nature

www.nature.com/nature

Vol 463 | Issue no. 7277 | 7 January 2010

A decade for psychiatric disorders

There are many ways in which the understanding and treatment of conditions such as schizophrenia are ripe for a revolution.

A media circus surrounded President Bill Clinton's visit to a New York medical centre in 2004 for a quadruple heart bypass. Yet barely a whisper was heard about other high-profile individuals' visits there for the treatment of psychiatric disorders.

In Britain, the public donates £500 million (US$800 million) each year to charities for cancer research. For mental-health research, the figure is a few million, and most of that is for work on neurodegenerative diseases such as Alzheimer's, rather than for earlier-onset conditions that can undermine people's entire lives, such as depressive disorders.

It is time for such disparities to be addressed in a more coherent and aggressive way than in the past. The stigma of psychiatric disorders is misplaced, their burdens on society are significantly greater than more publicized diseases in developed and developing nations alike, and biomedical science is poised to make significant strides. The timescales are daunting and the challenges great — human neurons are less accessible than tumour cells, separating genetic and environmental influences is tough, and the diagnosis of the conditions is highly problematic. There is much to be done, and a decade is the timescale over which enhanced commitment is required.

The problem of stigma persists. In some countries, progress in this regard has been made with depression: a few high-profile and brave sufferers in some Western countries have stood up and identified themselves. By contrast, schizophrenia, when covered by the media at all, is mostly associated with murders carried out by a tiny minority of sufferers who have an acute form of the condition.

Research challenges

Schizophrenia — a combination of delusions, reduced motivation and diminished cognitive functions — exemplifies many of the research challenges posed by psychiatric disorders as a whole. The extreme behaviours covered by the media are far from typical. Population studies indicate that the lifetime prevalence of all psychotic disorders (whose sufferers experience some sort of misperception of reality) is as much as 3%. Schizophrenia is controllable by medication and cognitive therapy, with a significant chance (a few tens of per cent) of beneficial positive outcomes.

Frustratingly, the effectiveness of medications has stalled. Nobody understands the links between the symptoms of schizophrenia and the crude physiological pathologies that have so far been documented: a decrease in white brain matter, for example, and altered function of the neurotransmitter dopamine. The medications, which are often aimed at the dopamine systems associated with delusions, have advanced over the decades not in their efficacy but in a reduction of their debilitating side effects.

Both diagnosis and drugs primarily address a late stage in the development of schizophrenia — the presentation of delusions. The earlier stages are much less well defined and are ambiguous in that, as currently characterized, they could lead to a number of alternative conditions. Here, above all, is where progress is needed in the form of reliable biomarkers to identify those at risk and to allow biomedical or cognitive interventions to prevent or mitigate the development of the disorders. Early intervention would lead to better outcomes.

A deeper understanding of the underlying biology is essential to improve diagnoses and therapies. New techniques — genome-wide association studies, imaging and the optical manipulation of neural circuits — are ushering in an era in which the neural circuitry underlying cognitive dysfunctions, for example, will be delineated. Tantalizingly, work in genetics is indicating how non-specific some genes are for schizophrenia, having associations in common with bipolar disorder and with autism. This suggests that the earlier stages of psychiatric disorders are multivalent, reinforcing the hope that early detection, coupled with a clearer understanding of the environmental factors, may allow prevention.

> "Early detection and a clearer understanding of environmental factors may allow prevention of psychiatric disorders."

Environmental influence

Too little fundamental research is devoted to environmental factors. About 80% of the pattern of schizophrenia in populations seems to be determined by genetics, but part of that genetic influence lies in susceptibility to environmental influences. The remaining 20% of direct environmental influence is also ripe for more extensive investigation — epidemiological studies point to social stress (associated, for example, with migration or urbanization) as a significant influence, albeit in a minority of schizophrenia sufferers. As stated in a recent review of schizophrenia, a "worldwide challenge is to bring together the various disciplines that are needed to examine models of disease causation based on various aspects of gene–environment interplay" (J. van Os and S. Kapur *Lancet* **374,** 635–645; 2009).

Of course it won't be just the basic biology of molecules and their circuits that will be essential in understanding the mechanisms of schizophrenia. There is a higher level of explanation required to understand, for example, delusions and their persistence.

Whether for schizophrenia, depression, autism or any other psychiatric disorders, it is clear, as Tom Insel, head of the US National Institute of Mental Health has emphasized (T. R. Insel *J. Clin. Invest.* **119,** 700–705; 2009), that understanding of these conditions is entering a scientific phase more penetratingly insightful than has hitherto been possible. But Insel also highlights the disruptive impact of the science on the practices of clinical psychiatrists — as biological insights develop, the crudity of current psychiatric diagnoses will become all too clear. Yet the exposure of many psychiatrists to contemporary biology is shallow at best. That, too, will need to change over the next decade. ■

NEWS BRIEFING

M. DUENAS/EPA/CORBIS

ERUPTIONS MARK NEW YEAR

Colombia's Galeras volcano (pictured) erupted on 2 January; no injuries were reported as *Nature* went to press. The same day, the Nyamuragira volcano in the Democratic Republic of the Congo began spewing ash and lava into Virunga National Park, threatening rare chimpanzees.

● POLICY

Japan's budget: Science projects in Japan emerged with minor scars from the new government's 2010 budget, after researchers protested against recommendations for drastic cuts. A supercomputer project fell just ¥4 billion (US$43 million) short of hoped-for funding, delaying its scheduled start by 7 months to June 2012, and Earth and oceanographic research spending dropped by ¥16 billion to ¥54 billion. But the SPring-8 synchrotron's support was largely unscathed; a basic-science budget — including a controversial grant programme for innovative research — jumped by ¥47 billion to ¥341 billion; and funding for green projects such as low-carbon technology almost tripled to ¥10 billion.

● RESEARCH

Space shortlist: Candidates for NASA's newest mid-size mission to visit another body in the Solar System were narrowed down from eight to three on 29 December. The proposals are landing on Venus; bringing back material from the Moon's south pole; or fetching a sample of a near-Earth asteroid. The winning mission will be selected in 2011 and it must be ready for launch before 2019.

Lab deaths: Police and internal investigations continue at India's Bhabha Atomic Research Centre — a nuclear-research facility near Mumbai — after the deaths of two young PhD students in an explosion and fire on 29 December. An analytical-chemistry laboratory was gutted by the blaze, but no radioactive material was involved, said Swapnesh Malhotra, a spokesman for the government's Department of Atomic Energy.

Primate-breeding dispute: A court has halted construction of a controversial facility in Guayama, Puerto Rico, that would breed primates for biomedical research. Bioculture, a company based in Mauritius that supplies primates to US and UK labs, should not have been awarded permits to build the facility on land reserved for agricultural use, a judge ruled in a statement published on 30 December. Bioculture says that it will appeal. Local citizens, concerned by the island's history of failing to contain primates brought there for research, filed the lawsuit, supported by US and UK antivivisection groups.

THE WEEK AHEAD

9–10 JANUARY

Last year's International Year of Astronomy officially ends with the 'Astronomy Beyond 2009' meeting in Padua, Italy.

▸ www.beyond2009.org

11 JANUARY

German chancellor Angela Merkel launches the United Nations' International Year of Biodiversity in Berlin.

▸ www.cbd.int/2010/welcome

12–15 JANUARY

Britain's Royal Society in London hosts the triennial conference of the Interacademy Panel on International Issues, a network of world science academies.

▸ go.nature.com/5YsESP

● BUSINESS

Buyout: Switzerland-based Novartis announced on 4 January a take-over of eye-care company Alcon, also based in Switzerland. It will pick up a 52% stake for $28.1 billion — adding to the 25% it bought for $10.4 billion in 2008 — offering other shareholders stock worth $11.2 billion. Novartis is diversifying in part to prepare for the 2011 loss of patent protection on its blockbuster blood-pressure drug Diovan.

BUSINESS WATCH

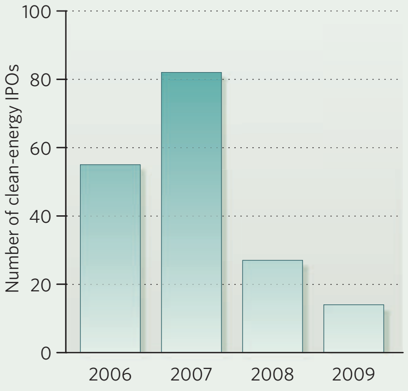
The new decade will see a bumper crop of clean-technology companies offering their stock to the public for the first time, analysts predict. It wouldn't be hard to improve on 2009, when few companies in the sector registered for an initial public offering (IPO).

London-based analysts New Energy Finance tracked only 13 clean-energy IPOs last year, compared with 82 in 2007 (see chart). Lithium-ion battery company A123Systems of Watertown, Massachusetts, was the highest-profile IPO of 2009, and, in December, solar firm Solyndra of Fremont, California, and biofuels start-up Codexis of San Francisco, California, both registered for IPOs. Five other solar firms made IPOs last year, compared with 13 in 2007 and 11 in 2008. New Energy Finance's Jenny Chase thinks solar IPOs will return to 2007 levels this year, with Miasolé, Nanosolar and Sunfilm probably among those in the thin-film sector. Tesla Motors, the electric-car firm based in San Carlos, California, is another company hotly tipped for a 2010 IPO.

"There's a certain amount of pent-up demand from public-sector investors," says Michael Holman of New York City-based Lux Research. But for venture capitalists and businesses seeking to offload their private-sector investments, most opportunities remain in mergers and acquisitions, he says.

AFTER THE IPO CRASH

Clean-energy public offerings may pick up in 2010



SOURCE: NEW ENERGY FINANCE

NEWS

New year, new science

Nature looks at what key events may come from the research world in 2010.

Planck peeks at the Universe's origin

The first detailed images of the cosmic microwave background sent back by the European Space Agency's Planck mission could alter theories about the origins and structure of the early Universe. Full results won't be officially released until 2012.

Life, but not as we know it

Surely this will be the year when genome pioneer Craig Venter and his team reveal they have booted up a laboratory-made genome inside a living bacterial cell, to create what will be billed as synthetic life.

An Antarctic time machine

An ice core from Antarctica could provide the sort of year-by-year climate records already gathered in the Northern Hemisphere from Greenland. The West Antarctic Ice Sheet Divide Ice Core project is in the final stages of pulling up a 3.4-kilometre-long climate record that covers the past 40,000 years in enough detail to compare how the north and south polar regions warm up, or chill, in relation to one another.

How many species will join the Rajah Brooke's birdwing butterfly on the protected list?

J. CARMICHAEL JR/NHPA

Stopping species loss

The United Nations has proclaimed 2010 the International Year of Biodiversity, to culminate in an October summit in Nagoya, Japan, that hopes to establish strategies to prevent biodiversity loss — probably by setting out ways to try to halt the current decline by 2050. New ideas are sorely needed: this year, 120 countries will miss a goal set by a 2002 accord to achieve a 'significant reduction' in biodiversity loss.

