiation at specific DNA elements[33]. The transition into a scanning state may be rate limiting in the scanning cycle, as the PIC is detected upstream of the TSS.

GTF depletion in genes and their termini

Although it is clear that GTFs assemble in promoter regions of mRNA genes and disengage Pol II within ~150 bp of the TSS[34–37], it is less

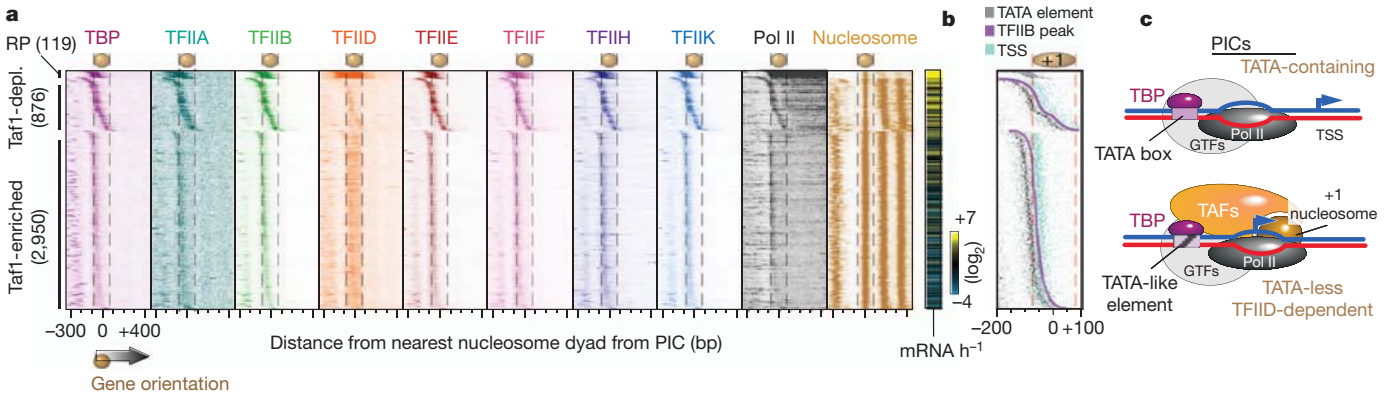


Figure 3 | PIC organization in relation to TFIID and the +1 nucleosome. **a**, GTF occupancy around the nearest nucleosome position (essentially +1) to an mRNA PIC, which were sorted by the distance between the two. Unfiltered tags on each strand were shifted in the 3′ direction by a fixed distance (~8 bp depending on each GTF, 73 bp for nucleosomes), so as to reflect better the points of crosslinking. Taf1-depleted (Taf1-depl.) and Taf1-enriched genes were determined as being distinct clusters when GTF occupancies of all genes were clustered by *k*-means (see Fig. 5a). For all graphs of this type, image resolution is less than the number of rows, resulting in some averaging and thus the appearance of less variance across adjacent rows. See Supplementary Data 2 for underlying values, which can be visualized in Treeview. RP indicates ribosomal protein genes. The right panel shows transcription frequency[48]. The nucleosome borders are denoted by vertical dashed black lines. **b**, Same as panel **a**, but showing an overlay of TATA elements, TFIIB and TSS. **c**, Model of PIC organization at TATA-box-containing and TATA-less/TFIID-dependent genes.

clear to what extent they assemble across the body of genes, or at genes that are either transcriptionally silent or classified as 'dubious'. A whole genome view of GTFs and Pol II is presented in Fig. 4a. Remarkably, PICs were almost entirely excluded from coding regions, regardless of gene activity, whereas Pol II was enriched across gene bodies, as expected. Approximately 90% of dubious open reading frames (ORFs) lacked a canonical PIC organization or contained PICs within the ORF, and thus are unlikely to be coding. Thus, coding region PIC-driven initiation, whether in the sense or antisense direction, is infrequent on the scale of what is seen at mRNA promoters. Moreover, the observed GTF pattern makes clear that Pol II disengages all GTFs at the promoter.

Much less is known of the fate of Pol II at the ends of genes, as it undergoes termination. To examine the 3′ ends of genes, without the complications associated with nearby mRNA promoters, we separated 3′ ends into those having nearby 3′ or 5′ ends of an adjacent gene (Fig. 4b). Within terminal intergenic regions, GTFs were highly depleted, indicating that PICs rarely exist at the 3′ ends of genes at levels seen for mRNA genes (although lower levels do exist).

Remarkably, we find a highly correlated enrichment of Pol II and TFIIH (Ssl2) but not TFIIK (Kin28), at the end of genes within 3′ NFRs (Fig. 4b). Such a physical separation of the TFIIH/Ssl2 and TFIIK/Kin28 submodules of holo-TFIIH in genome-wide binding experiments has not previously been reported, but may be in accord with their biochemistry[19–21]. However, Ssl2 is biochemically separable from the TFIIH core, which therefore prompted us to examine additional core TFIIH subunits, Ssl1 and Tfb1. Surprisingly, both were absent from the ends of genes (Supplementary Fig. 6), although they were present within PICs at promoters. These results suggest that Ssl2, a 3′–5′ helicase, operates independently of holo-TFIIH at the ends of genes. Consistent with a possible role of Ssl2 in transcription termination, Ssl2 has functional interactions with the Hsp90 protein chaperone[38], which has been implicated in the disassembly of the transcription machinery[39].

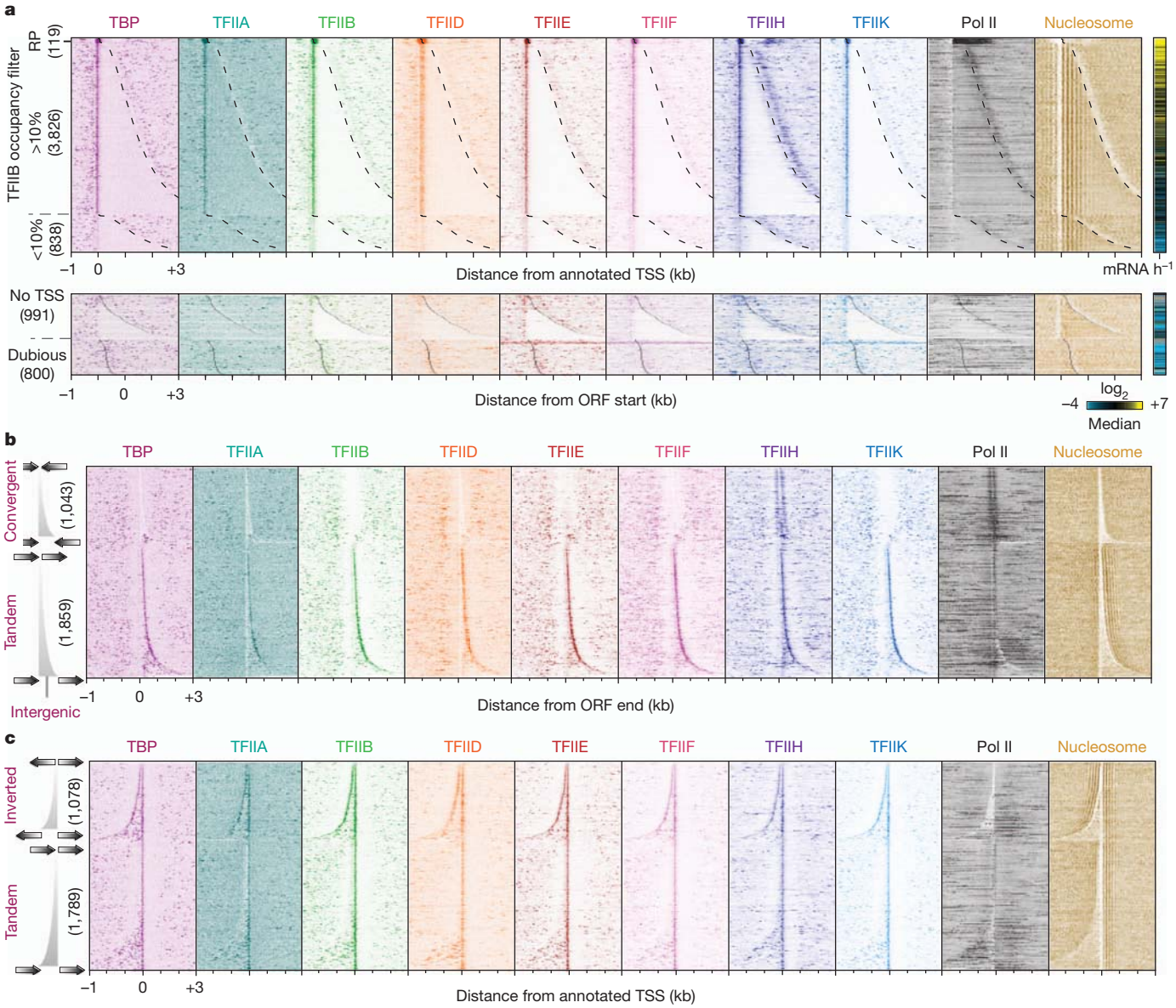


Figure 4 | Genomic view of PICs in relation to genes. **a**, GTF occupancy around transcript and ORF start sites[47], sorted by gene length. See Supplementary Data 3–5 for underlying values. Transcript or ORF ends are indicated by black dashed and solid lines, respectively. The right panel shows transcription frequency[48]. **b**, GTF distribution around the 3′ ends of mRNA genes, sorted by intergenic length, and sectioned by convergent versus tandem gene-pairs. Occupancy at eight reported looped genes[49,50] is shown in Supplementary Fig. 5. **c**, GTF distribution around the TSS of mRNA genes, sorted by intergenic length, and sectioned by inverted versus tandem gene-pairs.

Divergent transcription from distinct PICs

In contrast to coding regions, PICs were abundant in intergenic regions, far beyond what could be accounted for at mRNA promoters (Fig. 4c). Divergent transcription, in which mRNA and non-coding (nc)RNA initiation occurs within the same region but elongates in opposite directions, is well established in eukaryotes[40,41]. However, it has been unclear whether divergent transcription originates from the same PIC site. Conceivably, the entire PIC or a portion thereof might assemble in either direction. As shown in Fig. 4c, even the shortest (~120 bp) 5′ intergenic regions of mRNA genes with inverted orientation were associated with two PICs, one for each mRNA direction. Thus, divergent mRNA transcription originates from two distinct PICs, even when arising from the same NFR.

We next examined the composition and location of PICs associated with mRNA and ncRNA (variously classified as cryptic unstable transcripts (CUTs), stable uncharacterized transcripts (SUTs), Xrn1-sensitive unstable transcripts (XUTs) and antisense (AS))[42–44]. We also examined orphan PICs, which we defined as being >160 bp from any annotated TSS or ORF start site. Nearly all had the same relative composition of GTFs, including Taf1 depletion or enrichment (Fig. 5a), although mRNA PICs generally had higher occupancy levels. GTFs had highly correlated occupancies at all PICs (see Supplementary Fig. 7 for mRNA PICs). ncRNA PICs were generally organized around an adjacent nucleosome, as seen for mRNA PICs (Supplementary Fig. 8). Thus, all mRNA, ncRNA and orphan PICs are compositionally homogeneous with regards to the GTFs (excluding TAFs).

To visualize better the context of the low-occupancy ncRNA PICs with mRNA genes, we marked ncRNA PIC locations by their TATA element, and plotted their directionality with respect to nearby mRNA (Fig. 5b). We observed a general trend where ncRNA and mRNA PICs were positioned in opposite directions 150–200 bp apart. This places both PICs within the same NFR, and thus within the same canonical nucleosome architecture, as seen for two divergent mRNA PICs that share the same NFR (Supplementary Fig. 9).