A flood of genomes

The completed Neanderthal genome and the genomes of remaining primates will count among the highlights of another year of ever-cheaper DNA-crunching. Following last year's comprehensive portraits of cancer genomes, medically minded sequencing will continue to focus on the causes of specific diseases, and on spotting more human genetic variants.

Mexico City: the new Copenhagen

Starting in late November, Mexico will be the venue for the next major round of United Nations climate-policy wrangling, where an overdue formal agreement to succeed the Kyoto Protocol on climate change may finally be hammered out. Before then, attention will focus on the action that individual countries

Shorter NIH grant form launches

For many, a new year means a fresh round of paperwork, but in the United States this year many biomedical researchers have something of a reprieve. The length of applications for most grants at the National Institutes of Health (NIH) — including mainstay 'R01' grants — has been slashed by more than half, from 25 pages to 12.

The streamlined form comes into effect this month and is part of a major overhaul of peer review at the agency, based in Bethesda, Maryland, which will this year fund some US$16 billion in research project grants. The aim, says Sally Rockey, acting deputy director of the NIH's office of extramural research, is to reduce the administrative burden on applicants and reviewers, and to focus on "the essentials of the science".

The shorter form "really forces applicants to concentrate on getting their point across more rapidly and to organize their application" accordingly, Rockey says. With the longer applications, she adds, "reviewers often focused on the minutiae within the methodologies instead of focusing on the overarching concept and impact".

To try to move away from this, other changes to the process include a new scoring scale for peer reviewers, plus 'enhanced' review criteria to emphasize the big picture over the details.

Reviews of the shorter application form are mixed as the 5 February R01 submission deadline approaches. Some researchers agree that its brevity encourages clarity. "It's asking us to be more focused and concise in our explanations," says Joan Teno, a professor of community health and medicine at Brown University in Providence, Rhode Island, who is seeking an R01 to refine a data-gathering tool she developed to measure the quality of US hospice care. "That's a really good thing."

"In the past, I would have easily put in at least ten figures. That's impossible now."

David Wieczorek, a molecular geneticist at the University of Cincinnati in Ohio, concurs that the abbreviated format tends to focus both the mind and the writing. But, he adds, "once you get past a certain [word] limit, it becomes very difficult: what do you include and what do you exclude?" Late last month, his draft application for a five-year grant to study mouse models of heart disease ran to 13 pages, and he was bracing himself to find a final page of trims.

The former 25-page length for R01 applications stood out among funding agencies and foundations worldwide. Project grant application forms at Britain's Wellcome Trust, for instance, are limited to 3,500 words plus up to five pages of figures; a recent informal sample of such applications found their average length to be 6–8 pages, including figures. The Investigator Awards that will replace project grants later this year at Wellcome

ANASPIDES PHOTOGRAPHY/I. D. WILLIAMS

CLUES TO TASMANIAN DEVIL CANCER
Tumours are found to arise from cells that insulate nerves, raising test hopes.
go.nature.com/wzHTgp

need to take on their commitments, on climate legislation in the United States and on international standards to monitor emissions and verify promised reductions.

Earth-like worlds elsewhere

As planet-hunters eagerly await the discovery of an Earth-like planet in the habitable zone around a Sun-like star, they may have to make do this year with an easier target: a potentially hospitable planet around a red dwarf star. NASA's Kepler telescope has already discovered previously unknown planets (see page 15).

Hope for HIV prevention

Early this year, the first clinical trial to use a gel incorporating an antiretroviral drug is expected to release its initial results; several large trials of other microbicides have failed to show benefit in blocking HIV. Early results are also due from long-anticipated trials that look at 'pre-exposure prophylaxis', or administering anti-HIV drugs before risky sex.

A perfect symmetry

Evidence for supersymmetry — the theory that every known fundamental particle has an undiscovered, superheavy partner — may be the most intriguing discovery to come from Europe's Large Hadron Collider near Geneva, Switzerland. The find would be even more bizarre than the anticipated Higgs boson, the particle thought to imbue matter with mass.

Quantum effects go large

Solid objects in physics laboratories could be seen to enter a superposition of states — the real-world version of Schrödinger's mythical cat that is dead and alive at the same time. The effect, predicted by quantum mechanics, has previously been seen in objects no bigger than ions, but could push into the macroscopic realm this year.

Cell reprogramming gets safer

Induced pluripotent stem cells will probably be created from adult cells using small molecules — lessening the risk of tumours, which comes with adding genetic material to a cell. Safer, more efficient reprogramming routes could lead to the field's first therapeutic applications.

Embryonic stem cells go clinical

The first clinical trials of therapies involving human embryonic stem cells (hESCs) could finally come this year. Biotech company Geron, of Menlo Park, California, plans to restart regulator-halted trials of an hESC-derived therapy for patients with spinal-cord injuries.

Space travel crosses frontiers

Among the year's planned space launches are Japan's Akatsuki, to orbit Venus, and China's second lunar probe, Chang'e 2. And as NASA looks set to choose a new direction for its human space-flight programme, a decision that could come early in 2010, the US space-shuttle fleet will make its final outings. These include the July launch of the Alpha Magnetic Spectrometer, an instrument to study cosmic rays for evidence of antimatter and dark matter.

X-rays with laser-sharp focus

X-ray free-electron lasers, which produce short pulses of coherent X-ray light, may start to assert their superiority over synchrotrons for imaging. They should enable researchers to make images of single biomolecules without having to crystallize them, and to create detailed movies of molecular events such as protein folding. Data will flow from the first of these facilities, at the SLAC National Accelerator Laboratory in Menlo Park, California.

Climate computing heats up

Expect increasingly realistic climate models from several recently launched supercomputers, including the Earth Simulator II in Yokohama, Japan, and Blizzard in Hamburg, Germany. As some of the world's 40 most powerful computers, they will improve on two of the largest uncertainties of current simulations: resolving local eddies in ocean circulation, and providing long-term forecasts of cloud behaviour. Blizzard will also incorporate Earth's carbon cycle into its climate models. ■

Richard Van Noorden

See also Opinion, page 26.

will have 8-page applications.

The NIH has been well aware of the problem. In 2008, a working group of internal and external experts strongly recommended shrinking the length of the R01 application. So did the broader research community: among more than 5,000 responses submitted to the NIH on the issue, 70% supported jettisoning the 25-page form in favour of something substantially shorter.

Last year's economic stimulus gave the agency a chance to pilot the abbreviated application by using it for Challenge Grant applications. Teno, who was a Challenge Grant reviewer, gives the updated process the thumbs up, singling out the clarity of the NIH's new instructions to reviewers. It "significantly reduces the workload and it improves the quality of doing the reviews", she says. "It used to be about five to six hours and I'm now down to three to four hours per application."

NIH

Sally Rockey is helping to reform the NIH grant process.

But Rita Nahta, a pharmacologist at Emory University School of Medicine in Atlanta, Georgia, is struggling with an issue presented by the shortened form that is likely to dog younger investigators: preliminary data take up precious space. And many of these can't be relegated to a list of references because, unlike data coming from more senior investigators, they are unpublished. "In the past, I would have easily put in at least ten figures," most unpublished, Nahta says. "That's impossible now."

Rockey says that the NIH has mechanisms in place to give young investigators special consideration — for instance, by dipping below the minimum percentile for grant approval to fund highly rated young scientists, and by clustering applications from young investigators for review against one other rather than against senior investigators.

Time will tell which of the changes work. "It's going to take several rounds of reviews before you are really going to know how successful this has been," says Wieczorek. The NIH launched a formal evaluation process in December, circulating an electronic survey of 35 questions to 4,500 investigators and reviewers, asking for feedback on already implemented changes to the peer-review process. It will ask for feedback as early as this summer on the changes in the length and structure of the application.

In the meantime, an overriding concern at the NIH is the prospect of a tidal wave of applications as researchers whose 12-page Challenge Grant proposals were rejected last year resubmit them next month as 12-page R01s, forcing the agency to turn down thousands of worthy projects. Last month, NIH director Francis Collins told his advisory committee that the drop in grant success rates that is likely to result is "the thing that keeps me awake at night more than any other". ■

Meredith Wadman

Chemists crack complex compound

Naturalistic approach vindicated as sponge molecule yields to synthesis in the lab.

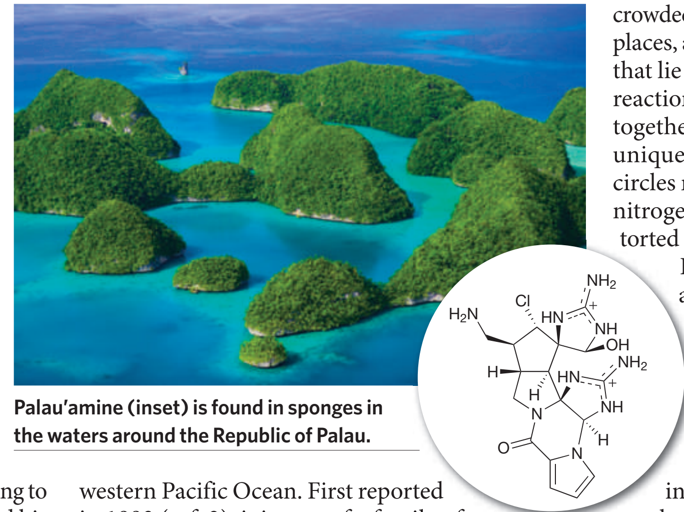
One of the most daunting challenges for synthetic chemists has finally been conquered. The effort to make palau'amine in the lab sparked heated competition for more than a decade between leading researchers, even though it may have little potential as a drug.

The yield of the 25-step synthesis, which was led by Phil Baran at the Scripps Research Institute in La Jolla, California, was just 0.015%: fewer than 2 in every 10,000 molecules of starting material made it through to the final product.

"Palau'amine is the pinnacle of technical difficulty," says organic chemist Patrick Harran of the University of California, Los Angeles, who has been trying to make the compound since 2002. "Phil and his students have set a standard against which all future work in the area will be judged."

But the synthesis, published last week in *Angewandte Chemie*[1], is more than a technical achievement. The procedure demonstrates the effectiveness of a set of guiding principles for efficient organic synthesis that was articulated by Baran's group several years ago and is now gaining adherents for its focus on brevity and simplicity.

Palau'amine was isolated from the sponge *Stylotella agminata*, which is found in the waters around the Republic of Palau in the western Pacific Ocean. First reported in 1993 (ref. 2), it is part of a family of compounds known as pyrrole-imidazole alkaloids, which may help to deter fish from snacking on the sponge or prevent microbes from taking up residence. The molecule has antitumour, antibacterial and antifungal activity at levels that are "OK, but not fantastic", says Matthias Köck, a marine natural-products chemist at the Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, Germany.



Palau'amine (inset) is found in sponges in the waters around the Republic of Palau.

B. KRIST/CORBIS

Thus, the main attraction in synthesizing the molecule is not its potential as a wonder drug, but the sheer challenge of making something so complex. The structure of palau'amine is crowded with spurs and joints in unusual places, and littered with nitrogen atoms that lie in wait to disrupt the chemical reactions used to stitch the compound together. At the molecule's heart lies a unique configuration of rings — two circles made up of carbon atoms and a nitrogen atom that are fused in a contorted configuration.

For more than a decade, chemists assumed that the rings were twisted into a bowl-like shape. But in 2004, Baran hypothesized that all the members of palau'amine's molecular family could be constructed with the same general strategy — implying that the accepted structure of palau'amine was wrong.

In 2007, his prediction was vindicated by three teams that independently worked out its true structure. Köck, who led the most detailed study, recalls that at first "almost no one believed us. Nearly everyone we spoke to thought we were misguided." But synthetic chemists soon switched their focus to the revised target. "Many groups had been chasing the wrong structure for years," says Köck.

Baran's synthesis adheres to a set of synthesis guidelines[3] that aims to exploit the target molecule's inherent reactivity and stay as close as possible to the way it is made in

Israeli government advisers threaten walkout

A rift in Israel's science establishment is threatening the country's long-term planning of civilian science. All 15 members of its National Council for Research and Development, now subordinate to the ministry of science, are poised to resign this month unless the council is given independent budgetary and administrative standing.

"If the council ceases to exist, we won't see any effects in the short term," comments Meir Zadok, director of the Israel Academy of Sciences and Humanities in Jerusalem. "But its responsibility is to aggregate information about research and development throughout the country and to look ten years ahead to see where the government needs to be involved."

In recent years the council has lent its support to efforts to increase research and university funding. Members have appeared in public forums and in the Knesset, Israel's parliament, to lobby for such funding.

The group was established in its current form in 2004 as part of the Israel Academy of Sciences and Humanities. In 2007, under the previous government, the council was transferred to the ministry of science when a science minister wanted to expand the ministry's responsibilities.

"The council's budget currently comes through the ministry of science, and it requires the approval of the ministry's officials for everything it does," says Meir Sheetrit, chairman of the science and technology committee in the Knesset. Council members chafed at being subject to the ministry's whims, and complained that their budget was being cut in favour of other areas under the ministry's purview, which included — until last spring — culture and sport.

In response, Sheetrit held hearings and drafted legislation to change the council's status to that of a government-run corporation. "Everywhere else in the world," he says, "national research councils are independent, with separate budgets, to ensure their objectivity."

But the government opposed the bill and, in mid-December, the Knesset rejected the legislation. As a result, council members plan to submit their resignations en masse in the next few weeks.

Council director Rony Dayan blames Daniel Hershkowitz, the

THE LOVE BUZZ
How mating mosquitoes harmonize their wing-beats.
go.nature.com/sqMeWx

J. GATHANY/CDC

nature, explains team member Ian Seiple. This includes using cascade reactions that can form many new chemical bonds in a single step, and avoiding the use of protecting groups to shield fragile parts of a molecule during synthesis because they increase the cost and complexity of the process.

Although none of the guidelines is new, applying them all within the same synthesis has become a hallmark of Baran's work. His goal is to prove that new drugs do not have to be built from the relatively limited pool of molecular motifs used by pharmaceutical companies.

The efforts to synthesize palau'amine have forced chemists to develop new reactions and techniques for assembling complicated molecules. Part of Baran's synthesis relies on a silver-based reagent, for example, that his lab invented to gently oxidize the half-built palau'amine molecule without disrupting its nitrogen atoms. That reagent is already being used by a pharmaceutical company to make a range of drug candidates, says Baran.

In the near future, he hopes to make grams of the compound instead of the few milligrams he has so far achieved, and to tweak his synthesis so that just one of the two possible mirror-image forms of the compound is produced. His team already has a working route that cuts ten steps from the beginning of the process. "For us, the story has just begun," Baran says. ■

Mark Peplow

1. Seiple, I. B. *et al. Angew. Chem. Int. Edn* doi:10.1002/anie.200907112 (2009).
2. Kinnel, R. B., Gehrken, H. P. & Scheuer, P. J. *J. Am. Chem. Soc.* **115,** 3376–3377 (1993).
3. Baran, P. S., Maimone, T. J. & Richter, J. M. *Nature* **446,** 404–408 (2007).

country's science minister, for torpedoing the legislation. "The officials in his ministry warned him that his small ministry would have trouble justifying its existence if it lost authority over the council," says Dayan.

Hershkowitz rejects that charge and says he supports independent status for the council. "The law that was submitted wasn't appropriate, however," he told *Nature*. "My goal is that the research council operates with independence, but there needs to be oversight to ensure proper management."

He declined to comment on how he would react to a full council resignation.

Hershkowitz says that he plans to draft legislation in line with his goals of continued ministry oversight. Sheetrit has already reintroduced his own bill, and plans to continue to push for its passage. ■

Haim Watzman

Kepler finds its first planets

WASHINGTON DC

Stars hum and throb, and the vibrations of this cosmic music could aid the NASA satellite Kepler in its goal of finding an Earth-like extrasolar planet.

On 4 January at a meeting of the American Astronomical Society in Washington DC, the Kepler team announced that it had identified five new planets. These are the first to be found by the 1-metre telescope, which stares continuously at one swathe of sky and looks for the dimming as a planet crosses a star and blocks some of its light. Hundreds more planet candidates await confirmation as the telescope gathers more data. These include some that orbit stars bright enough for their characteristic 'asteroseismology' vibrations to be detected, says Ronald Gilliland, a Kepler team member at the Space Telescope Science Institute in Baltimore, Maryland.

A precise understanding of these vibrations could allow astronomers to separate Earth-sized planets into two groups: those that are rocky and those that are watery, says Dimitar Sasselov, a co-investigator on the Kepler science team and an astronomer at Harvard University in Cambridge, Massachusetts. "It makes all the difference."

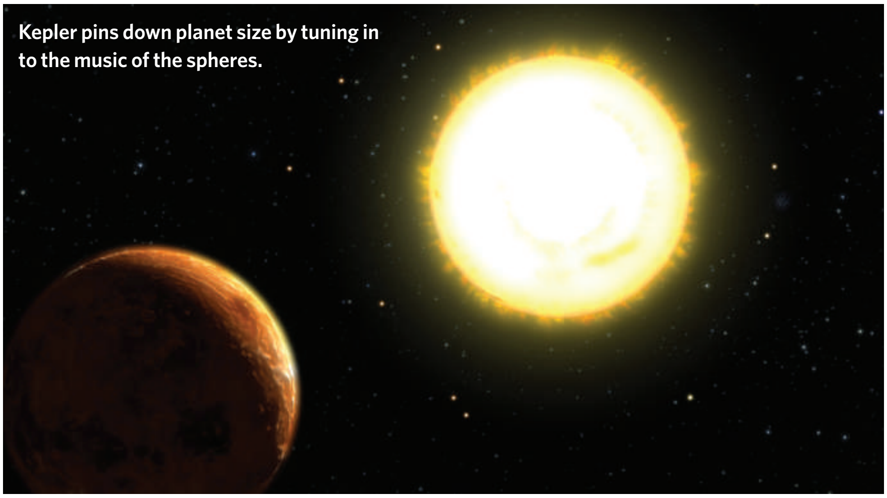
Because the core of an older star vibrates differently from that of a younger one, asteroseismology measurements can allow a precise determination of the age of the star system (and thus the planet). The data can also lead to a better estimate of the star's size — which in turn leads to more precision in the planet size. Gilliland says the extra precision could, when combined with ground-based measurements, help to determine the density of exoplanets as much as 50% better than before. Sasselov says that will be just enough of an improvement to discern the difference between a rocky planet like Earth, which is 0.06% water, and a water world like the recently discovered GJ 1214b, which is probably at least 50% water (D. Charbonneau *et al. Nature* 462, 891–894; 2009).

It is still early days for Kepler, which launched on 6 March 2009 from Cape Canaveral, Florida. The first five planets were discovered with just the first six weeks' science data, and they cross their parent stars repeatedly in short period orbits of a few days. Four are bigger than Jupiter — the largest planet in our Solar System — and one is about the size of Neptune.

Orbs more like Earth will be seen as the team shifts its attention to smaller planets in longer period orbits.

But Kepler only measures size. To understand density the team needs to measure mass as well, which comes from follow-up observations by ground-based astronomers. Sasselov says that even the giant 10-metre Keck telescopes in Hawaii lack an instrument sensitive enough to confirm an Earth-like planet if Kepler saw it. He is building a new instrument that he hopes to have installed on the 4.2-metre William Herschel Telescope in the Canary Islands by 2011 or 2012 — about the time when Kepler should have Earth-analogue candidates to check. ■

Eric Hand



Kepler pins down planet size by tuning in to the music of the spheres.

NASA/JPL-CALTECH

Floods linked to San Andreas quakes

Historical record underscores connections between reservoirs and seismic activity.

G. LUDWIG/CORBIS

A network of faults runs under California's shrinking Salton Sea, which was created in a flood a century ago.

SAN FRANCISCO
Geophysicists have linked historical earthquakes on the southern section of California's famed San Andreas fault to ancient floods from the nearby Colorado River.

The work has broad implications for understanding how floods or reservoirs relate to quakes — a topic that gained new relevance in 2008, after a massive earthquake in China's Sichuan province killed more than 80,000 people. Some geologists have proposed that impounding water behind a newly built dam there helped hasten the quake.

Now, new work in southern California suggests that at least three times in the past 2,000 years, the weight of river water spreading across floodplains seems to have helped trigger earthquakes in the region.

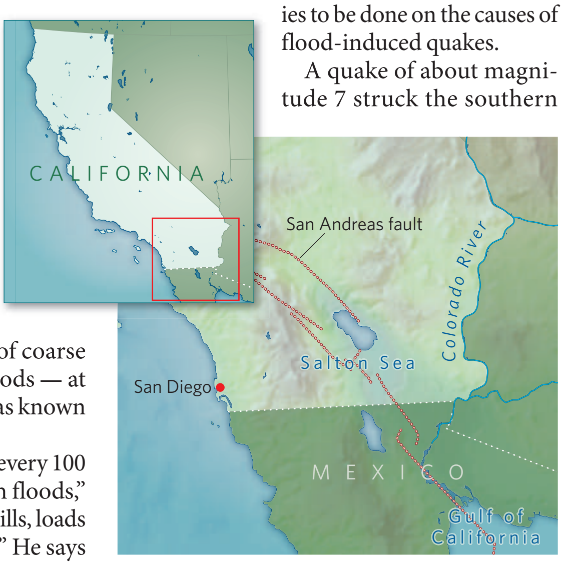
The findings could also be significant for the future of the Salton Sea, a lake about 200 kilometres east of San Diego that is the dwindling relic of a flood from a century ago. Various groups have long proposed pumping water back into the lake, from conservationists who want to restore wildlife habitat, to developers wanting to take advantage of the region's mild climate and scenic setting.