Low-occupancy ncRNA PICs were also found towards the 3′ ends of mRNA genes, of which the majority were antisense to the mRNA (Fig. 5b), and associated with low expression of the sense mRNA

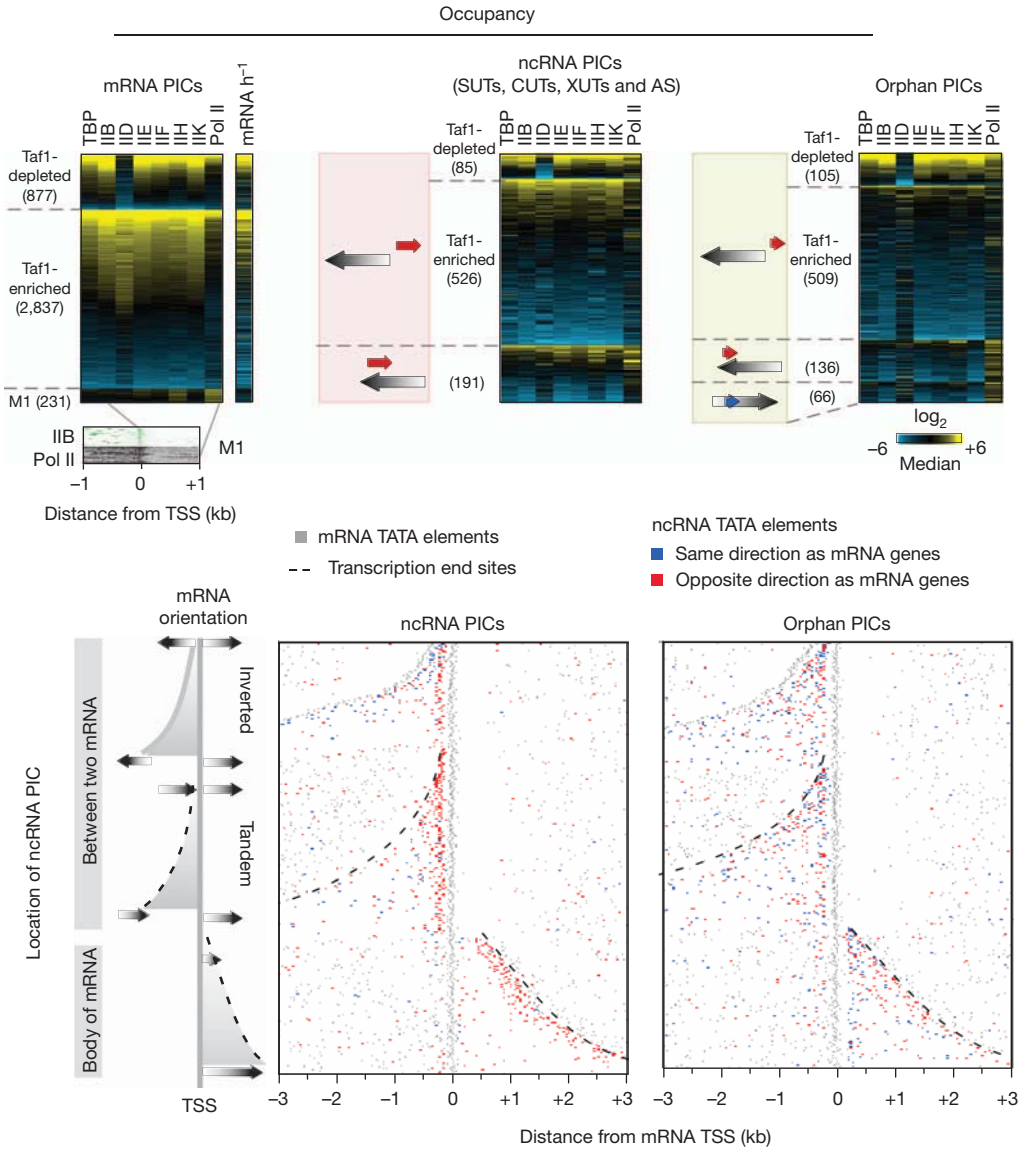


Figure 5 | Distribution of ncRNA PICs. a, GTF occupancy levels at PICs for mRNA, ncRNA and orphans, sectioned by *k*-means clustering. Occupancy levels of each GTF were median normalized, $\log_2$ transformed, then sorted by row median. The small M1 group represents a terminating polymerase originating from upstream. **b**, Distribution of ncRNA PICs relative to mRNA genes. mRNA genes were filtered to retain only those having a nearby ncRNA-associated PIC (defined as having >10% of the genome-wide TFIIB average). Plotted are the locations of TATA elements associated with the mRNA (grey), sense-directed ncRNA (blue), and antisense-directed ncRNA (red). Additional plot details are as described in Fig. 4.

(Supplementary Fig. 10a). Thus, ncRNAs, which tend to be antisense[40], are generally associated with repression when residing in gene bodies.

In total, we identified ~6,000 PICs in rapidly growing yeast cells, in which the PICs had an occupancy level of >10% of the genome average. More than 98% of these PICs had a TATA element precisely where TBP bound. Approximately 70% of the identified PICs were associated with mRNA genes (Supplementary Fig. 10b). The remaining ~30% were divided evenly between ncRNA and orphans. At lower detection thresholds, many more low-occupancy PICs could be identified. We do not believe that they represent technical noise, as they are highly enriched in NFRs in which mRNA and ncRNA PICs are found. They might produce low levels of promoter-specific basal transcription.

Unifying principles of PICs

Our data suggest that with the exception of TAFs, PICs are compositionally homogeneous in regards to GTFs at coding and non-coding Pol II transcription units in the yeast genome. PICs differ markedly in occupancy levels, which is in accord with their transcription frequency. PICs tend to form at NFR/nucleosome borders at the 5′ end (and to some extent at the 3′ end) of genes, where they direct either mRNA or ncRNA transcription away from the NFR. As such, an NFR may normally accommodate two divergently oriented PICs at markedly different occupancy and transcription levels. These occupancy levels do not strictly correlate, which suggests largely independent control of the two PICs.

PIC assembly, orientation and positioning may be contributed to in part by the resident TATA element, as well as through sequence-specific factors and co-factors. TFIID-regulated promoters may rely less on TATA element strength, and more on an NFR-adjacent nucleosome for PIC assembly, orientation and positioning. The adjacent nucleosome might also serve to impede a scanning Pol II so that it can productively select a TSS at a focused position just inside the nucleosome. At SAGA-regulated promoters, which tend to contain a consensus TATA box, nucleosome occupancy may be more competitive with PIC assembly, wherein the strength of TBP/TATA interactions would be more important for PIC assembly, orientation and positioning. As such, there would be no nearby nucleosome to impede polymerase scanning, which would allow TSS selection to be controlled by other factors, including DNA sequence.

The emergent concept of ncRNA and the difficulty of distinguishing random transcriptional noise from specific initiation raise the question as to what constitutes a gene[45]. The unambiguous and precise mapping of PIC locations across a genome, as described here, might help define the start of individual genes.

METHODS SUMMARY

Saccharomyces strains (BY4741) bearing TAP-tagged GTF or Pol II subunits (or untagged TBP) were grown to exponential phase in yeast extract peptone dextrose (YPD) media (30 °C to $OD_{600\,nm} = 0.8$), then subjected to 1% formaldehyde crosslinking for 15 min at 25 °C. Cells were harvested and washed. Sonicated chromatin was prepared by standard methods. Standard ChIP methods were used, followed by lambda exonuclease treatment and library construction, as described elsewhere[22]. Libraries were sequenced by an ABI SOLiD sequencer. Figures displaying strand-specific sequencing tags represent the raw data without normalization to input. TFIIB peak calls were made with GeneTrack software[46]. PICs ($n = 6{,}045$) were identified as having a TFIIB peak-pair in at least two out of four biological replicates and having ≥33 sequence tags (>10% of average TFIIB occupancy)[22]. PICs were assigned to the nearest TSS within ± 200 bp, with mRNA[47] having precedence over ncRNA. For this purpose, ORFs lacking a TSS (from the *Saccharomyces* Genome Database) were assigned a hypothetical TSS based on the genome-wide consensus. PICs of ncRNA were assigned to the nearest TSS within ± 200 bp of SUTs, CUTs, ASs and XUTs[42–44], with SUTs/CUTs having precedence over AS/XUTs. To assign directionality to orphan PICs, we compared nucleosome occupancy on the lower versus higher coordinate side of TFIIB locations. If the higher coordinate had higher nucleosome occupancy it was classified as 'sense', otherwise it was 'antisense'. We validated this method by applying it to mRNA PICs, and found >91% of the assignments to be correct. We searched for TATA elements between −80 to +20 bp of the midpoint of 6,045 TFIIB-bound locations on the sense strand, first by searching for the consensus TATAWAWR, then for one and then two mismatches to the consensus. The TATA element closest to −28 bp of a TFIIB peak had precedence if multiple elements were found.

Received 8 October; accepted 20 December 2011.
Published online 18 January 2012.

1. Green, M. R. TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes. *Trends Biochem. Sci.* **25,** 59–63 (2000).
2. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56,** 549–561 (1989).
3. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev.* **10,** 2657–2683 (1996).
4. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–335 (1996).
5. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nature Rev. Genet.* **10,** 161–172 (2009).
6. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116,** 699–709 (2004).
7. Huisinga, K. L. & Pugh, B. F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in *Saccharomyces cerevisiae*. *Mol. Cell* **13,** 573–585 (2004).
8. Lee, T. I. *et al.* Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature* **405,** 701–704 (2000).
9. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18,** 1084–1091 (2008).
10. Sugihara, F., Kasahara, K. & Kokubo, T. Highly redundant function of multiple AT-rich sequences as core promoter elements in the TATA-less RPS5 promoter of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **39,** 59–75 (2011).
11. Mohibullah, N. & Hahn, S. Site-specific cross-linking of TBP *in vivo* and *in vitro* reveals a direct functional interaction with the SAGA subunit Spt3. *Genes Dev.* **22,** 2994–3006 (2008).
12. Dudley, A. M., Rougeulle, C. & Winston, F. The Spt components of SAGA facilitate TBP binding to a promoter at a post-activator-binding step *in vivo*. *Genes Dev.* **13,** 2940–2945 (1999).
13. Bhaumik, S. R. & Green, M. R. Differential requirement of SAGA components for recruitment of TATA-box-binding protein to promoters *in vivo*. *Mol. Cell. Biol.* **22,** 7365–7371 (2002).
14. Kostrewa, D. *et al.* RNA polymerase II–TFIIB structure and mechanism of transcription initiation. *Nature* **462,** 323–330 (2009).
15. Liu, X., Bushnell, D. A., Wang, D., Calero, G. & Kornberg, R. D. Structure of an RNA polymerase II–TFIIB complex and the transcription initiation mechanism. *Science* **327,** 206–209 (2010).
16. Li, Y., Flanagan, P. M., Tschochner, H. & Kornberg, R. D. RNA polymerase II initiation factor interactions and transcription start site selection. *Science* **263,** 805–807 (1994).
17. Pardee, T. S., Bangur, C. S. & Ponticelli, A. S. The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. *J. Biol. Chem.* **273,** 17859–17864 (1998).
18. Yan, Q., Moreland, R. J., Conaway, J. W. & Conaway, R. C. Dual roles for transcription factor IIF in promoter escape by RNA polymerase II. *J. Biol. Chem.* **274,** 35668–35675 (1999).
19. Kim, T. K., Ebright, R. H. & Reinberg, D. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science* **288,** 1418–1421 (2000).
20. Keogh, M. C., Cho, E. J., Podolny, V. & Buratowski, S. Kin28 is found within TFIIH and a Kin28–Ccl1–Tfb3 trimer complex with differential sensitivities to T-loop phosphorylation. *Mol. Cell. Biol.* **22,** 1288–1297 (2002).
21. Svejstrup, J. Q., Feaver, W. J. & Kornberg, R. D. Subunits of yeast RNA polymerase II transcription factor TFIIH encoded by the *CCL1* gene. *J. Biol. Chem.* **271,** 643–645 (1996).
22. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single nucleotide resolution. *Cell* **147,** 1408–1419 (2011).
23. Giardina, C. & Lis, J. T. DNA melting on yeast RNA polymerase II promoters. *Science* **261,** 759–762 (1993).
24. Giardina, C. & Lis, J. T. Dynamic protein–DNA architecture of a yeast heat shock promoter. *Mol. Cell. Biol.* **15,** 2737–2744 (1995).
25. Forget, D., Langelier, M. F., Therien, C., Trinh, V. & Coulombe, B. Photo-cross-linking of a purified preinitiation complex reveals central roles for the RNA polymerase II mobile clamp and TFIIE in initiation mechanisms. *Mol. Cell. Biol.* **24,** 1122–1131 (2004).
26. Lagrange, T. *et al.* High-resolution mapping of nucleoprotein complexes by site-specific protein–DNA photocrosslinking: organization of the human TBP–TFIIA–TFIIB–DNA quaternary complex. *Proc. Natl Acad. Sci. USA* **93,** 10620–10625 (1996).
27. Chen, H. T. & Hahn, S. Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC. *Cell* **119,** 169–180 (2004).
28. Pal, M., Ponticelli, A. S. & Luse, D. S. The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol. Cell* **19,** 101–110 (2005).