The latest study, presented last month in San Francisco at a meeting of the American Geophysical Union, suggests that water-management plans should take the potential earthquake risk into consideration.

Between 2006 and 2008, a team led by geophysicist Daniel Brothers of the US Geological Survey's Woods Hole Science Center in Massachusetts conducted the first detailed seismic survey of the Salton Sea, uncovering a previously unknown fault system running beneath the lake (D. S. Brothers *et al. Nature Geosci.* **2,** 581–584; 2009). The team subsequently analysed data from 20-metre-deep cores pulled from the lake bed in 2003 during earlier work for the US Bureau of Reclamation. The cores showed layers of coarse sandy material laid down during floods — at the same time that seismic activity was known to have occurred.

"We found quakes happened about every 100 to 200 years and were correlated with floods," says Brothers. "The Colorado River spills, loads the crust and then there is a rupture." He says the team is "very confident" in its evidence for the existence of three flood-derived quakes, of roughly magnitude 6, which happened about 600 years ago, 1,100 years ago and 1,200–1,900 years ago. "Sediments don't lie," he says.

The team has other, scant evidence for a quake that may have occurred soon after the 1905 flood that created the sea. Patrick Williams, an earthquake geologist at San Diego State University, says the group's work is "really smart", and calls out for more follow-up studies to be done on the causes of flood-induced quakes.



A quake of about magnitude 7 struck the southern San Andreas fault about 300 years ago; the next is a century overdue. One possible reason is the Hoover Dam: since its completion in 1936, the lower Colorado no longer floods.

A new scheme, however, could bring a further influx of water to the Salton Sea region. Developers are eyeing a route to pump water through low basins and river channels that link the Salton Sea to the Gulf of California (see map). The idea is being pushed by consultant Gary Jennings, whose Utility Solutions Group in South Lake Tahoe, Nevada, hopes to be involved in the project. In about a month, the Salton Sea Authority, a regional restoration agency, will consider a planning agreement to move forward with Jennings's 'sea-to-sea' plan.

Jennings, who had not heard of the seismic work when asked by *Nature*, calls the new findings "very insightful". "If anyone raises an earthquake flag," he says, "we would say don't increase the size of the sea."

"The new studies may cause some serious problems" for the project, adds Douglas Barnum, science coordinator for the Salton Sea at the US Geological Survey in La Quinta, California. Yet even if some form of water management goes ahead, he says, the new geophysical data gathered on the sea will still be important for understanding the region. ■

Rex Dalton

THE ENERGY STORAGE PROBLEM

Renewable energy is not a viable option unless energy can be stored on a large scale. **David Lindley** looks at five ways to do that.

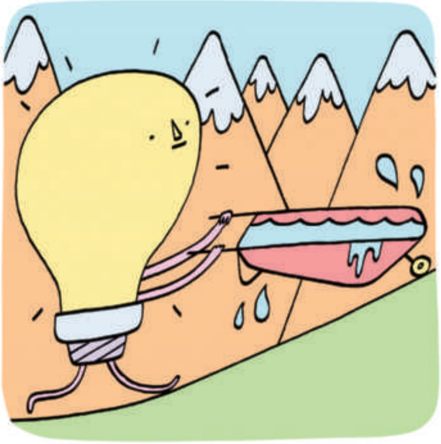
In February 2008, during a sudden cold snap, the normally relentless winds of west Texas fell silent and the thousands of wind turbines that dot that part of the state slowed to a halt. Local utility operators, unable to make up the shortfall with power from elsewhere in the grid, were forced to cut service to some users for up to an hour and a half before the winds picked up again.

That windless interval would have been a non-event if the utility companies had had a few hundred megawatt hours of energy stored away that they could draw on in emergencies. But they didn't. Ever ephemeral, electrical energy is difficult and expensive to store in large quantities.

The lack of good storage options has plagued utility operators for generations. Obligated to provide a steady supply of electricity to meet constantly varying demand, they have conventionally resorted to the costly and inefficient method of adjusting the output of a coal-fired plant, say, or by turning on a gas-powered 'peaker' plant during periods of high demand.

But that supply-based strategy is becoming less viable with the increased use of renewable energy sources with unpredictable output, notably solar arrays and wind farms. As the Texan example indicates, the power produced by these technologies is dependent on nature's whim, not human demand. "If we want to have a significant part of our energy come from renewable sources, storage is a must," says Ali Nourai, manager of energy storage at American Electric Power, a utility company in Columbus, Ohio, and chairman of the Electricity Storage Association, a trade association in Washington DC.

A number of technologies for energy storage already exist, including some that have been around for decades. The challenge is to make them robust, reliable and economically competitive — while matching the most suitable technology to each energy source or location. "Each technology has unique features," says Jillis Raadschelders, of energy consulting firm KEMA in Arnhem, the Netherlands. "There will never be a winning technology." Choosing the right technology means looking at each one in some detail.



ILLUSTRATIONS BY MATTHEW HODSON

PUSHING WATER UPHILL

The need for storage is particularly acute in densely populated northern Europe, where many countries are building offshore turbines to harness the winds blowing across the North Sea. Denmark already gets about 20% of its electricity from land- and sea-based wind farms, and it is aiming to increase that figure to 50% by 2025. Because the North Sea winds can drop to low levels for days at a time, however, countries such as Denmark and the Netherlands are increasing their grid connectivity to Norway, which gets the vast majority of its power from hydroelectric plants. Norway's mountain reservoirs provide back-up power capacity, and also offer substantial amounts of pumped storage hydroelectricity, in which water is pumped uphill to a reservoir using surplus electricity, and released downhill again to turn a generator when power is needed. Pumped hydroelectricity has a storage efficiency of 70–85%, and it is the most mature and widespread technology being used for large-scale electricity storage. China, Japan and the United States, for example, have numerous installations with generating capacities ranging from tens of megawatts (MW) to several gigawatts (GW). Pumped storage hydroelectricity is a particularly good match for wind power because water pumped into an upper reservoir will stay there for a long time, making up for potentially large gaps in wind generation.

But in its conventional form, pumped storage hydroelectricity requires mountains, so opportunities are limited by geography. Building such storage also tends to be expensive and environmentally destructive, and installing high-voltage transmission lines to connect remote storage sites to grids often triggers opposition on environmental grounds.

If the capacity of pumped storage hydroelectricity is to grow significantly, it will have to leave the mountains. One innovative concept by KEMA would put wind turbines and pumped hydro in the same place: an 'energy island' in a shallow part of the North Sea. An area of perhaps 60 square kilometres would be ringed off by a dyke or levee to create an artificial lake. Wind turbines would stand on the encircling dyke, and any excess power would be used to pump water out of the lake and into the surrounding sea. Letting sea water flow back in would regenerate the stored electricity. In the absence of wind, KEMA estimates that the energy island could supply an average of 1,500 MW for as long as 12 hours.

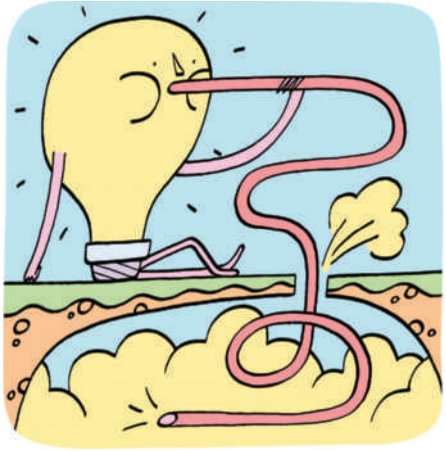
"If pumped storage hydroelectricity capacity is to grow significantly, it will have to leave the mountains."

SQUEEZING AIR UNDERGROUND

In the farmlands near Huntorf, Germany, about 100 kilometres southwest of Hamburg, an ordinary-looking industrial installation performs an unusual task: when demand for electricity in the local grid is low, the plant uses excess power to compress air and pump it into two underground salt caverns with a combined volume of more than 300,000 cubic metres. Then, at times of high demand, the compressed air is allowed to expand through turbines on the surface to regenerate the electricity.

The Huntorf plant, which has been working since 1978, can supply almost 300 MW of reserve power for up to three hours, and comes into operation about 100 times a year. But it has not exactly spawned a legion of imitators. A similar but smaller plant in McIntosh, Alabama, came online in 1991, and efforts to build another such system in Iowa, begun in 2002, are only now at the point of acquiring land for test drilling.

The problem is that these compressed-air energy storage (CAES) facilities are considerably more complex in practice than they are in principle. Gas heats up when it is compressed, which limits how much air can be pumped underground before it becomes too hot to be stored safely. Moreover, the longer that hot air is left in place, the more of its heat — which represents a substantial fraction of the input energy — is lost into the walls of the surrounding cavern. And then when it is released again, the expanding air cools down. In the Huntorf and McIntosh facilities, in fact, the released air is fed into a standard natural gas turbine, boosting its efficiency. So the net effect of the air-compression system is to boost the efficiency of a more or less conventional natural-gas-fuelled power plant.



In the near term this kind of hybrid system "makes a great deal of sense", says Haresh Kamath, a researcher at the Electric Power Research Institute (EPRI) in Palo Alto, California, especially as more electricity from renewable sources is available to recharge the system at night. Looking further down the road, however, the EPRI and others are also researching improvements that would turn CAES into a true energy storage system, requiring no fossil fuel. Such an 'advanced adiabatic' system would capture and store the heat of compression and then use it to reheat the released air, which would spin a turbine directly without any additional fuel. Metal foundries and blast furnaces have for years captured waste heat in stacks of refractory bricks or similar materials, says Christoph Jakiel, a researcher at MAN Turbo in Oberhausen, Germany. So applying the technique to compressed-air storage should be straightforward.

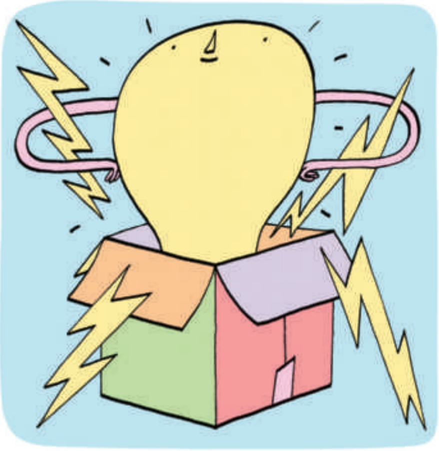
He estimates that the efficiency of such a system to be something under 80%, comparable with pumped storage hydroelectric systems. The overall costs of construction and operation would also be about the same. Suitable locations should not be hard to find in most regions of the world, says Jakiel. Salt caverns are not uncommon, and the proposed Iowa Stored Energy Park, should it ever be built, will pump compressed air into an aquifer.

ELECTRICITY IN A BOX

Large-scale battery storage would be a solved problem already if utility companies could use the ubiquitous lead-acid technology that has been the basis of car batteries for nearly a century. Unfortunately, lead-acid batteries have a low energy density — they are bulky and heavy for the amount of energy they store — and they do not stand up well to repeated charge–discharge cycles.