29. Kuehner, J. N. & Brow, D. A. Quantitative analysis of *in vivo* initiator selection by yeast RNA polymerase II supports a scanning model. *J. Biol. Chem.* **281,** 14119–14128 (2006).
30. Matangkasombut, O., Buratowski, R. M., Swilling, N. W. & Buratowski, S. Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. *Genes Dev.* **14,** 951–962 (2000).
31. Koerber, R. T., Rhee, H. S., Jiang, C. & Pugh, B. F. Interaction of transcriptional regulators with specific nucleosomes across the *Saccharomyces* genome. *Mol. Cell* **35,** 889–902 (2009).
32. Vermeulen, M. *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131,** 58–69 (2007).
33. Faitar, S. L., Brodie, S. A. & Ponticelli, A. S. Promoter-specific shifts in transcription initiation conferred by yeast TFIIB mutations are determined by the sequence in the immediate vicinity of the start sites. *Mol. Cell. Biol.* **21,** 4427–4440 (2001).
34. Mayer, A. *et al.* Uniform transitions of the general RNA polymerase II transcription complex. *Nature Struct. Mol. Biol.* **17,** 1272–1278 (2010).
35. Ahn, S. H., Keogh, M. C. & Buratowski, S. Ctk1 promotes dissociation of basal transcription factors from elongating RNA polymerase II. *EMBO J.* **28,** 205–212 (2009).
36. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Struct. Mol. Biol.* **11,** 394–403 (2004).
37. Yudkovsky, N., Ranish, J. A. & Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408,** 225–229 (2000).
38. Flom, G., Weekes, J. & Johnson, J. L. Novel interaction of the Hsp90 chaperone machine with Ssl2, an essential DNA helicase in *Saccharomyces cerevisiae. Curr. Genet.* **47,** 368–380 (2005).
39. Freeman, B. C. & Yamamoto, K. R. Disassembly of transcriptional regulatory complexes by molecular chaperones. *Science* **296,** 2232–2235 (2002).
40. Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Rev. Genet.* **10,** 833–844 (2009).
41. Wei, W., Pelechano, V., Jarvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27,** 267–276 (2011).
42. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037 (2009).
43. Granovskaia, M. V. *et al.* High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol.* **11,** R24 (2010).
44. van Dijk, E. L. *et al.* XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475,** 114–117 (2011).
45. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17,** 669–681 (2007).
46. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack—a genomic data processing and visualization framework. *Bioinformatics* **24,** 1305–1306 (2008).
47. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103,** 5320–5325 (2006).
48. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728 (1998).
49. Ansari, A. & Hampsey, M. A role for the CPF 3′-end processing machinery in RNAP II-dependent gene looping. *Genes Dev.* **19,** 2969–2978 (2005).
50. Singh, B. N. & Hampsey, M. A transcription-independent role for TFIIB in gene looping. *Mol. Cell* **27,** 806–816 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. Albert and Y. Li for bioinformatic support, and members of the Pugh laboratory and the Penn State Center for Eukaryotic Gene Regulation for valuable discussions. Sequencing was performed at the Penn State Genomics Core Facility. This work was supported by National Institutes of Health grant GM059055.

Author Contributions H.S.R. performed the experiments and conducted data analyses. H.S.R. and B.F.P. conceived the experiments, analyses, and co-wrote the manuscript.

Author Information Sequencing data have been deposited at the NCBI Sequence Read Archive under accession number SRA046523. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.F.P. (bfp2@psu.edu).

ARTICLE

doi:10.1038/nature10802

DNA breaks and chromosome pulverization from errors in mitosis

Karen Crasta[1,2,3], Neil J. Ganem[1,2,3], Regina Dagher[1,2,3], Alexandra B. Lantermann[1], Elena V. Ivanova[4], Yunfeng Pan[5], Luigi Nezi[1], Alexei Protopopov[4], Dipanjan Chowdhury[5] & David Pellman[1,2,3]

The involvement of whole-chromosome aneuploidy in tumorigenesis is the subject of debate, in large part because of the lack of insight into underlying mechanisms. Here we identify a mechanism by which errors in mitotic chromosome segregation generate DNA breaks via the formation of structures called micronuclei. Whole-chromosome-containing micronuclei form when mitotic errors produce lagging chromosomes. We tracked the fate of newly generated micronuclei and found that they undergo defective and asynchronous DNA replication, resulting in DNA damage and often extensive fragmentation of the chromosome in the micronucleus. Micronuclei can persist in cells over several generations but the chromosome in the micronucleus can also be distributed to daughter nuclei. Thus, chromosome segregation errors potentially lead to mutations and chromosome rearrangements that can integrate into the genome. Pulverization of chromosomes in micronuclei may also be one explanation for 'chromothripsis' in cancer and developmental disorders, where isolated chromosomes or chromosome arms undergo massive local DNA breakage and rearrangement.

Whole-chromosome aneuploidy is a major feature of cancer genomes, yet its role in tumour development remains controversial[1,2]. This contrasts with chromosome breaks and rearrangements, which are known to produce cancer-causing mutations. Recent genetic evidence demonstrates that increased rates of whole-chromosome mis-segregation can accelerate oncogenesis[3–6]; however, the only established mechanism by which whole-chromosome segregation errors promote tumorigenesis is by facilitating the loss of heterozygosity for tumour suppressors[7]. Intriguingly, two animal models where whole-chromosome segregation errors result in robust tumour development also display extensive structural alterations in chromosomes[6,8]. This raises the question of whether errors in mitosis can predispose to DNA damage. We considered the possibility that segregation of chromosomes into micronuclei might produce DNA damage. Whole-chromosome-containing micronuclei form from anaphase lagging chromosomes[9–13]; micronuclei can also be generated from acentric fragments of chromosomes[11]. Micronuclei have many features of primary nuclei, but much controversy surrounds their actual composition and functional properties. Studies differ on whether micronuclei are transcriptionally active, replicate DNA, mount a normal DNA damage response, or assemble normal nuclear envelopes; moreover, the ultimate fate of chromosomes trapped within micronuclei remains unclear[11,14,15].

DNA damage in micronuclei

To determine if newly formed whole-chromosome micronuclei develop DNA damage, we generated micronuclei in synchronized cells and tracked them through the cell cycle. As a first synchronization approach, micronuclei were generated in non-transformed RPE-1 and transformed U2OS cells by release from nocodazole-induced microtubule depolymerization. When mitotic cells are released from nocodazole, spindles reassemble abnormally, producing merotelic kinetochore attachments (one kinetochore attached to two opposite spindle poles), lagging chromosomes, and ~10% of cells with micronuclei[16]. Because prolonged mitotic arrest causes DNA damage[17–19], RPE-1 cells were arrested for a short (6 h) interval with nocodazole. After release from the 6 h nocodazole block, neither the primary nuclei nor the newly formed kinetochore-positive micronuclei showed significant DNA damage during the subsequent G1 phase as measured by damage-dependent phosphorylation of the histone variant H2AX[20] (γ-H2AX foci formation; Fig. 1a–c and Supplementary Fig. 2a–f), TdT-mediated dUTP nick end labelling (TUNEL) to detect DNA breaks[21] (Fig. 1d and Supplementary Fig. 2e, f), and visualization of DNA breaks with the comet assay[22] (Supplementary Fig. 3d, e).

Because aneuploidy can produce a p53-dependent G1 cell-cycle arrest[23,24], p53 was silenced by RNA interference (RNAi) to allow us to monitor the fate of micronucleated RPE-1 cells at later stages of the cell cycle. As expected, after S-phase entry, low-level DNA damage was detected in both the micronuclei and the primary nucleus[25]; however, in G2 phase cells, most micronuclei showed DNA damage whereas almost none was detected in the primary nucleus (Fig. 1a–d). Similar results were observed in U2OS cells (Supplementary Fig. 3) and in cells where merotelic kinetochore–microtubule attachments were generated by knockdown of the kinetochore-associated microtubule depolymerase MCAK[26] or the kinetochore protein NUF2 (ref. 27) (Supplementary Fig. 4a–c). MCAK knockdown does not delay cells in mitosis, demonstrating that the acquisition of DNA damage in micronuclei is independent of mitotic arrest[17–19]. This damage did not represent activation of apoptosis because it was not accompanied by caspase-3 activation, and it was not suppressed by a pan-caspase inhibitor (Supplementary Fig. 5).

As a completely independent method for generating micronuclei, we used a human cell line (HT1080) carrying a human artificial chromosome (HAC) with a kinetochore that could be conditionally inactivated[28]. In this system, kinetochore assembly on the HAC is blocked by washout of doxycycline from the medium; consequently, the HAC is unable to attach to the mitotic spindle and is left behind at anaphase,

[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02115, USA. [2]Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [3]Howard Hughes Medical Institute, 450 Brookline Avenue, Boston, Massachusetts 02115, USA. [4]Belfer Institute for Applied Cancer Science, Dana-Farber Cancer Institute, 4 Blackfan Street, Boston, Massachusetts 02115, USA. [5]Department of Radiation Oncology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02115, USA.

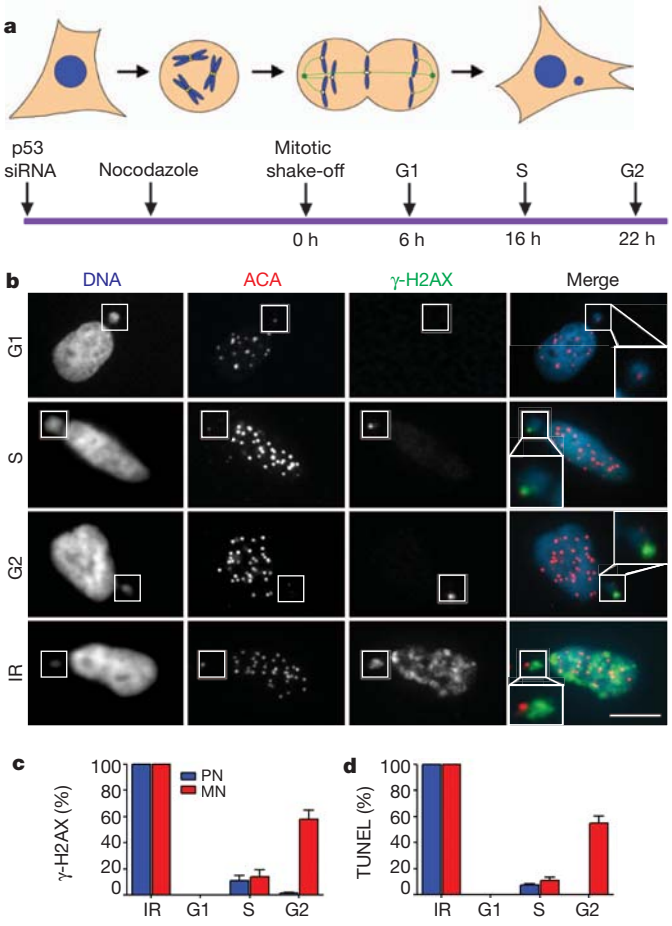


Figure 1 | Micronuclei from lagging chromosomes develop DNA breaks. **a**, Schematic of the experiment. **b**, Representative images of micronucleated G1 (6 h), S (16 h), G2 (22 h) and irradiated RPE-1 cells (IR) labelled for DNA (blue), centromeres (ACA, red) and γ-H2AX (green). Insets show enlarged images of micronuclei. Scale bars, 10 μm. **c**, **d**, Percentage of primary nuclei (PN, blue bars) and centric micronuclei (MN, red bars) with γ-H2AX foci (**c**) and TUNEL labelling (**d**) (3 experiments, $n = 100$). Errors bars indicate s.e.m.

ultimately reforming as a micronucleus (Fig. 2a). Results obtained with the HAC-containing micronuclei (Fig. 2b–d) were in agreement with the results obtained by the other synchronization methods described above. Taken together, micronuclei do not exhibit significant DNA damage during G1, but a large fraction of micronuclei acquire DNA damage during S phase that persists into G2.