A better solution is the sodium–sulphur (NaS) battery, which stores energy by chemically dissociating sodium polysulphide into sodium and sulphur. The energy can then be released by allowing the two elements to react again. NaS batteries have a high energy density and can last through thousands of charge–discharge cycles. Their chief drawback is that the sodium and sulphur have to be kept in separate reservoirs in the molten state, at about 300 °C. Also, the batteries suffer irreparable damage if they discharge completely and grow cold. The resulting need for a robust container, along with other technical requirements, means that NaS batteries cost about US$3,000 per kilowatt (kW) of available power. That compares unfavourably with standard gas-powered plants, which cost about $1,000/kW. Nonetheless, NaS batteries have been developed commercially by NGK Insulators in Nagoya, Japan. Japan now has an installed capacity able to supply its grid with about 300 MW when extra power is needed, for up to six hours at a stretch. Other countries are also picking up the pace. The United States, for example, has about 10 MW of NaS capacity in place and a similar amount on the way, led by companies such as American Electric Power and Xcel Energy in Minneapolis, Minnesota.

In the future, large-scale NaS storage could face a challenge from lithium-ion technology. Already in widespread use for mobile phones and laptops, and under development for electric cars, lithium-ion batteries have a high energy density and efficiencies of more than 90%. Their big drawback is cost, which is in part driven by safety considerations: the batteries use a lithium salt in an organic solution, which is flammable, necessitating robust construction to minimize fire hazards. Lithium-ion batteries made for consumer electronics currently cost a few hundred dollars per stored kW hour. But for widespread vehicle applications, that cost must come down closer to $100 per kW hour, and for grid applications it needs to be lower still.



Yet Nourai, for one, remains optimistic. Safety issues are more easily and cheaply met for batteries in secure, fixed installations than in hand-held devices, he says. And in Asia, especially, there is strong support for lithium-ion technology and keen competition between manufacturers, which he hopes will lead to dramatic cost reductions. In China, he recently saw a cargo-container-sized lithium-ion installation, and expects to see capacities of a megawatt or more in the coming years.

At the Massachusetts Institute of Technology in Cambridge, materials chemist Donald Sadoway is trying a more radical approach to reducing the cost. "I want a battery that's dirt cheap," says Sadoway, "and the way to do it is to build it from dirt" — that is, from the most abundant elements in Earth's crust. Although there is little new to discover about the electrochemistry of these elements — such as silicon, iron and aluminium — a battery involves two reactions, one at each electrode, along with an electrolyte that supports the appropriate ion transfer. That makes for a huge and largely untested combination of possible compounds and reactions to search through. The quest is made feasible, says Sadoway, by supercomputers that can quickly assess proposed battery chemistries, freeing researchers from the need to synthesize and test actual materials. In the coming decade, he says, "I'm optimistic that the rate of discovery will accelerate."

TAKING ELECTRICITY FOR A SPIN

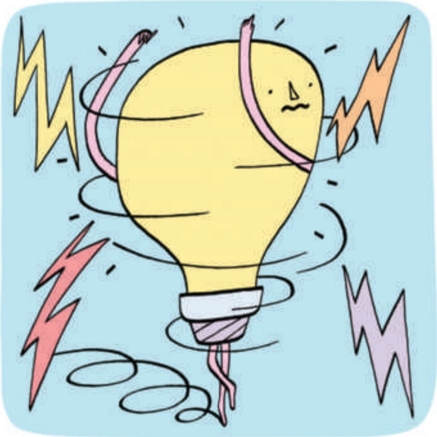
Conceptually, at least, one of the most straightforward ways to store energy is in a spinning flywheel: electrical energy gets converted into the kinetic energy of rotation by running it through a motor, which accelerates the flywheel. And the kinetic energy is extracted when it is needed by coupling the flywheel to a generator, which slows the wheel down and produces electricity.

Again, however, the reality is more complex — the flywheel has to spin very fast yet be strong enough to keep from flying apart. Flywheel storage systems are commercially available as uninterruptible power supplies that can deliver modest amounts of power for seconds or minutes, but they are not competitive for the longer storage times needed by the electric utility companies.

One big advantage of flywheels is that they can absorb the energy within seconds or minutes, and give it back just as quickly. This is exactly what is needed for regulating the frequency of a power grid, which is supposed to be maintained at an even 50 or 60 cycles per second, depending on the country, but which tends to drop whenever short-term increases in the load cause the turbines to slow down. Keeping it stable is a challenge for utility companies everywhere.

With that in mind, Beacon Power of Tyngsboro, Massachusetts, has spent the past decade developing a high-tech flywheel that is optimized for frequency regulation. Measuring about 2 metres tall and 1 metre in diameter, the flywheel consists of a cylindrical aluminium core, which houses the motor and generator, and a carbon-fibre composite rim. It is suspended on magnetic bearings inside a vacuum-sealed chamber, where it can spin at up to 16,000 revolutions per minute. The devices are designed to run for 20 years or more with no maintenance, says Matthew Lazarewicz, Beacon's chief technical officer. They can store energy with an efficiency of 85%, he says, and can spin up and down for perhaps millions of cycles during their working life, making them far more durable than batteries.

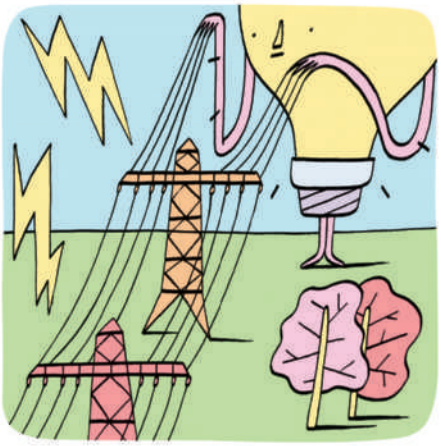
The challenge now is to bring the cost down, which Beacon hopes to do thanks to a project it has recently undertaken with loan guarantees from the US Department of Energy. In Stephentown, New York, Beacon has begun construction of a $70-million, 20-MW, 200-wheel flywheel farm that will help to regulate frequency in the regional power grid. The budget includes a number of one-time costs related to establishing its qualifications for the federal loan guarantees. The company estimates that future plants of this size will cost less than $50 million — a price it hopes to bring down to about $30 million. In late November, the energy department awarded Beacon $24 million for half the cost of a new 20-MW plant to be built outside Chicago, Illinois.



INTEGRATING WITH A SMART GRID

There are a number of even more exotic technologies that could become candidates for large-scale energy storage — assuming that researchers can eventually get the cost down to a competitive level. Examples include 'ultracapacitors' that can store huge amounts of electrical charge in atoms-thick layers next to the electrodes, and coils of superconducting wire able to store large amounts of circulating current indefinitely.

But by far the most cost-effective approach to large-scale electric energy storage is to minimize the need for it. That is one of the goals set out earlier this year in the US stimulus bill, which allocated $4.3 billion to research and development in renewable-energy generation, energy efficiency and, most especially, a 'smart grid'. Instead of simply adjusting the supply of electricity in response to the vagaries of unpredictable demand, a smart grid would constantly adjust demand as well. When demand hits a peak, for example, the grid might start cutting power for household refrigerators, office air-conditioning systems and other non-urgent uses — just for a moment in each case, and nothing that anyone would notice, but enough to smooth out variations in the overall load.



In that kind of system, says Nourai, storage and smart-grid technologies would work together, evening out the usual peaks and troughs in grid load to a greater extent than either could achieve alone. "Variation is never going to go away but with storage it can be much flatter," he says. He sees a future in which even fairly small communities could be "net zero", meaning that on average they make as much electricity as they need, and maintain a reliable supply by exchanging modest amounts of power back and forth with neighbouring communities. Local interconnections would be low-voltage lines, and long-distance high-voltage lines would be needed only to connect wind farms or solar arrays in remote areas with populated regions. That transformation, Nourai says, "will change the way we think about, run and plan" the storage of electricity. ■

David Lindley is a freelance writer in Alexandria, Virginia.

PHOTOS BY M.SCHROPE

Wish you were here

An annual excursion to an exclusive Caribbean island has yielded an impressive body of ecological fieldwork. Just don't call it a holiday, says **Mark Schrope**.

It's supper time on Guana Island in the British Virgin Islands and a dozen diners are relaxing on a candlelit veranda atop a cliff overlooking the Atlantic Ocean. They're enjoying a four-course meal and several bottles of wine. The privilege of staying at this very private resort — the only sign of civilization on the 340-hectare island — typically costs guests between US$700 and $8,000 per night. But these aren't typical guests. They're biologists, and if the constant talk of taxonomy isn't enough to prove it, they have physical evidence. After the meal, zoologist James 'Skip' Lazell (pictured above), produces from his shirt pocket a small bag containing a live snake, *Liophis exigua*. It had been captured that evening and would soon be measured and returned to where it was found.

Lazell, president of the non-profit organization the Conservation Agency in Jamestown, Rhode Island, has been coming to Guana every year for nearly three decades to lead a study of this distinctly unspoiled island's flora and fauna. He and a rotating crew of collaborators have produced what is arguably the most comprehensive, long-term record of the natural history of an island in the Caribbean, where high-volume tourism and frequent catastrophic weather shape the ecosystem. They have revealed a remarkably diverse ecological cast: including some 12 species of reptiles, 100 birds, 160 fungi, 330 plants and hundreds of insects, several of them new.

"I can't really think of another place where exactly these kinds of data for a whole bunch of species all at the same time have been gathered year after year," says Daniel Simberloff, an ecologist at the University of Tennessee Knoxville. That record has lent itself to increasingly sophisticated ecological modelling approaches that may ultimately help to predict the regional effects of global climate change. Moreover, the work challenges the prevailing theory on what drives island biodiversity.

> "Guana is the least screwed up of all the islands on the Greater Puerto Rico Bank."
> — James Lazell

Development on the island has been minimal, making the place attractive not only to the 30 or so guests the island can house at any given time, but also to scientists: the island offers a relatively unspoiled baseline for healthy Caribbean plant and animal life. "It is in the best ecological condition and the least screwed up, certainly, of all the islands on the Greater Puerto Rico Bank," Lazell says. Most other islands in the region have undergone significant development and human settlement for at least some portion of their history.

Welcome to the island

But on Guana, before the resort there was only a short-lived Quaker plantation in the 1700s and a native American settlement centuries before that. Gloria Jarecki, who with her husband, Henry, bought Guana in 1975, says the plan had always been to maintain as small a footprint as possible, "and try to pass [the island] on in the condition we found it in or better". So they were receptive when Lazell first proposed a long-term study of the island in the early 1980s. A foundation run by the Jareckis covers the cost of the team's accommodation in October, during tourism's low season for the area. Scientists typically only have to pay for travel. (The author of this article was similarly accommodated.)

One can be forgiven for assuming, as some of their colleagues have, that the whole enterprise is a tropical junket. In addition to the dinners, the team meets most afternoons on a palm-lined, white-sand beach with a well-stocked bar. "It's about as cushy as you can get on the domestic end," says Clint Boal, an ornithologist with the US Geological Survey based at Texas Tech University in Lubbock, and long-time Guana team member. But the days on Guana are anything but pampered. The vast majority of the island is accessible only by rugged footpaths, some of which descend precipitously down boulder-strewn gulches. The scientists typically slog kilometres every day with their gear, conducting general surveys or working at regular study plots. "It is a very difficult place to work in many ways," Boal says.

For the first few years, the goal was mainly to catalogue the plant and animal species. In parallel with the survey work, Lazell also worked with the owners to preserve the ecosystem and restore it as much as possible to its pre-human condition, recommending efforts such as limiting the population of wild sheep and removing cats, dogs and invasive plant species such as Australian pines. He also established a programme to replace animals thought lost, based on historic and archaeological records, including flamingos, red-legged tortoises, and critically endangered stout iguanas.

The team also watches island populations. For birds and reptiles, the researchers measure survival rates and demographics — data that could reveal the factors most important to maintaining a healthy Caribbean island. To monitor the birds, Boal and his colleagues spread lightweight mist nets across bird thoroughfares. After weighing captured birds and taking other measurements, researchers mark them with leg tags.

That doesn't work for reptiles, however, as many have legs that grow substantially and some don't have any. So, the team began implanting radio-frequency identification devices in 2001. Brent Bibles, a population biologist at Utah State University in Vernal, points to an adult stout iguana (*Cyclura pinguis*) on a trail near the resort. These reintroduced animals now number in the hundreds, from the eight that were brought to the island in 1984 (J. D. Lazell *Island: Fact and Theory in Nature* Univ. California Press, 2005).

The team hopes soon to begin recapturing adults. Bibles, who has bite marks all over his right hand from an encounter with a metre-long Puerto Rican racer snake (*Alsophis portoricensis*), is excited by the prospect. But at up to 30 kilograms, the iguanas, he says, are, "kind of nasty to handle".

Barry Valentine, a retired entomologist from Ohio State University in Columbus, has been collecting insects since he was ten. In 1946 he joined a group of adventurous undergraduates at the University of Alabama in Tuscaloosa. In his book *Naturalist* (Island Press, 1994), Edward O. Wilson describes chasing down specimens while hanging off the hood of Valentine's car. Still, it's difficult to imagine Valentine any more exuberant in his work than he remains today at 85. On a clear night, Valentine flips on a black light at his cottage and darts after interesting visitors.

Tropical futures

In recent years Valentine has found four new beetle species, but also some puzzling shifts in the number and types of various insect groups on the island. Most years he collects about 40 ground beetles, for instance; this year he found just six. He is working with others in the Guana team to see whether they can spot corresponding trends between weather patterns and animal populations. "There are always questions being raised," says Valentine, "Which is one of the things that makes it so much fun."

Some species of birds show troubling patterns with fewer young and lower body mass. Because the Lazell team's data set stretches back so far and covers so many groups, they can look for explanations for shifts in populations. "That's exactly the kind of data you need to demonstrate the impacts of climate change," says Simberloff.

A team including Barry Valentine (top) catalogues insects, reptiles and birds on Guana Island.

Boal, Bibles and others are looking at rainfall data from the National Oceanic and Atmospheric Administration to see if they can spot correlations between drought years and population declines, for example. Others are looking at water retention in the reptiles on the island for clues to how a changing climate could affect animals.

To Lazell, though, the most important application of the Guana data set, along with his related, but less extensive, studies on other islands around the world, is to inform the ongoing debate over whether mathematical formulae can reliably predict the diversity on a given island. One of the most prominent theories, proposed in the 1960s by Wilson and Robert MacArthur (R. H. MacArthur & E. O. Wilson *The Theory of Island Biogeography* Princeton Univ. Press, 1967), correlates the number of species on an island with its area. But the number of species on Guana far exceeds the theory's predictions, even ignoring reintroduced species. "We took the MacArthur–Wilson species–area notion and blew it out of the water," says Lazell, who is not known for mincing his words.

Lazell argues that human impact, although difficult to quantify, may be the main controlling factor of diversity. MacArthur and Wilson "were looking at islands that were just wrecked", he says, and that skewed their results. Overall, Lazell says that ecology may be too complex for broadly applicable formulas. Gad Perry, a herpetologist and conservation biologist at Texas Tech who helps to coordinate the Guana programme, has similar suspicions. "There are people that say this is not rocket science, but rocket science is so much simpler," he says. "Our systems have a lot more variables and that's what makes them interesting."

Simberloff, who was a student of Wilson's, agrees that a universal theory of biodiversity will probably remain elusive. "I don't think anyone would say that Guana resolutely rejects any prominent theories," he says, "but it could be that in a few years people reviewing literature on some particular theory might use Guana as one of several examples to say, 'Look, it doesn't hold up as strongly as we hoped.'"

For his part, Lazell says that more studies like those conducted at Guana may be needed. But, he says, "there's not much more I can do in one lifetime". Although he's crisscrossed the island more times than he can count, his health now confines him mainly to the resort, where he spends most of his time facilitating the work of others, and reigning as the self-proclaimed "Curmudgeon in Chief". As for Guana itself, with its trails through virgin forest and its collegial, candlelit dinners, his conclusions are difficult to argue. "This is my favourite place to do work," he says, "You can't beat it." ■

Mark Schrope is a freelance writer based in Florida.

Tomorrow never knows

Science should focus more on understanding the present and less on predicting the future, argues **Daniel Sarewitz**.

On humid summer mornings weather forecasters will often predict an afternoon thunderstorm, but sensible people know they probably won't have to cancel plans for a picnic. The forecast may be accurate, but people's understanding of how to interpret and contextualize the information is what makes it valuable to them.

Indeed, weather predictions are uniquely useful and useable. The US National Weather Service issues millions of them annually, affording continual opportunity for assessing and improving performance. Organizations that communicate forecasts have learned to tune information to diverse users, be they average citizens, ship captains or the airline industry. The predicted events are discrete and familiar, and they occur soon after the prediction, so that response options — to delay the flight, or not — are unambiguous, and disagreements have little time to emerge.

When these qualities are lacking, however, big challenges to decision-making ensue. In February 1997, the weather service predicted that the Red River would reach a record flood crest of 15 metres in Grand Forks, North Dakota. The town prepared for that height, but in April the crest passed 16 metres, and the result was US$1.5 billion in flood damage. The prediction was within the historical range of forecast accuracy (R. A. Pielke Jr *Appl. Behav. Sci. Rev.* **7**, 83–101; 1999), but information providers, communicators and users lacked the experience and judgement to respond appropriately to the prediction.

In the wake of the disaster, residents were willing to abandon low-lying areas of town as part of a new $409-million flood-control system for floods of up to 19 metres. What had they learned from their experience? Not to depend on predictions. And when high floods struck again last March, the town stayed dry.

False belief

Predictions are not instructions that people simply follow to make better decisions. They are pieces of an intricate puzzle that may sometimes contribute to improved decisions. For complex, long-term problems such as climate change or nuclear-waste disposal, the accuracy of predictions is often unknowable, uncertainties are difficult to characterize and people commonly disagree about the outcomes they desire and the means to achieve them. For such problems, the belief that improved scientific predictions will compel appropriate behaviour and lead to desired outcomes is false.

This conclusion flies in the face of the instincts and interests of scientists and decision-makers. Scientists are attracted to the intellectual challenge of making predictions, and recognize that promising to provide predictions appeals to the interests of the policy-makers who fund them. And decision-makers would prefer to hand over responsibility for the future to scientists — who would also take the blame when wrong.

For example, regional climate predictions are now being offered by scientists as a next logical step in applying science to the global-warming problem. As explained on the website of the Climate Variability and Predictability project of the World Climate Research Programme: "The increased confidence in attribution of global scale climate change to human induced greenhouse emissions, and the expectation that such changes will increase in future, has led to an increased demand in predictions of regional climate change to guide adaptation." The seductive but dangerous logic is driven by the confluence of the "increased demand" of decision-makers, and the high-prestige science of climate modellers who believe that society needs more of what they've been doing anyway (see *Nature* **453**, 268–269; 2008). But this logic confuses the distinct tasks of bringing a problem to public attention and figuring out how to address the societal conditions that determine the consequences of the problem.

"The right lessons for the future of climate science come from the failure to predict earthquakes."

Hurricane Katrina in 2005 provides cautionary insight. The likelihood of such a storm had been appreciated for decades and, in the days leading up to the disaster, the storm's path was accurately predicted. But New Orleans's fate had long been sealed by a lethal combination of socioeconomic and racial inequity, regional environmental degradation, unwise development patterns and engineering failure. Science had delivered ever more knowledge about regional climate behaviour and ever more accurate hurricane-track predictions, but this was not what the city needed to avoid catastrophe.

In contrast, from a societal perspective, perhaps the best thing that ever happened in the field of earthquake research was the recognition that earthquake prediction was likely to be impossible. In recent decades, the priorities of the US Geological Survey's earthquake-hazard programme have moved away from prediction and towards the assessment, communication and reduction of vulnerabilities. This evolution has demanded closer collaboration between scientists and diverse regional and state decision-makers, to provide information that can help improve construction practices, land-use decisions, disaster-response plans and public awareness.

If wise decisions depended on accurate predictions, then in most areas of human endeavour wise decisions would be impossible. Indeed, predictions may even be an impediment to wisdom. They can narrow the view of the future, drawing attention to some conditions, events and timescales at the expense of others, thereby narrowing response options and flexibility as well.

This difficulty is on spectacular yet unacknowledged display in the climate-change arena. The recently concluded United Nations climate-change conference shows that the world's attention is focused on global warming, but also that clear progress towards addressing the problem is incredibly difficult to achieve. A central obstacle is that predictions of long-term doom have created a politics that demands immense costs to be borne in the near term, in return for highly uncertain benefits that accrue only in a dimly seen future.

Science could help untangle this politically impossible dilemma by moving away from its obsession with predicting the long-term future of the climate to focus instead on the many opportunities for reducing present vulnerabilities to a broad range of today's — and tomorrow's — climate impacts. Such a change in focus would promise benefits to society in the short term and thus help transform climate politics. Strange as it may seem, the right lessons for the future of climate science come not from the success in predicting thunderstorms, floods and hurricanes, but from the failure to predict earthquakes. ■

Daniel Sarewitz, co-director of the Consortium for Science, Policy and Outcomes at Arizona State University, is based in Washington DC.
e-mail: dsarewitz@gmail.com.

See go.nature.com/ILx8PC for more columns.

CORRESPONDENCE

Climate e-mails: man's mark is clear in thermometer record

We welcome debate about the ethics of science prompted by the language of some of the hacked e-mails from the UK Climatic Research Unit (*Nature* **462,** 545; 2009). Rightly or not, this has created concerns about the scientific process. But it is critical to point out that no grounds have arisen to doubt the validity of the thermometer-based temperature record since 1850.