Defective DNA replication in micronuclei

To determine directly whether acquisition of DNA damage in micronuclei requires DNA replication, synchronized micronucleated cells were released into medium containing thymidine to block DNA replication. Blocking DNA replication with thymidine abolished the acquisition of DNA damage (Fig. 3a–c), demonstrating that the breaks in micronuclei occur in a replication-dependent manner.

We next tested whether the acquisition of DNA damage in micronuclei is mediated by aberrant DNA replication. This possibility was suggested because micronuclei in G2 cells that had γ-H2AX foci were almost always additionally labelled for phosphorylated RPA2 (Ser 33)[29], a marker of DNA replication stress (Supplementary Fig. 4d). Moreover, the characteristic 5-bromodeoxyuridine (BrdU) labelling patterns of early, mid and late S-phase cells[30] suggested that DNA replication in micronuclei is inefficient and asynchronous with the primary nucleus (Fig. 3d). By measuring the fluorescence intensity of BrdU relative to DNA area from single focal plane confocal images, DNA replication in micronuclei was detected at a markedly reduced level throughout S phase when compared to the primary nuclei (Fig. 3e). Notably, pulse-labelling with BrdU demonstrated asynchronous DNA replication of the micronuclei relative to the primary nucleus, with ~30% of micronuclei replicating DNA in G2 cells (22 h; Fig. 3d).

The compromised DNA replication observed in newly formed micronuclei led us to test whether complexes necessary to license DNA replication origins are recruited to chromosomes in micronuclei. The first step in assembling replication origins is the loading of the origin recognition complex (ORC) after chromosome decondensation during telophase[31]. Quantitative fluorescence imaging suggested that ORC recruitment is equally efficient in the micronuclei as in the primary nucleus (Supplementary Fig. 6a). By contrast, newly generated micronuclei showed significant reduction in the recruitment of replicative DNA helicase components MCM2 and MCM3 (refs 32, 33) as well as the replication initiation factor CDT1 (ref. 31) (Fig. 4a, b and Supplementary Fig. 6b, c). We observed an increased recruitment of MCM subunits into micronuclei as cells progressed through the cell cycle, but the levels of MCM subunits never exceeded 20% of that observed in the primary nucleus (Fig. 4b).

To examine whether the DNA damage response in micronuclei was also abnormal, cells were treated with replication inhibitor (aphidicolin) or exposed to 2 Gy irradiation to induce damage and then labelled to detect components of the DNA damage response. After aphidicolin treatment, micronuclei showed clear signal for γ-H2AX and its kinase ATR (ataxia telangectasia mutated-related), but downstream constituents of the DNA damage response such as 53BP1 were not efficiently recruited, as previously reported[34] (Supplementary Fig. 7). A similar defect in the recruitment of DNA-damage-response factors into micronuclei was also observed after irradiation (Supplementary Fig. 8a). This recruitment defect had clear consequences for DNA damage repair: whereas γ-H2AX foci were resolved in primary nuclei by 6 h after 2 Gy irradiation, γ-H2AX foci persisted in micronuclei for >24 h (Supplementary Fig. 8b). The failure of micronuclei to recruit MCM subunits and components of the DNA repair machinery may be due, at least in part, to a defect in nucleocytoplasmic transport. Newly formed micronuclei have a marked reduction in the density of nuclear pores[11,14,15] (Supplementary Fig. 9) and exhibit strongly reduced nuclear import (Fig. 4c, d and Supplementary Fig. 10).

Micronuclei and chromosome pulverization

Next we tested the prediction that abnormal DNA replication in micronuclei could generate chromosome breaks. Chromosome spreads were prepared from non-transformed cells in the first cell cycle after release from nocodazole or from DMSO-treated controls (Fig. 5a–f and Supplementary Fig. 11a). Notably, 7.6% of chromosome spreads prepared from micronucleated cells released from nocodazole block ($n = 722$) exhibited chromosomes with a 'pulverized' appearance, characterized by fragments of 4′,6-diamidino-2-phenylindole (DAPI)-stained material that were often clustered into a discrete location on the spread[35] (Fig. 5a, d (yellow arrows), i and Supplementary Fig. 11b). By contrast, pulverization was exceedingly rare (0.14%) in controls. Multicolour fluorescence chromosome painting (SKY) confirmed that 34 of 47 spreads with pulverized chromosomes were composed of fragments from a single chromosome and another eight were from two chromosomes (Fig. 5a–i and Supplementary Fig. 11). Despite this marked disruption to chromosomal structure, only one cell harboured non-reciprocal chromosome translocations 24 h after nocodazole release ($n = 101$), whereas no translocations were observed in control cells (data not shown). At 72 h after nocodazole release ($n = 100$), we still did not detect chromosome translocations, but we did detect truncated or derivative single chromosomes in three cells. Although our experiments show that whole-chromosome mis-segregation does not produce frequent non-reciprocal translocations, we do not exclude the possibility that the DNA damage acquired in micronuclei could result in translocations, albeit at a frequency lower than ~1/100 per generation.

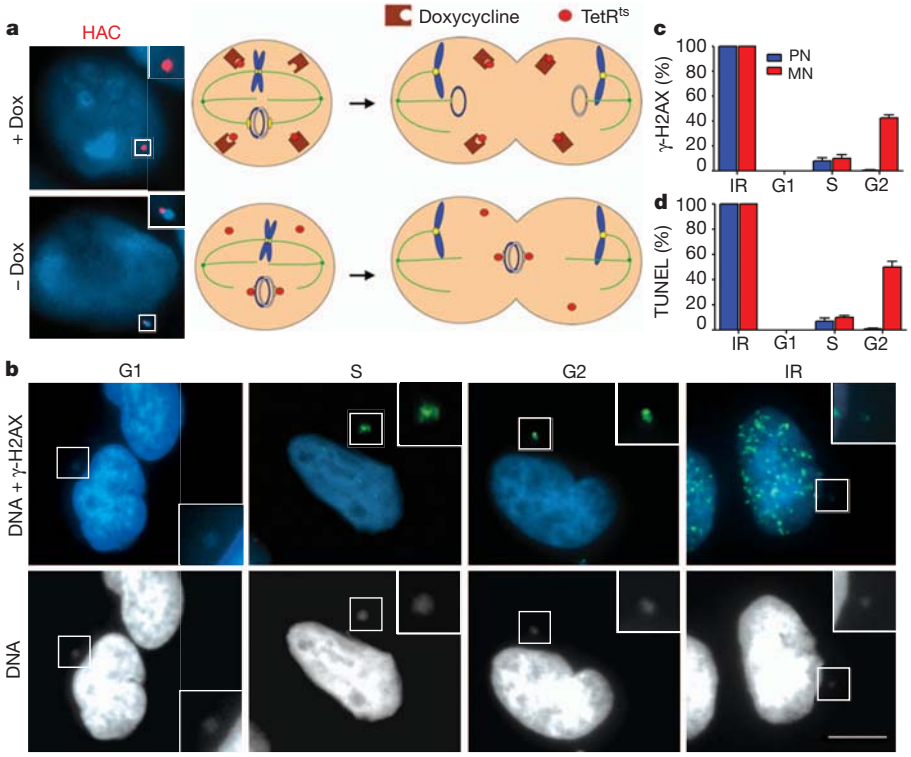


Figure 2 | DNA breaks in a HAC targeted to a micronucleus. **a**, Right: schematic; left: fluorescence *in situ* hybridization images of HAC (red) in a primary nucleus (+Dox) or micronucleus (−Dox). **b**, Images of micronucleated cells as in Fig. 1b (enlarged and brightened in insets). Scale bars, 10 μm. **c**, **d**, Percentage of primary nuclei (blue bars) and centric micronuclei (red bars) with γ-H2AX foci (**c**) and TUNEL labelling (**d**) (3 experiments, $n = 100$). Errors bars indicate s.e.m.

Chromosome pulverization has been observed in cell-fusion experiments where chromosomes from an S-phase cell are pulverized because of exposure to signals from mitotic cytoplasm[36]. It has also been observed when aberrant late-replicating chromosome translocations are generated[37]. The mechanism of pulverization involves compaction of partially replicated chromosomes induced by mitotic cyclin-dependent kinase activity, termed premature chromosome compaction[35,38,39]. By labelling micronucleated and control cells for 2 h with BrdU immediately before chromosome condensation and mitotic arrest, we found that 25 of 30 pulverized chromosomes were undergoing DNA replication (Fig. 5h). The fact that only micronuclei incorporate BrdU in G2 cells (Figs 3d and 5g) demonstrates that the pulverized chromosomes are derived from micronuclei.

The fate of chromosomes in micronuclei

We next addressed the potential for chromosome aberrations acquired in micronuclei to be reincorporated into the genome. We used long-term live-cell imaging to determine the fate of spontaneously arising or experimentally induced micronuclei in a variety of cell lines (Supplementary Table 1). Imaging analysis revealed that the vast majority of micronuclei (525 of 541; 97%) are stably maintained during interphase. Although micronuclei may be extruded in some cells[11], extrusion was not detected in our imaging experiments. Furthermore, micronuclei were not visibly degraded[11], did not co-localize with lysosomes (Supplementary Fig. 12), and did not appear to fuse with the primary nucleus. By contrast, our experiments indicated that after nuclear envelope breakdown some micronuclei might join the other mitotic chromosomes and be distributed to daughter nuclei. Of the 331 micronucleated cells that were tracked as they underwent cell division, 16% (53 of 331) gave rise to daughter cells

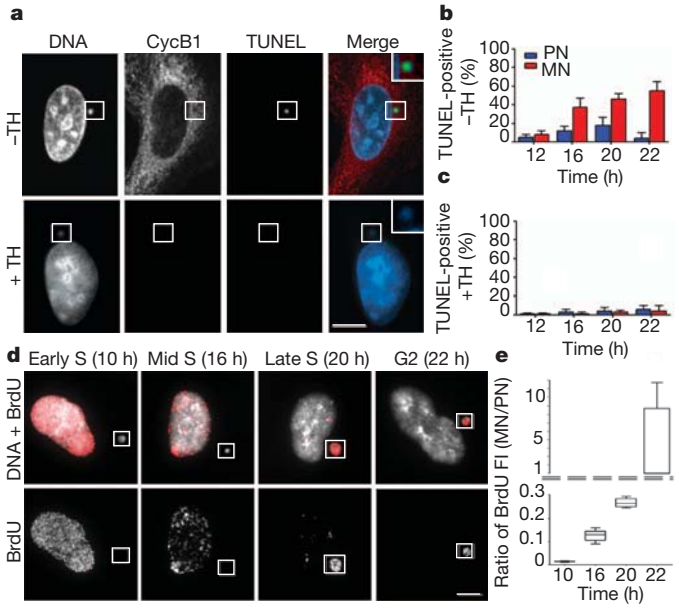


Figure 3 | DNA damage in micronuclei results from aberrant DNA replication. **a**–**c**, DNA replication requirement for acquisition of DNA damage in micronuclei. **a**, RPE-1 cells were synchronized as in Fig. 1a and released into media with (+TH) or without (−TH) 2 mM thymidine. Cells were co-labelled for TUNEL (green) and cyclin B1 (red). Shown are images from the 22 h sample. Scale bar, 10 μm. **b**, **c**, Percentage of TUNEL-positive primary nuclei (blue bars) and micronuclei (red bars) with (**c**) or without (**b**) thymidine treatment; (3 experiments, $n = 100$). Errors bars indicate s.e.m. **d**, Inefficient and asynchronous DNA replication in micronuclei. RPE-1 cells as in Fig. 1a were pulse-labelled with BrdU and labelled: DNA (white) and BrdU (red). **e**, The ratio of BrdU incorporation in micronuclei relative to primary nuclei after a 30-min pulse label. Normalized fluorescence intensity (FI) measurements are as shown in the box and whisker plots (2 experiments, $n = 50$). Boxes represent upper and lower quartiles, lines within boxes represent median, and the error bars comprise the whiskers which extend to the maximum and minimum value data sets.