Both the detection of climate change and its attribution to human activities rely on the thermometer-based temperature record (compiled by the Climatic Research Unit and other institutions). They do not rely on proxy reconstructions of temperature over the past millennium, which are based on indirect evidence such as tree rings. Reconstructions contribute less to our understanding of climate than the thermometer record because of uncertainty both in these reconstructions and in the drivers of climate change before the twentieth century.

Unfortunately, the mainstream media have confused the two. The thermometer record shows unequivocally that Earth is warming, and provides the main evidence that this is caused by human activity. This crucial record remains unchallenged.

Commentators have suggested that the e-mails disclose a 'team mentality' among prominent climate scientists. Some people may have gone too far in promoting particular viewpoints, so an independent enquiry and open discussion should help to re-establish public confidence. However, it is absurd to suggest that there is some kind of global conspiracy involving all climate scientists.

We and our colleagues have worked with the scientists at the centre of this controversy. We have examined, used and at times criticized their data and results — just as they, at times, have criticized ours. Our disagreements have no bearing on our respect for other aspects of their work.

Hans von Storch Institute for Coastal Research, 21502 Geesthacht, Germany
e-mail: hvonstorch@web.de
Myles Allen Department of Physics, University of Oxford,
Oxford OX1 3PU, UK

Climate e-mails: lack of data sharing is a real concern

Your Editorial (*Nature* **462,** 545; 2009) castigates "denialists" for making "endless, time-consuming demands for information under the US and UK Freedom of Information Acts". But you do not mention the reason — that the Climatic Research Unit at the University of East Anglia has systematically tried to avoid revealing data and code.

Science relies upon open analysis of data and methods, and the UK Natural Environment Research Council (NERC) has a clear data-sharing policy that expects scientists "to cooperate in validating and publishing [data] in their entirety". The university's leaked e-mails imply a concerted effort to avoid data sharing, which both violates the best practice defined in NERC policy and prevents verification of the results obtained by the unit. Asking for scientific data and code should not lead to anyone being branded as part of the "climate-change-denialist fringe".

David R. Bell Molecular Toxicology, School of Biology, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: david.bell@nottingham.ac.uk

Step up aspirations to save biodiversity as 2010 begins

The Convention on Biological Diversity's post-2010 targets are likely to aim for a halt in biodiversity loss by 2050 and for 'more modest' interim targets for 2020 (*Nature* **461,** 1037; 2009). Global aspirations need to be much higher than this to avert the accelerating and catastrophic decline in the variety of life forms on Earth.

A 2050 vision should aim both to arrest this loss and to restore the populations, habitats and ecological cycles that support biodiversity and ecosystem services. A 40-year horizon should be about right, given that the restoration of forests, wetlands, coral reefs and other habitats depends on species and processes that often have decades-long generational periods.

We should, at the very least, aim to maintain biodiversity and the health of ecosystems as they are now — in particular, by setting an intermediate target to prevent further extinctions.

The deadline for achieving 'more modest' targets should be 2015, not 2020. That would synchronise it with the Millennium Development Goals (J. D. Sachs *et al. Science* **325,** 1502–1503; 2009) and with the timeframe of political cycles, which would help to ensure that elected politicians successfully deliver the target to their constituencies.

Ashok Khosla International Union for Conservation of Nature (IUCN),
22 Olaf Palme Marg,
New Delhi 110057, India
e-mail: president@iucn.org
Julia Marton-Lefèvre IUCN,
Rue Mauverney 28, Gland 1196,
Switzerland

Icelandic genetic database not at risk from bankruptcy

Although the Icelandic genomics company deCODE has filed for bankruptcy, this does not, as you put it in your News story (*Nature* **462,** 401; 2009), leave the "fate of its valuable genetic database unclear".

As chief executive of deCODE, I can state that its Iceland-based subsidiary Islensk Erfdagreining continues to perform all of the company's human genetics work, managing its population resources, conducting its research and services, and processing its tests and genome scans.

It is Islensk Erfdagreining's scientists and laboratories that are licensed to undertake this work. We continue to operate under the same data and privacy protections as usual, rooted in the Icelandic community and within a tried and tested regulatory environment.

Nor should scientists be "lamenting the prospect of losing deCODE's vast database of genetic and medical information". Islensk Erfdagreining will probably be sold to another group of investors as a going concern. Such a change in ownership of the operating company will have no bearing on the terms under which Islensk Erfdagreining manages and analyses genetic samples and data.

Islensk Erfdagreining does not own these samples or the data. They are owned by the individuals who provide them and are only used for the specific purpose, whether research or testing, agreed upon with those individuals and under the regulatory protections under which we work.

These resources cannot therefore be sold and are not for sale. The genetics operation of Islensk Erfdagreining cannot be put in a box and dispatched elsewhere.

Kári Stefánsson deCODE genetics,
Sturlugata 8, 101 Reykjavik,
Iceland
e-mail: kstefans@decode.is

Contributions may be submitted to correspondence@nature.com. Please refer to the Guide to Authors at go.nature.com/cMCHno. They should be no longer than 300 words. Published contributions are edited. Science publishing issues are regularly featured at http://blogs.nature.com/nautilus, where we welcome comments and debate.

OPINION

2020 visions

For the first issue of the new decade, *Nature* asked a selection of leading researchers and policy-makers where their fields will be ten years from now. We invited them to identify the key questions their disciplines face, the major roadblocks and the pressing next steps. Visit go.nature.com/htW8uM to respond and to add your vision.

ILLUSTRATIONS BY JESSE LEFKOWITZ

Search

Peter Norvig*
Director of research at Google

Internet search as we know it is just one decade old; by 2020 it will have evolved far beyond its current bounds. Content will be a mix of text, speech, still and video images, histories of interactions with colleagues, friends, information sources and their automated proxies, and tracks of sensor readings from Global Positioning System devices, medical devices and other embedded sensors in our environment.

The majority of search queries will be spoken, not typed, and an experimental minority will be through direct monitoring of brain signals. Users will decide how much of their lives they want to share with search engines, and in what ways.

The results we get back will be a synthesis, not just a list. For example, today if I ask 'compare approaches to nuclear fusion', the major search engines agree that a general encyclopaedia article on fusion power comes first, followed by other similar articles. A decade from now, the result will summarize the major approaches, contrast their differences, automatically translate any foreign documents into my language, and then rank the results by efficacy or place them in a table or chart as appropriate. If I then ask for 'background mathematics for fusion theory', I will get an outline for an impromptu course concentrating on the necessary complex analysis, customized to specific applications in fusion and to my level of mathematical understanding. If I stumble, the course will be readjusted to fit my needs, or perhaps the search engine will connect me to a tutor or another student in a similar plight. Interaction with search engines will be an ongoing conversation; one that is integrated with the other ongoing tasks of our lives.

"An experimental minority of search queries will be through the direct monitoring of brain signals."

One big challenge for search engines is to implement a measure of quality that is not based solely on popularity. Search engines must determine both relevance (is the item pertinent to the user's query?) and quality (is the item inherently accurate, useful and understandable, independent of the query?). Current relevance measures do reasonably well. Measures of quality require better models of the concepts and relations expressed in documents and how they relate to the reality of the world, as well as models of the trustworthiness of authors. Thus, a site that claims that the Moon landings were a hoax and seems to have a coherent argument structure will be judged to be lower quality than a legitimate astronomy site, because the premises of the hoax argument are at odds with reality. Understanding and improving these models is a key challenge for the coming decade.

Microbiome

David A. Relman
Chief of infectious diseases at Veterans Affairs Palo Alto Health Care System, Palo Alto, California

The human body is one of the most important ecological study sites of the coming decade. Humans depend on the microbial communities that colonize them for a surprising suite of benefits. These include: extracting energy from food, educating the immune system and protection from pathogens. Yet, despite the recent attention to this indigenous microbiota, we are relatively ignorant of what our 'extended self' comprises or how it works.

The human body consists of multiple microbial habitats, studied and defined so far on the basis of gross anatomical features, such as the skin, mouth, intestines and vagina. Only a subset of the relevant habitats and habitat boundaries across the human landscape have been identified — and important biology often takes place at such boundaries. Over the next ten years, molecular microbial surveys need to capture rare species and assess diversity at multiple spatial scales.

Although the organization of the human microbiota is not random, little is known about the rules that govern its assembly. What are the contributions of early exposures, dispersal, and species interactions? Is there selection at the community level, and if so, how? And, most importantly, how does the human body control community composition? With answers to these questions, the assembly of the microbial community, for example on the tooth surface or intestinal mucosa, could be guided towards states that confer health.

Equally pressing questions concern the stability and robustness of human microbial communities. How well do these communities resist perturbation by forces such as antibiotics, or return to their prior state after disturbance? How many healthy states are there? Does community resilience determine or predict human health? What mechanisms underlie resilience, and how can they be measured or reinforced?

Answering these questions requires understanding the functional properties of the microbiota. This means coupling DNA sequencing with direct measurements of RNA, protein products, functional assays and associated environmental variables. The existing national and international projects to map the human microbiome are a good start. In addition this field needs well-controlled, longitudinal clinical studies; non-disruptive, minimally invasive sampling methods; management and analysis strategies for complex, multi-dimensional data; and new partnerships between microbiologists, ecologists, clinicians, physiologists and technologists.

Personalized medicine

David B. Goldstein*
Duke University

Over the past decade, powerful genotyping tools have allowed geneticists to look at common variation across the entire human genome to identify the risk factors behind many diseases. Two striking findings will define the study of disease for the decade to come. First, common genetic variation seems to have only a limited role in determining people's predisposition to many common diseases. Second, gene variants that are very rare in the general population can have outsized effects on predisposition.

For example, rare mutations that cause the elimination of chunks of the genome can raise the risk of diseases such as schizophrenia, epilepsy or autism by up to twentyfold. Some researchers view these major risk factors as aberrations. My guess is that as more genomes are sequenced, many other high-impact risk factors will be identified.

If so, here's one confident but uncomfortable prediction of what personalized genomics could look like in 2020. The identification of major risk factors for disease is bound to substantially increase interest in embryonic and other screening programmes. Society has largely already accepted this principle for mutations that lead inevitably to serious health conditions. Will it be so accommodating of those who want to screen out embryos that carry, say, a twentyfold increased risk of a serious but unspecified neuropsychiatric disease?

"We are relatively ignorant of what our 'extended self' comprises or how it works."

Some advances will be relatively uncontroversial, such as the development of tailored therapeutic drugs based on genetic differences that are otherwise innocuous. Others will be transformational, such as the identification of definitive genetic risk factors that provide new drug targets for conditions that are often poorly treated such as schizophrenia, epilepsy and cancers. Over the next decade millions of people could have their genomes sequenced. Many will be given an indication of the risks they face. Serious consideration about how to handle the practical and ethical implications of such predictive power should begin now.