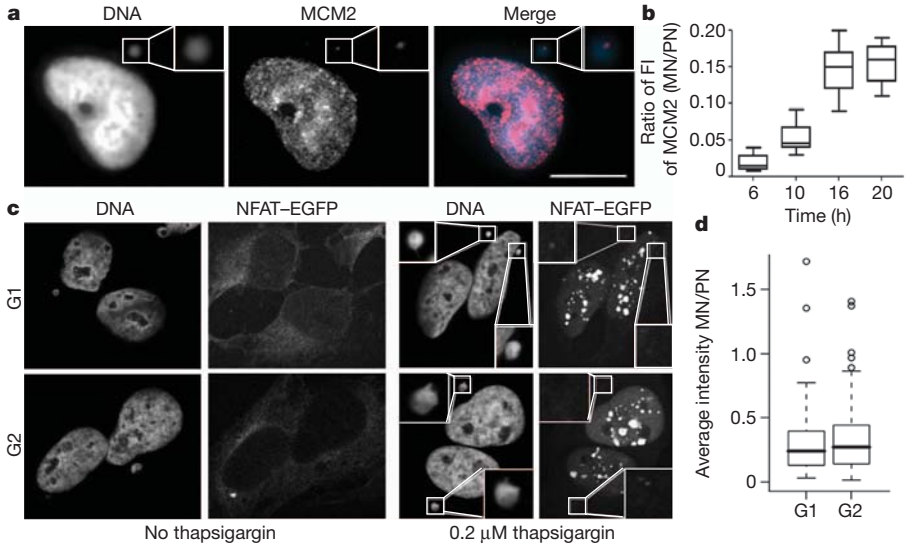


Figure 4 | Defective MCM2-7 complex recruitment, DNA damage response and nucleocytoplasmic transport in micronuclei. **a**, Impaired MCM2 recruitment into micronuclei. G1 RPE-1 cells were synchronized as in Fig. 1a and stained for chromatin-bound MCM2 and DNA. Scale bar, 10 μm. **b**, Relative MCM2 fluorescence intensity (FI) as in Fig. 3e. Approximate cell cycle stage of time points: 6 h, G1; 10 h, early S phase;16 h, mid S phase; 20 h, late S phase; (2 experiments, $n = 50$). **c**, **d**, Micronuclei are partially defective for nuclear import of NFATc1–EGFP. **c**, Representative images of micronucleated U2OS cells stably expressing NFATc1–EGFP in G1 and G2 with or without treatment with 0.2 μM thapsigargin for 10 min. **d**, Ratio of NFATc1–EGFP fluorescence intensity between micronuclei and primary nuclei; (3 experiments, $n = 30$). For **b** and **d**, error bars comprise the whiskers which extend to the maximum and minimum value data sets.

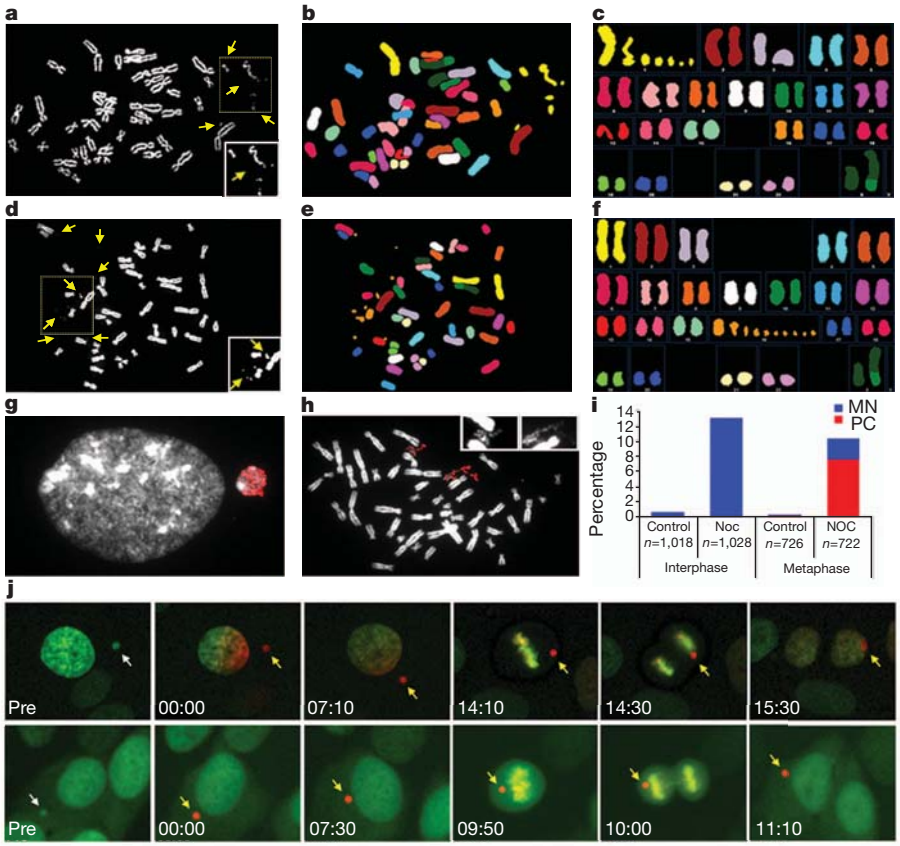


Figure 5 | The fate of chromosomes in micronuclei. Micronuclei were induced in RPE-1 cells as described in Fig. 1a (after p53 knockdown) and chromosome spreads were prepared 24 h after nocodazole release. **a**–**c**, Pulverization of chromosome 1 demonstrated by DAPI staining (chromosome fragments, brightened in inset, yellow arrows) (**a**), SKY probes (pseudo-coloured) (**b**), and aligned SKY karyotype (**c**). **d**–**f**, Pulverization of chromosome 16, viewed as in **a**–**c**. **g**, A BrdU-positive micronuclei-containing G2 RPE-1 cell. 2 h BrdU pulse label. DNA, white; BrdU, red. **h**, Selective BrdU labelling of a pulverized chromosome. **i**, The percentage of cells with intact micronuclei (blue bars) or pulverized chromosomes (PC, red bars) from control or nocodazole-released (NOC) RPE-1 cells. Interphase is 18-h sample and metaphase is 24 h. **j**, Images from live-cell experiment showing a pre-converted green fluorescent micronucleus (white arrows) photo-converted to a red-fluorescent micronucleus (yellow arrows) imaged through mitosis. After anaphase, the micronucleus either reincorporated with the primary nucleus or failed to reincorporate and reformed as a micronucleus. Top row: reincorporation of micronuclei into primary nucleus; bottom row: no reincorporation. Time is shown as h:min.

that lacked micronuclei (Supplementary Table 1 and Supplementary Movies 1 and 2).

To test directly whether micronuclei reincorporate during mitosis, we generated a U2OS cell line stably expressing H2B fused to the photoconvertible fluorescent protein Kaede[40], which converts from green to red fluorescence emission after ultraviolet illumination. This enabled us to photoactivate selectively and subsequently track a single chromosome encapsulated in a micronucleus throughout mitosis (Fig. 5j and Supplementary Movies 3 and 4). This experiment demonstrated that chromosomes within micronuclei reincorporate into daughter nuclei at a significant frequency (11 of 29; 38%). The remaining micronuclei persisted in cells well into the second generation.

Discussion

Together, our findings indicate a novel mechanism by which whole-chromosome segregation errors can cause chromosome breaks and potentially cancer-causing mutations. Mitotic errors can lead to lagging chromosomes which in turn can be partitioned into micronuclei[13]. This can produce DNA damage in two ways. First, DNA damage can result as a direct consequence of aberrant DNA replication, potentially due to a reduced density of replication origins. Second, chromosome pulverization due to premature chromosome compaction can occur if a micronucleated cell enters mitosis with micronuclei still undergoing DNA replication. This may not be the only mechanism by which whole-chromosome mis-segregation can cause DNA damage, as it was recently proposed that lagging chromosomes can be damaged by the cytokinetic furrow[41]. In contrast to previous studies that have suggested that most micronuclei are inactive, discarded or degraded[11], we find that micronuclei persist over several generations but that the chromosomes within the micronuclei can be segregated at a significant frequency into daughter cell nuclei. Thus, DNA rearrangements and mutations acquired in micronuclei could be incorporated into the genome of a developing cancer cell.

Our experiments suggest a mechanism that could explain the recently discovered phenomenon of chromothripsis[42,43]. Chromothripsis was discovered by cancer genome sequencing and DNA copy number analysis[42]. It is defined by small-scale DNA copy number changes and extensive intrachromosomal rearrangements that are restricted to a single chromosome or chromosome arm. Two non-exclusive models have been proposed for chromothripsis are: (1) fragmentation of a chromosome followed by stitching together of the resulting fragments by non-homologous end joining[42]; and (2) aberrant DNA replication resulting in fork stalling and template switching or microhomology-mediated break-induced replication[43,44]. Partitioning of a chromosome into a spatially isolated micronucleus can explain why a single chromosome would be subject to aberrant DNA replication and can explain how extensive DNA damage can be restricted to a single chromosome. Because micronuclei may persist as discrete entities over several generations, damage to the chromosome need not occur all at once[42,43], but could accrue from multiple failed attempts at DNA replication. Chromothripsis restricted to a chromosome arm could be generated from acentric chromosome fragments, as chromosome pulverization was previously correlated with the presence of micronuclei in cancer cell lines after DNA damage[35,45]. Although chromothripsis seems to be relatively rare, DNA damage in micronuclei could be more common. We propose that chromothripsis may be an extreme outcome of a mutagenesis mechanism that could be widespread in human cancer.

METHODS SUMMARY

All cell lines were maintained at 37 °C with 5% CO_2 atmosphere. To generate micronuclei, cells were treated with nocodazole (100 ng ml^{-1}) for 6 h followed by mitotic shake-off and nocodazole washout. Alternatively, micronuclei were generated by depletion of MCAK or NUF2 by short interfering RNA (siRNA). Micronuclei were also generated in HT1080 cells by inactivation of the kinetochore of the human artificial chromosome by washout of doxycycline. Detailed descriptions of immunofluorescence, RNAi, BrdU labelling, comet assays, inactivation of the human artificial chromosomes (HAC), flow cytometry, quantitative analysis of nuclear import, chromosome spreads, spectral karyotyping, antibodies used, photoconversion of H2B-Kaede and long-term live-cell imaging experiments can be found in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 August; accepted 20 December 2011.
Published online 18 January 2012.

1. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nature Rev. Genet.* http://dx.doi.org/10.1038/nrg3123 (in the press).
2. Weaver, B. A. & Cleveland, D. W. Aneuploidy: instigator and inhibitor of tumorigenesis. *Cancer Res.* **67,** 10103–10105 (2007).
3. Schvartzman, J. M., Sotillo, R. & Benezra, R. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. *Nature Rev. Cancer* **10,** 102–115 (2010).
4. Ricke, R. M., van Ree, J. H. & van Deursen, J. M. Whole chromosome instability and cancer: a complex relationship. *Trends Genet.* **24,** 457–466 (2008).
5. Weaver, B. A., Silk, A. D., Montagna, C., Verdier-Pinard, P. & Cleveland, D. W. Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell* **11,** 25–36 (2007).
6. Sotillo, R. *et al.* Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. *Cancer Cell* **11,** 9–23 (2007).
7. Baker, D. J., Jin, F., Jeganathan, K. B. & van Deursen, J. M. Whole chromosome instability caused by Bub1 insufficiency drives tumorigenesis through tumor suppressor gene loss of heterozygosity. *Cancer Cell* **16,** 475–486 (2009).
8. Fujiwara, T. *et al.* Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* **437,** 1043–1047 (2005).
9. Cimini, D. Merotelic kinetochore orientation, aneuploidy, and cancer. *Biochim. Biophys. Acta* **1786,** 32–40 (2008).
10. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460,** 278–282 (2009).
11. Terradas, M., Martin, M., Tusell, L. & Genesca, A. Genetic activities in micronuclei: is the DNA entrapped in micronuclei lost for the cell? *Mutat. Res.* **705,** 60–67 (2010).
12. Gisselsson, D. *et al.* Abnormal nuclear shape in solid tumors reflects mitotic instability. *Am. J. Pathol.* **158,** 199–206 (2001).
13. Thompson, S. L. & Compton, D. A. Chromosome missegregation in human cells arises through specific types of kinetochore-microtubule attachment errors. *Proc. Natl Acad. Sci. USA* **108,** 17974–17978 (2011).
14. Hoffelder, D. R. *et al.* Resolution of anaphase bridges in cancer cells. *Chromosoma* **112,** 389–397 (2004).
15. Géraud, G. *et al.* Three-dimensional organization of micronuclei induced by colchicine in PtK1 cells. *Exp. Cell Res.* **181,** 27–39 (1989).
16. Cimini, D. *et al.* Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J. Cell Biol.* **153,** 517–528 (2001).
17. Dalton, W. B., Nandan, M. O., Moore, R. T. & Yang, V. W. Human cancer cells commonly acquire DNA damage during mitotic arrest. *Cancer Res.* **67,** 11487–11492 (2007).
18. Quignon, F. *et al.* Sustained mitotic block elicits DNA breaks: one-step alteration of ploidy and chromosome integrity in mammalian cells. *Oncogene* **26,** 165–172 (2007).
19. Orth, J. D., Loewer, A., Lahav, G. & Mitchison, T. J. Prolonged mitotic arrest induces partial activation of apoptosis, DNA damage and p53 activation. *Mol. Biol. Cell* http://dx.doi.org/10.1091/mbc.E11-09-0781 (2011).
20. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J. Biol. Chem.* **273,** 5858–5868 (1998).
21. Gavrieli, Y., Sherman, Y. & Ben-Sasson, S. A. Identification of programmed cell death *in situ* via specific labeling of nuclear DNA fragmentation. *J. Cell Biol.* **119,** 493–501 (1992).
22. Singh, N. P., McCoy, M. T., Tice, R. R. & Schneider, E. L. A simple technique for quantitation of low levels of DNA damage in individual cells. *Exp. Cell Res.* **175,** 184–191 (1988).
23. Thompson, S. L. & Compton, D. A. Proliferation of aneuploid human cells is limited by a p53-dependent mechanism. *J. Cell Biol.* **188,** 369–381 (2010).
24. Li, M. *et al.* The ATM-p53 pathway suppresses aneuploidy-induced tumorigenesis. *Proc. Natl Acad. Sci. USA* **107,** 14188–14193 (2010).
25. Branzei, D. & Foiani, M. Maintaining genome stability at the replication fork. *Nature Rev. Mol. Cell Biol.* **11,** 208–219 (2010).
26. Kline-Smith, S. L., Khodjakov, A., Hergert, P. & Walczak, C. E. Depletion of centromeric MCAK leads to chromosome congression and segregation defects due to improper kinetochore attachments. *Mol. Biol. Cell* **15,** 1146–1159 (2004).
27. DeLuca, J. G., Moree, B., Hickey, J. M., Kilmartin, J. V. & Salmon, E. D. hNuf2 inhibition blocks stable kinetochore-microtubule attachment and induces mitotic cell death in HeLa cells. *J. Cell Biol.* **159,** 549–555 (2002).
28. Nakano, M. *et al.* Inactivation of a human kinetochore by specific targeting of chromatin modifiers. *Dev. Cell* **14,** 507–522 (2008).
29. Binz, S. K., Sheehan, A. M. & Wold, M. S. Replication protein A phosphorylation and the cellular response to DNA damage. *DNA Repair* **3,** 1015–1024 (2004).
30. O'Keefe, R. T., Henderson, S. C. & Spector, D. L. Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. *J. Cell Biol.* **116,** 1095–1110 (1992).
31. DePamphilis, M. L. & Bell, S. D. *Genome Duplication* (Garland Science, 2011).

32. Dimitrova, D. S., Prokhorova, T. A., Blow, J. J., Todorov, I. T. & Gilbert, D. M. Mammalian nuclei become licensed for DNA replication during late telophase. *J. Cell Sci.* **115,** 51–59 (2002).
33. Mendez, J. & Stillman, B. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol. Cell. Biol.* **20,** 8602–8612 (2000).
34. Terradas, M., Martin, M., Tusell, L. & Genesca, A. DNA lesions sequestered in micronuclei induce a local defective-damage response. *DNA Repair* **8,** 1225–1234 (2009).
35. Obe, G., Beek, B. & Vaidya, V. G. The human leukocyte test system. III. Premature chromosome condensation from chemically and x-ray induced micronuclei. *Mutat. Res.* **27,** 89–101 (1975).
36. Johnson, R. T. & Rao, P. N. Mammalian cell fusion: induction of premature chromosome condensation in interphase nuclei. *Nature* **226,** 717–722 (1970).
37. Smith, L., Plug, A. & Thayer, M. Delayed replication timing leads to delayed mitotic chromosome condensation and chromosomal instability of chromosome translocations. *Proc. Natl Acad. Sci. USA* **98,** 13300–13305 (2001).
38. Nichols, W. W., Levan, A., Aula, P. & Norrby, E. Chromosome damage associated with the measles virus *in vitro*. *Hereditas* **54,** 101–118 (1965).
39. Kato, H. & Sandberg, A. A. Chromosome pulverization in human binucleate cells following colcemid treatment. *J. Cell Biol.* **34,** 35–45 (1967).
40. Ando, R., Hama, H., Yamamoto-Hino, M., Mizuno, H. & Miyawaki, A. An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. *Proc. Natl Acad. Sci. USA* **99,** 12651–12656 (2002).
41. Janssen, A., van der Burg, M., Szuhai, K., Kops, G. J. & Medema, R. H. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* **333,** 1895–1898 (2011).
42. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
43. Liu, P. *et al.* Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146,** 889–903 (2011).
44. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10,** 551–564 (2009).
45. Sen, S., Hittelman, W. N., Teeter, L. D. & Kuo, M. T. Model for the formation of double minutes from prematurely condensed chromosomes of replicating micronuclei in drug-treated Chinese hamster ovary cells undergoing DNA amplification. *Cancer Res.* **49,** 6731–6737 (1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. D'Andrea, M. E. McLaughlin, T. A. Rapoport, J. Walters and Pellman laboratory members for discussions and/or comments on the manuscript; L. Cameron for advice and help on microscopy; H. Li for irradiation of cells; M. Nitta and M. Hennessy for preliminary experiments; and V. Larionov, B. Stillman, J. Ellenberg, I. Mattaj, A. Miyawaki and T. Kuroda for reagents. D.P. was supported by the Howard Hughes Medical Institute and the NIH (GM083299); K.C. was a fellow of A*STAR Singapore; N.J.G. was a fellow of the Leukemia and Lymphoma Society; Y.P. and D.C. were funded by the NIH (1R01CA142698-01).

Author Contributions D.P. conceived the project; K.C., N.J.G., R.D., A.B.L., D.C. and D.P. designed the experiments; D.P., K.C. and N.J.G. wrote the manuscript with edits from all authors; K.C. contributed Figs. 1–3, 4a, b and Supplementary Figs 2–8 with help from R.D. N.J.G. contributed Fig. 5j, Supplementary Fig. 12, Supplementary Table 1 and Supplementary Movies. R.D., E.V.I. and A.P. contributed Fig. 5a–i and Supplementary Figs 1 and 11; A.B.L. contributed Fig. 4c, d and Supplementary Figs 9 and 10; Y.P. and D.C. contributed Supplementary Fig. 3d; L.N. contributed Fig. 2a.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to david_pellman@dfci.harvard.edu.

METHODS

Cell culture. All cell lines were maintained at 37° C with 5% CO_2 atmosphere in DMEM (U20S, HT1080), DMEM:F12 (hTERT RPE-l), or DMEM:F12 medium without phenol red (H2B–GFP and H2B–mRFP U2OS, H2B–GFP and H2B–mRFP RPE-1, H2B–GFP Caco2, H2B–GFP SCC-114, H2B–GFP HeLa, H2B–mRFP MDA-231). All media were supplemented with 10% FBS, 100 IU ml^{-1} penicillin and 100 μg ml^{-1} streptomycin. HT1080 cells bearing the alphoid^{tetO} human artificial chromosome and expressing the tetracycline repressor TetR fused to the transcriptional silencer tTS (a gift from V. Larionov) were also supplemented with 1 μg ml^{-1} doxycycline (Sigma), 0.5 mg ml^{-1} G418 (geneticin, Invitrogen) and 4 μg ml^{-1} blasticidin S (Invitrogen). U2OS cells stably expressing human NFATc1 fused to the C terminus of EGFP (Thermo Scientific; R04-017-01) were supplemented with 0.5 mg ml^{-1} G418.

Generation of whole-chromosome micronuclei. To generate cells with whole-chromosome micronuclei, hTERT RPE-1 and U2OS cells were treated with 100 ng ml^{-1} nocodazole (Sigma) for 6 h. Mitotic cells were collected, washed twice with fresh medium containing 10% FBS, and then plated into medium containing 20% FBS where they completed cell division. To overcome the aneuploidy-induced p53-dependent G1 arrest associated with nocodazole washout experiments, RPE-1 cells were transfected with 50 nmol p53 siRNA (Smartpool, Dharmacon) using Lipofectamine RNAi Max (Invitrogen) 12 h before the nocodazole treatment, enabling cell-cycle progression of the subsequent aneuploid daughter cells with micronuclei. For NUF2 and MCAK knockdown, U2OS cells were serum starved for 2 days before transfection with 50 nmol NUF2 or MCAK siRNAs (Smartpool, Dharmacon). Cells were then washed in fresh medium containing 10% FBS and 24 h later treated with 4 μM dihydrocytochalasin B (DCB; Sigma) for 16 h to inhibit cytokinesis. The subsequent arrested tetraploid cells depleted of NUF2 or MCAK (many of which harboured micronuclei) were then treated for 6 h with 10 μM SB203580 (Sigma), a p38 inhibitor that promotes cell-cycle progression of tetraploids. Sequence information of the small interference RNA (siRNA) pools used from Dharmacon are as follows: human *TP53* ON-TARGETplus SMARTpool siRNA L-003329-00-0005, (J-003329-14) 5′-GAAAUUUGCGUGUGGAGUA-3′, (J-003329-15) 5′-GUGCAGCUGUGGGUUGAUU-3′, (J-003329-16) 5′-GCAGUCAGAUCCUAGCGUC-3′, (J-003329-17) 5′-GGAGAAUAUUUCACCCUUC-3′; human *KIF2C*/*MCAK* ON-TARGETplus SMARTpool siRNA L-004955-00-0005, (J-004955-06) 5′-GGCAUAAGCUCCUGUGAAU-3′, (J-004955-07) 5′-CCAACGCAGUAAUGGUUUA-3′, (J-004955-08) 5′-GCAAGCAACAGGUGCAAGU-3′, (J-004955-09) 5′-UGACUGAUCCUAUCGAAGA-3′; human *CDCA1*/*NUF2* ON-TARGETplus SMARTpool siRNA L-005289-00-0005, J-005289-06 5′-GAACGAGUAACCACAAUUA-3′, (J-005289-07) 5′-UAGCUGAGAUUGUGAUUCA-3′, (J-005289-08) 5′-GGAUUGCAAUAAAGUUCAA-3′, (J-005289-09) 5′-AAACGAUAGUGCUGCAAGA-3′.

Generation of micronuclei containing the HAC. HT1080 cells carrying the HAC were cultured in doxycycline-free medium and treated with 100 ng ml^{-1} nocodazole for 6 h to synchronize cells in mitosis. The absence of doxycycline enables TetR binding to TetO, which induces inactivation of the HAC centromere by tTS and the subsequent formation of a micronucleus after release from mitotic arrest. Cells released from mitotic arrest were synchronized at G0/G1 by serum starvation for 24 h, then released into the cell cycle with fresh medium containing 10% serum for further analysis.

Flow cytometry. Cells were fixed with 70% ethanol at 4 °C followed by incubation with 250 μg ml^{-1} RNase A and 10 μg ml^{-1} propidium iodide at 37 °C for 30 min. FACS analysis was performed with a FACScalibur flow cytometer (Becton Dickinson) and data analysed with CellQuest software.

Indirect immunofluorescence microscopy. For most experiments, cells were seeded on glass coverslips, washed in CSK buffer (10 mM PIPES, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$), pre-extracted with CSK buffer containing 0.5% Triton X-100 for 5 min at 4 °C, and fixed in PBS containing 4% paraformaldehyde for 15 min. After fixation, cells were permeabilized in PBS-0.2% Triton X-100 for 5 min, blocked in blocking buffer (PBS containing 5% FBS, 2% BSA and 0.1% Triton X-100) for 30 min, and then incubated with primary antibodies at room temperature for 1 h or overnight at 4 °C. Cells were washed with PBS-0.1% Triton X-100 and incubated with fluorescence-conjugated secondary antibodies (1:1,000, Molecular Probes) at room temperature for 1 h. Cells were also stained for DNA with Hoechst 33342 (1:5,000, Invitrogen) in PBS. More specific immunostaining protocols were as follows: for γ-H2AX and other DNA damage-response proteins, as in ref. 46; for ELYS and mAb414, ref. 47; for cyclin B1/γ-H2AX/BrdU, ref. 48, and for MCM2, CDT1 and ORC2, ref. 49. Images for most experiments were collected with a Yokogawa CSU-22 spinning disk confocal mounted on a Zeiss Axiovert microscope using 488, 561 and 640 nm laser light. A series of 0.3 μm optical sections was acquired using a ×100 1.4 NA Plan Apo objective with an Orca ER CCD camera (Hamamatsu Photonics). For nuclear pore quantification, images were collected on a Leica SP5 laser scanning confocal microscope with a 405 nm laser and a white light laser tuned to 488 nm, 568 nm and 647 nm using a ×63 1.4 NA Plan Apo objective. *Z*-stacks were collected with a 0.3 μm step size with pinhole at 1 Airy unit. Photomultipliers were calibrated to ensure linear range. For quantification of MCM2 and BrdU signals, mean fluorescence intensity within the primary nuclei or micronuclei was measured using ImageJ software. For quantification of γ-H2AX, primary nuclei were classified as positive when five or more foci were detected, with the threshold of foci intensity set against the irradiated sample. The same threshold was applied to score γ-H2AX foci in micronuclei. Acquisition parameters, shutters, filter positions and focus were controlled by Slidebook software (Intelligent Imaging Innovations). Images presented in figures are maximum intensity projections of entire *z*-stacks, unless otherwise stated.

Antibodies for immunofluorescence. Samples were incubated with primary antibodies for human anti-centromere ACA (1:1,000; Antibodies, Inc.), rabbit γ-H2AX-Ser 139 (1:500, Cell Signaling), mouse γ-H2AX (clone JBW301, 1:500, Upstate), rabbit 53BP1 (1:500, Cell Signaling), rabbit phospho-53BP1-Ser 1778 (1:500, Cell Signaling), rabbit phospho-CHK1-Ser 317 (1:500, Cell Signaling), rabbit phospho-ATM-Ser 1981 (1:500, Cell Signaling), rabbit ATR (1:500, Cell Signaling), rabbit MRE11 (1:300, Abcam), rabbit RAD51 (1:100, Santa Cruz), mouse BRCA1 (1:300, Santa Cruz), mouse RPA2 (1:500; Abcam), rabbit phospho-RPA2-S33 (1:500, Bethyl Laboratories), MCM2 and ORC2 (1:300, gifts from B. Stillman), rabbit CDT1 (1:500, Bethyl Laboratories), rabbit geminin (1:500; Santa Cruz), mouse monoclonal antibody 414 (1:5,000, Abcam), ELYS/MEL28 (1:1,000, gift from I. Mattaj), mouse lamin A/C (1:300, Abcam), rabbit cyclin A (1:300; Santa Cruz), rabbit cyclin B1 (1:500; Santa Cruz), rabbit cyclin D1 (1:300; Santa Cruz) and mouse BrdU (Sigma). Secondary antibodies used were Alexa Fluor 488 (green), 594 (red) and 647 (far red) from Molecular Probes.

TUNEL assay. TUNEL assay was performed according to the manufacturer's instructions (*In situ* cell death detection kit, Roche).

BrdU labelling. Cells were pre-labelled with 10 μM BrdU for 30 min and subsequently fixed and permeabilized according to the manufacturers' conditions (5-bromo-2-deoxyuridine labelling and detection kit 1, Roche). Cells were then stained with anti-BrdU (1 h, diluted 1:300) and processed for immunofluorescence. In the box and whiskers plot, the box represents upper and lower quartiles, line within box represents median and the whiskers extend to the highest and lowest value data sets.

Long-term live-cell imaging. Labelled cells were grown on glass-bottom 12-well tissue culture dishes (Mattek) and imaged on either a Nikon TE2000-E2 or Nikon Ti-E inverted microscope. Both microscopes were equipped with a cooled CCD camera (TE2000, Orca ER, Hamamatsu; Ti-E, Coolsnap HQ2, Photometrics), a precision motorized stage (Bioprecision, Ludl), and Nikon Perfect Focus, and both were enclosed within temperature and CO_2-controlled environments that maintained an atmosphere of 37 °C and 3–5% humidified CO_2. GFP or RFP images were captured at multiple points every 5–10 min for 3–5 days with either ×10 (0.3 NA) or ×20 (0.5, 0.75 NA) objectives, and exposure to fluorescent light was minimized to the greatest extent possible (all image acquisition used neutral density filters and/or 2 × 2 binning). Cell viability was confirmed by the continuous observation of mitotic cells throughout the duration of experiments. Captured images from each experiment were analysed using NIS-Elements software.

Imaging of H2B-Kaede. Tandem repeats of Kaede (a gift of A. Miyawaki) were fused to H2B and cloned into the pLenti6/V5 lentiviral vector (Invitrogen). A U2OS cell line stably expressing H2B-Kaede was generated by lentiviral infection followed by repeated FACS sorting of GFP^+ cells. For tracking experiments, pre-converted micronuclei (GFP^+) were identified and photoconverted with a brief (1–3 s) pulse of 350–400 nm ultraviolet light from a DAPI filter cube. Micronuclei that efficiently converted (RFP^+) were subsequently imaged through the completion of the following mitosis using the same microscope setup as described above. Images were acquired every 10 min with a ×20 objective and 2 × 2 binning.

Chromosome spreads/spectral karyotyping. RPE-1 cells were treated with p53 siRNA (as previously described) for 16 h and then grown in medium supplemented with nocodazole (100 ng ml^{-1}) or vehicle control (DMSO) for 6 h. Mitotic cells were collected by shake-off, washed thoroughly in fresh medium without nocodazole, and re-plated in culture medium with 20% FBS to complete mitosis. FACS analysis demonstrated that the subsequent daughter cells progressed to S/G2-phase 18–20 h after release. At this point, interphase samples were collected and fixed to observe the fraction of cells containing micronuclei. In parallel, 100 ng ml^{-1} colcemid (Invitrogen) was added to the DMSO or nocodazole-treated cells for an additional 6 h and then fixed. All samples were similarly processed for fixation: cells were pelleted and re-suspended in a hypotonic solution of 0.075 M KCl for 18 min, fixed in Carnoy's fixative (3:1 methanol:glacial acetic acid) and washed four times with Carnoy's fixative. All fixed samples were spread on slides for staining or hybridization. Mitotic samples were stained with Giemsa or Hoechst to visualize chromosomal abnormalities. Spectral karyotyping (SKY) was performed

on interphase and mitotic samples (from both DMSO- and nocodazole-treated samples) according to the SkyPaint DNA kit H-5 for human chromosomes procedure (Applied Spectral Imaging, SKY000029) and imaged on a Nikon Eclipse E6000 microscope equipped with the SD300 Spectracube and Spectral Imaging acquisition software. To determine if pulverized chromosomes were derived from late-replicating chromosomes in micronuclei, the above procedure was also performed with addition of BrdU labelling reagent (Roche Applied Science) only in the last 2 h of colcemid treatment before fixation. Mitotic samples from the BrdU samples were treated according to the 5-bromo-2′-deoxy-uridine labelling and detection kit I protocol (Roche Applied Science, 11296736001), except that the anti-BrdU antibody was diluted 1:50.

Single-cell gel electrophoresis (comet) assay. Single-cell comet assays were performed according to the manufacturer's instructions (Trevigen). Briefly, U2OS cells were collected at G1, S and G2 phases of the cell cycle after nocodazole washout was used to induce micronuclei. As controls, G1 cells were irradiated at 10 Gy. Cells were re-suspended in cold PBS at 2×10^5 cells ml^{-1}, mixed with low-melt agarose (1:10 ratio) and spread on frosted glass slides. After the agarose solidified, the slides were sequentially placed in lysis and alkaline solutions (Trevigen). Slides were then subjected to electrophoresis at 12 V for 10 min in 1× TBE buffer, fixed with 70% ethanol, and stained with DAPI. Nuclei were visualized using epifluorescent illumination on a Zeiss microscope and images were analysed with the NIH Image J program. DNA damage was quantified for 50 cells with micronuclei for each experimental condition by determining the tail DNA percentage using Comet Score (TriTek) software.

Quantitative analysis of nuclear pore complexes. To determine nuclear pore complex density in primary nuclei and micronuclei, cells were pre-extracted, fixed, permeabilized and immunostained with monoclonal 414 and ELYS antibodies. Confocal images were collected on a Leica SP5 laser scanning confocal with both a 405-nm and white light laser (at 488nm and 568nm) using a ×63 Plan Apo 1.4 NA oil objective. *Z*-stacks were collected with a 0.3 μm step size with pinhole at 1 Airy unit. Photomultipliers were calibrated to ensure linear range. Single layers of the *Z*-stack showing optimal nuclear pore staining for the primary nucleus and the micronuclei were chosen for quantitative analysis using MetaMorph. The areas of primary nuclei and micronuclei were judged from Hoechst staining and used to determine the perinuclear rim area. Total fluorescence intensity of monoclonal antibody 414 and ELYS was measured in the perinuclear rim area. To compare nuclear pore densities between primary nuclei and micronuclei, the total nuclear pore fluorescence was divided by the perinuclear rim area of primary nuclei and micronuclei, resulting in an average intensity ratio.

Nuclear import. To measure nuclear import, we used U2OS cells stably expressing the fusion protein NFATc1–EGFP (Thermo-Fisher Scientific). The inactive transcription factor NFATc1 resides in the cytosol. Elevated calcium levels in the cytosol lead to the dephosphorylation of NFATc1 by calcineurin, thereby inducing its rapid translocation into the nucleus. To increase the calcium level in the cytoplasm, U2OS cells released from nocodazole washout were treated with 0.2 μM thapsigargin, an inhibitor of the sarco-endoplasmic reticulum Ca^{2+} ATPase. Nuclear import of NFATc1 was then measured in the primary nuclei and micronuclei of U2OS cells at both 6 h and 20 h after nocodazole release. Cells were subsequently fixed with 2% paraformaldehyde and import was quantified based on nuclear fluorescence. To measure nuclear import of IBB-DiHcRed[50], U2OS cells were transiently transfected with the IBB-DiHcRed-plasmid (gift from J. Ellenberg) using Lipofectamine 2000 (Invitrogen). Cells were then synchronized for 10 h with nocodazole, washed as previously described, and released into the cell cycle. Cells were fixed with 2% paraformaldehyde both 6 h and 20 h later, and import was quantified based on nuclear fluorescence.

46. Bekker-Jensen, S. *et al.* Spatial organization of the mammalian genome surveillance machinery in response to DNA strand breaks. *J. Cell Biol.* **173,** 195–206 (2006).
47. Franz, C. *et al.* MEL-28/ELYS is required for the recruitment of nucleoporins to chromatin and postmitotic nuclear pore complex assembly. *EMBO Rep.* **8,** 165–172 (2007).
48. Peterson-Roth, E., Reynolds, M., Quievryn, G. & Zhitkovich, A. Mismatch repair proteins are activators of toxic responses to chromium-DNA damage. *Mol. Cell. Biol.* **25,** 3596–3607 (2005).
49. Prasanth, S. G., Mendez, J., Prasanth, K. V. & Stillman, B. Dynamics of pre-replication complex proteins during the cell division cycle. *Phil. Trans. R. Soc. Lond. B* **359,** 7–16 (2004).
50. Dultz, E. *et al.* Systematic kinetic analysis of mitotic dis- and reassembly of the nuclear pore in living cells. *J. Cell Biol.* **180,** 857–865 (2008).

prostate or blood cancers — also display epigenetic changes that are crucial for tumour development and maintenance. Future experiments, including integrated genomic analyses, may provide the answer.

The underlying mechanism of alterations in gene expression by epigenetic means in retinoblastoma is still unclear. But Zhang and co-workers' observation[3] that 13% of these tumours show recurrent mutations in the *BCOR* gene offers a possible explanation. The BCOR protein is highly expressed in the fetal retina and is essential for eye development, as evidenced by a congenital eye disorder, syndromic microphthalmia, which results from a heritable mutation in *BCOR*[7]. BCOR associates with proteins that repress gene expression epigenetically[8]. This raises the possibility that loss of BCOR function, due to acquired mutations in its gene, may contribute to an altered epigenetic landscape in RB1-deficient retinal cells. Consequently, inappropriate expression of crucial genes may impair cell maturation and so facilitate the progression of retinoblastoma. It will be of interest to determine whether BCOR is part of the repression machinery that silences SYK expression in the normal retina, and whether acquired *BCOR* mutations provide at least one route to altered expression of SYK in retinoblastoma.

Nevertheless, the current work — using a comprehensive, integrated genomics approach — is notable not just for demonstrating that epigenetic alterations have a predominant role in the progression of retinoblastoma. The new possibilities it raises for therapies in this childhood malignancy, and possibly in other types of tumour, are equally noteworthy. ■

Julien Sage *is in the Departments of Pediatrics and of Genetics and* **Michael L. Cleary** *is in the Departments of Pathology and of Pediatrics, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California 94305, USA.*
e-mails: julsage@stanford.edu; michael.cleary@stanford.edu

1. Chinnam, M. & Goodrich, D. W. *Curr. Top. Dev. Biol.* **94,** 129–169 (2011).
2. Manning, A. L. & Dyson, N. J. *Trends Cell Biol.* **21,** 433–441 (2011).
3. Zhang, J. *et al. Nature* **481,** 329–334 (2012).
4. Stratton, M. R. *Science* **331,** 1553–1558 (2011).
5. Parsons, D. W. *et al. Science* **331,** 435–439 (2011).
6. Hansen, K. D. *et al. Nature Genet.* **43,** 768–775 (2011).
7. Ng, D. *et al. Nature Genet.* **36,** 411–416 (2004).
8. Fan, Z. *et al. Nature Cell Biol.* **11,** 1002–1009 (2009).

EVOLUTIONARY BIOLOGY

A ratchet for protein complexity

Molecular machines containing related protein subunits are common in cells. Reconstruction of ancient proteins suggests that this type of complexity can evolve in the absence of any initial selective advantage. SEE LETTER P.360

W. FORD DOOLITTLE

Organisms and cells are bewilderingly complicated, and the molecular machines that perform many basic cellular functions are often giant, multisubunit, multifunctional protein complexes with tangled evolutionary histories. It is generally assumed that such complexes arose by the stepwise accretion of individual proteins, each addition representing a selective advantage by adding to or refining the machine's performance. But on page 360 of this issue, Thornton and colleagues[1] argue against this standard explanation in one particular instance — that of a ring-shaped protein complex in fungi. The authors show how evolutionary processes entailing loss of function rather than gain might initially drive a system towards complexity, independently of selection*.

Many cellular molecular machines contain several proteins that self-assemble into a multi-subunit ring. In simple cases, rings are homo-oligomeric; that is, all the subunits are identical and thus probably the products of a single gene. In more complex (hetero-oligomeric) examples, the protein molecules in the ring are different, but may be related. Often, hetero-oligomeric rings seem to have arisen from homo-oligomeric complexes after a gene encoding a single subunit became duplicated, producing two or more gene copies (called paralogues), with each copy subsequently evolving to encode a slightly different protein.

The question that Thornton and colleagues address is, why does this happen? Do the structurally distinguishable subunits and/or their specific pattern of assembly confer improved or additional functions on the protein complex, with selection for their enhanced performance being the evolutionary driving force? Alternatively, might neutral processes be responsible, at least initially?

To understand how the latter possibility might come about, imagine that a gene encoding, for example, eight subunits of a homo-oligomeric ring undergoes duplication (Fig. 1). The two paralogous gene copies that this duplication produces — and the protein subunits they encode (call them A and B) — will of necessity begin to diverge through accumulation of neutral mutations. These mutations might not at first affect any of the subunits' functions, so that functional octameric rings could continue to form by a random sampling

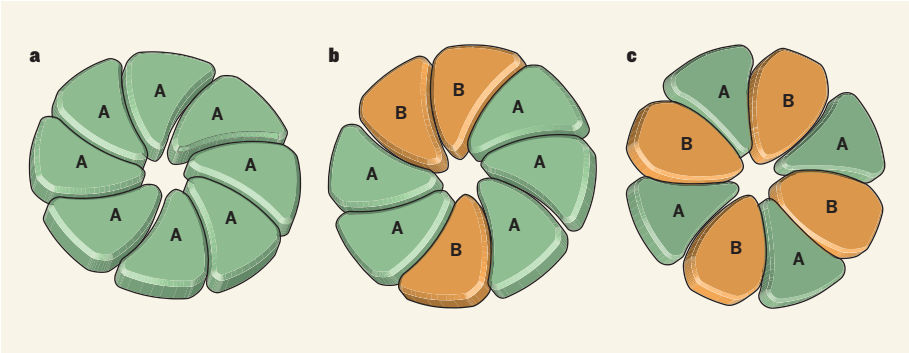


Figure 1 | Evolution of complexity in a protein ring. **a**, Homo-oligomeric protein complexes contain identical subunits (A) that are typically encoded by a single gene. **b**, If the gene is duplicated, the two gene copies can diverge through the accumulation of neutral mutations, generating structurally distinguishable but functionally unaltered subunits A and B, which can form functional hetero-oligomeric rings by random mixture. **c**, If additional mutations prevent subunits from binding to others of their own type, functional rings could still be formed by alternating subunit types. As further mutations accumulate, the probability of returning to the initial homo-oligomeric situation becomes very small. Thus, the subunit composition of a protein association may become complex in the absence of initial selective advantages. Thornton and colleagues[1] provide experimental evidence suggesting that this type of process has occurred in the evolution of a six-membered protein ring that forms part of the vacuolar H^+-ATPase enzymes in fungi.

*This article and the paper[1] under discussion were published online on 8 January 2012.

of A and B subunits. But further mutations might result in each subunit losing the ability to bind others of its own type, A to A or B to B. However, as long as A still binds B and B still binds A, functional rings can form.

As A and B continue to co-evolve through the acquisition of further neutral mutations, the likelihood of reversing back along the mutational path to the initial single-gene homo-oligomeric complex becomes vanishingly small. We can thus see the homo-oligomer to hetero-oligomer transition as an inexorable evolutionary ratchet. And even if there were some slight selective disadvantage to using a hetero-oligomeric complex (for example, complex formation could become slower), this could be overcome in small populations — in which the efficiency of selection is reduced and mildly harmful mutations can be fixed by chance[2].

Some chaperonins — molecular machines that assist other proteins to fold into functional configurations — have until recently provided the best evidence for such a scenario. Microorganisms known as archaea have one, two or three paralogous genes encoding chaperonin subunits that form rings: homo-oligomeric eight-membered rings when there is a single gene, or hetero-oligomeric eight- or nine-membered rings when there are two or three genes. Such paralogy seems to reoccur frequently in archaea, and a neutral co-evolutionary picture of the sort described above has been suggested[3] to explain it.

Comparative bioinformatic studies[4] support the idea that hetero-oligomeric archaeal chaperonins consisting of two different subunits are not functionally distinct from the homo-oligomers, and thus did arise neutrally. Adaptive changes can follow later, however. Indeed, there is direct evidence for specialized roles for the eight distinct chaperonin subunits of eukaryotes (organisms such as plants, animals and fungi)[5].

So a more realistic picture might be as follows. Rarely, gene duplication is quickly followed by functional differentiation (as proposed in early formulations of the hypothesis of evolution by gene duplication[6]). Selection against loss of either duplicate would then be ensured early on. Perhaps more often, though, the paralogous condition might first be locked in by the above-described co-evolutionary ratchet, allowing functional differences (if any) to be acquired later (if ever). That is what Thornton and colleagues[1] argue has happened in the case of a six-membered hetero-oligomeric ring that forms part of the enzymes known as vacuolar H^+-ATPases (V-ATPases) in fungi. Their argument is especially compelling because it rests not only on bioinformatics, but also on experiments, including the synthesis and testing of proteins that the authors propose are ancestral to the V-ATPase subunits.

In most eukaryotes, the V-ATPase six-membered ring is made up of five copies of the protein Vma3 and one of its paralogue Vma16. In fungi such as budding yeast, by contrast, one of the five Vma3 proteins is replaced by Vma11, which is a fungus-specific paralogue of Vma3. The abilities of these subunits to bind to one another[7] are such that Vma11 must lie between Vma16 and Vma3 in the ring, and Vma3 cannot form the interface with the side of Vma16 that Vma11 can. This represents a sort of functional degeneration, because the last common ancestor of Vma3 and Vma11 must have been able to interface with either side of Vma16, as Vma3 must still do in organisms other than fungi.

Thornton and colleagues[1] synthesized proteins that matched the phylogenetically inferred ancestor of Vma3 and Vma11, as well as that of Vma16 and other intermediates in the predicted evolutionary pathway. They demonstrate that Anc.3-11, the inferred ancestor of fungal Vma3 and Vma11, does indeed interface with either side of Vma16. The authors tested the functionality of these and other inferred evolutionary ancestors and intermediates by expressing them in yeast mutants that lacked current versions of one or more of the three genes. According to this picture, when Anc.3-11 underwent duplication, one paralogue (which was to become modern fungal Vma3) lost the ability to interact with the 'anticlockwise' side of Vma16. The other paralogue (which became Vma11) lost the ability to interact with the 'clockwise' side of Vma16 and/or the anticlockwise side of Vma3.

A general theory, sometimes called constructive neutral evolution, to explain how neutral processes might drive a system towards complexity is more than a decade old[8–10]. But the study by Thornton and colleagues[1] may provide the most compelling experimental evidence to date. Of course, one can never prove that some subtle, unidentified selective advantage was not involved in the evolution of the V-ATPase protein ring, but neutrality would seem the most justifiable default hypothesis. Thus, a neutral theory of molecular evolution, normally invoked for nucleotide substitutions, may also apply to certain higher-order structures such as multi-subunit protein rings. How general such neutral mutational drives to complexity might be is one of evolutionary theory's deeper unanswered questions. ■

W. Ford Doolittle *is in the Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada.*
e-mail: ford@dal.ca

1. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. *Nature* **481,** 360–364 (2012).
2. Lynch, M. *Proc. Natl Acad. Sci. USA* **104,** 8597–8604 (2007).
3. Archibald, J. M., Logsdon, J. M. Jr & Doolittle, W. F. *Curr. Biol.* **9,** 1053–1056 (1999).
4. Ruano-Rubio, V. & Fares, M. A. *Mol. Biol. Evol.* **24,** 1384–1396 (2007).
5. Fares, M. A. & Wolfe, K. H. *Mol. Biol. Evol.* **20,** 1588–1597 (2003).
6. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
7. Wang, Y., Cipriano, D. J. & Forgac, M. *J. Biol. Chem.* **282,** 34058–34065 (2007).
8. Stoltzfus, A. *J. Mol. Evol.* **49,** 169–181 (1999).
9. Force, A. *et al. Genetics* **151,** 1531–1545 (1999).
10. Gray, M. W., Lukeš, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. F. *Science* **330,** 920–921 (2011).

GALAXY FORMATION

Distant dwarfs

Astronomers value the smallest galaxies for the clues they can provide about how galaxies form. But only those nearest to us are easily detectable. A neat technique has allowed one such object to be found at a large distance. SEE LETTER P.341

ROBERT W. SCHMIDT

Some objects do not shine brightly enough to be seen easily. Examples of this are the small and extremely faint galaxies called dwarf or satellite galaxies. Only when they are near to our Galaxy might we be lucky enough to detect the light of their stars. Satellite galaxies are of great interest because simulations of the process of galaxy formation predict that there are many more satellites in the Local Group, which includes the Milky Way and other galaxies in its vicinity, than have yet been found[1–3]. This begs the question of whether there is something wrong with our understanding of how galaxies form. On page 341 of this issue, Vegetti *et al.*[4] report the discovery of a satellite in the vicinity of a galaxy located at a cosmological distance from Earth that allows them to test the simulation predictions for galaxies beyond those of the Local Group.

The object studied by Vegetti *et al.* is a distant elliptical galaxy that acts as a gravitational lens: light from a background source is deflected from its otherwise straight path by a massive foreground object, the lens. When compared with camera lenses, gravitational lenses have rather dismal optical properties because they can lead to large distortions and multiple images of the background object. But from an astronomer's point of view this is a good thing: the mass of the gravitational lens can be measured by reconstructing the background object from the distorted images.