Energy

Daniel M. Kammen
Director of the Renewable and Appropriate Energy Laboratory, University of California, Berkeley

By 2020, humankind needs to be solidly on to the path of a low-carbon society — one dominated by efficient and clean energy technologies. It is essential to put a price on carbon emissions, through either well-managed cap-and-trade schemes or carbon taxes. Creative financing will also be needed so that homes and businesses can buy into energy efficiency and renewable energy services without having to pay up front. An example is the Property-Assessed Clean Energy financing mechanism, which my lab is helping to design and promote (http://rael.berkeley.edu/financing).

Government funding of research is crucial. Several renewable technologies are ready for explosive growth. Energy-efficiency targets could help to reduce demand by encouraging innovations such as net-positive-energy buildings and electric vehicles. Research into solar energy — in particular how to store and distribute it efficiently — can address needs in rich and poor communities alike. Deployed widely, these kinds of solutions and the development of a smart grid would mean that by 2020 the world would be on the way to an energy system in which solar, wind, nuclear, geothermal and hydroelectric power will supply more than 80% of electricity.

Mental health

Daniel R. Weinberger
Senior investigator, US National Institute of Mental Health

The search over the past decade for genes behind mental illness has led to the realization that mental disorders are not discrete conditions with specific causes. Rather, they are the result of interactions between risk factors that affect development; psychiatric symptoms can arise from many causes and are more interrelated than current disease models allow. By 2020, this insight, which has been slow to take hold, will have transformed how doctors understand and treat psychiatric conditions.

Finding specific genes for mental illness now seems a pipe dream. A more realistic endeavour for the next ten years is to look for genes that code for basic cellular and brain functions that modulate our responses to the environment and that come together in particular ways in individuals at increased risk. Many hundreds of genes may contribute to raised vulnerability, and such defects may affect brain development and function independently of any specific psychiatric diagnosis. There is no straight road to psychiatric illness, but a highly diverse network of developmental pathways.

This approach will lead to diagnosis and treatment based on a proper grasp of the underlying biology, rather than on an interpretation of symptoms. Psychiatric research is poised to realize Sigmund Freud's dream of a biological psychology, but it will require new applications of old thinking (see also page 9).

Hominin palaeontology

Leslie C. Aiello
President, Wenner-Gren Foundation for Anthropological Research

Most of the recent effort in hominin palaeontology has been focused on Africa and Europe. But the announcement in 2004 of the small hominin *Homo floresiensis* in Indonesia was a warning that we are naive to assume we know more than the basic outline of human evolutionary history. If *H. floresiensis* is indeed a surviving remnant of early *Homo* that left Africa around 2 million years ago, we have to reject the long-standing idea that *Homo erectus* was the first African emigrant. We also must reject many hypotheses concerning the prerequisites for this emigration, such as a relatively large brain size, large body size and human-like limb proportions. Importantly, we must confront our relative ignorance about human evolution outside Europe and Africa.

One of the big challenges for the next decade is to begin to fill in the large gaps in our knowledge about human evolution in Asia. We need strong and creative international collaborations that have the financial, institutional and governmental support to carry out the necessary research and interpretation. The field needs large, focused support in Asia similar to that given to research in eastern Africa by the Turkana Basin Institute in Stony Brook, New York, the Max Planck Institute for Evolutionary Anthropology in Leipzig and the Heidelberg Academy of Sciences and Humanities programme on 'The Role of Culture in Early Expansions of Humans'. Fossil hunting is a high-risk venture and expeditions may not always produce the desired results. However, the number of new hominin species discovered in Africa and Europe in the recent past suggests that a similar effort in Asia would not go unrewarded.

"Corporate lobbying must be restrained, it is one of the greatest dangers to sustainable development."

Hopefully, by 2020, we will have many more pieces of the big puzzle of human evolution — how and why did hominins evolve and disperse worldwide over a period of around 6 million years? Advances can be expected from areas such as genetics, isotope analyses and palaeoclimate research, as well as from fossil discoveries. But we cannot answer the key questions about human evolution without working towards a more geographically complete fossil record.

Synthetic biology

George Church
Professor of genetics, Harvard Medical School

In the past decade, the cost of reading and writing DNA has dropped a million-fold, outstripping even Moore's law for exponentially increasing computer power. The challenge for the next decade will be to integrate molecular engineering and computing to make complex systems. The development of engineering standards for biological parts, such as how pieces of DNA snap together, will permit computer-aided design (CAD) at levels of abstraction from atomic to population scales. Biologists will have access to tools that will allow them to arrange atoms to optimize catalysis, for example, or arrange populations of organisms to cooperate in making a chemical.

The obvious application will be in manufacturing and delivering drugs more efficiently. However, these treatments might be superseded by smarter ones, such as oral vaccines and 'programmable' personal stem cells or bacteria (which exploit sensors, logic and actuators harvested from natural and lab evolution) that could, for example, sense a nearby tumour, coordinate an attack and drill into the cancer cells to release toxins. Another application is in the production of chemicals, biofuels and foods — for example, the development of parasite-resistant crops or photosynthetic organisms that can double their biomass in just three hours. As costs drop, such technology will allow developing nations to leapfrog fertilizer-wasting, fossil-fuel-intensive and disease-rife farming for cleaner, more efficient systems, just as they are leapfrogging costly landlines in favour of mobile-phone networks.

Synthetic biology is already having an impact beyond its field, and by 2020 this will have increased significantly. Myriad technologies will be possible, such as nano-memory devices that harness the ability of certain bacteria to navigate Earth's weak magnetic field using magnetite nanoparticles. As electronic chips hit conventional manufacturing limits, they will be replaced by atomically precise and fault-tolerant biological circuits. Three-dimensional 'bio-printers' could make nearly all manufactured goods much less expensive. The grand challenge will be to anticipate the many unintended consequences of the synthetic biology revolution — ecological, economic and social — and to safeguard against them.

Universities

John L. Hennessy
President, Stanford University

The world faces increasingly complex challenges, such as maintaining our ecosystem while supporting 9 billion to 10 billion people, reducing poverty, increasing peace and security, and improving human health in both the developed and developing world. Universities must have a role in seeking solutions for these problems and in educating the next generation of leaders to tackle them.

Perhaps the largest threat to our research universities over the next decade is the financial challenge facing governments. In the United States, for example, budget deficits have caused many states to reduce their funding for public universities, and at the federal level, there is likely to be no growth or a cut in funding for research programmes.

To address these financial and intellectual challenges, universities need to be willing to change how they see their research and teaching mission. The scale and complexity of today's global problems demand a more collaborative, multidisciplinary approach.

Traditionally, universities have been structured around disciplines and departments. The agencies that fund research often reflect that structure in their financial support of projects. That rigidity can be a barrier to innovation, and to the need to educate students for a more collaborative working environment.

Therefore, universities and funding agencies need to encourage working across disciplines — for example, through academic centres based around broad themes rather than narrow fields. The challenge will be to do this without abandoning the traditional disciplines and the role they have in ensuring excellence.

As financial pressures increase, institutions may be forced to make difficult decisions — prioritizing areas in which they have sufficient existing strength or student interest and collaborating with peer institutions that have greater capability in other fields. Continuing support for fledgling cross-disciplinary efforts in difficult financial circumstances will require vigilance.

Universities have a dual charge: to advance the boundaries of knowledge and to educate students. Through this dual role we have the potential to make contributions that can shape the future. The challenge of the next decade is to live up to that potential.

Global governance

Jeffrey Sachs

Director, the Earth Institute

By 2020, the world needs an effective system of global governance for managing sustainable development. It will require systematic improvements in four areas.

First, politics must take account of technical expertise. In international negotiations such as the Copenhagen climate process, negotiators spend a lot of time arguing over the legalities of agreements but little time discussing technological options. There is a tendency to announce targets without technical strategies, and then to miss the targets. The United Nations should follow up the Copenhagen meeting by setting up expert groups to support the practical tasks of climate-change mitigation and adaptation. Within a few years, a new world environment organization should be established to oversee and provide technical support for the major treaties.

Second, public and private investments in new technologies should be managed as part of an integrated system. Almost all environmental challenges, from greenhouse-gas emissions to the depletion of groundwater resources, demand technological transformation. Achieving this will need a mix of public and private enterprise. The public sector will be responsible for issues such as monitoring, regulation and public safety and awareness; the private sector will take the lead in profit-oriented investments, in particular in research and development. Both sides will need to harmonize their actions and seek effective partnerships.

Third, corporate lobbying must be restrained: it is one of the greatest dangers to sustainable development. In the United States, corporate influence through lobbying, campaign funding and misleading advocacy campaigns has been an enormous obstacle to effective regulation of the economy and environment. For example, heavy lobbying by Wall Street contributed to the financial deregulation that helped cause the recent crisis, and pressure from the energy industry has delayed action on climate change. Some countries have successfully constrained such influence through public financing of elections and other means. The United States should follow suit.

Finally, global financing for poorer countries must improve if international agreements on climate, land use and biodiversity are to succeed. The record of aid delivery to poor countries is dismal. Rich countries regularly promise support that never arrives. Two proposals have been made that could improve things: a small tax on cross-border financial transactions, and a global levy on carbon emissions. Both should be implemented alongside more traditional forms of aid to secure a more reliable source of development finance.

Astronomy

Adam Burrows

Vice-chair of the Board on Physics and Astronomy of the US National Research Council

Key questions for the coming decade include determining the nature of the dark matter that permeates the Universe — it would be a major embarrassment if the dark matter paradigm was not verified within 40 years of its inception by the direct detection of the associated weakly interacting particles. Some people single out the nature of dark energy as the most fundamental puzzle confronting astronomy. Others want to know how tenuous gas and dust is converted into dense stars and planets and how many Earthlike — and habitable — planets populate the Galaxy. Answers to all these questions could be found by 2020, but the astronomy community must decide which to prioritize.

Prioritization will not be easy. Future technologies will inevitably be more complicated and expensive. Ground-based astronomy has become big science and space-based astronomy struggles to balance creativity and affordability. As a consequence, the operating budgets for current telescopes are constraining future telescope construction projects. Moreover, the life cycle costs of the James Webb Space Telescope, due to be launched in 2014, have and will continue to cut into budgets for smaller, cheaper and more nimble astrophysics missions.

To craft an exciting and integrated strategy for achieving in the next decade the promise of the previous one, the United States has embarked on its decadal survey of astronomy, due to be completed before the end of 2010. Astronomers have submitted an avalanche of public white papers articulating the scientific and engineering cases for missions and facilities that collectively could cost more than US$70 billion. This exceeds the likely funding for new initiatives in the next decade by at least four to five times. Therefore, the survey committee is charged with finding the right balance between large and small projects and the proper mix of ground and space observatories. It must also address the coordination between public and private telescopes, a unique feature of American astronomy, as well as determine the optimal suite of instruments for those telescopes.

This is astronomy's golden age. The potential for startling breakthroughs remains great, but considerable money and skill will be necessary to realize even a fraction of it. Will humankind baulk on the threshold of a comprehensive understanding of the Universe? Policy decisions in the next year or two may well decide the issue.

Drug discovery

Gary P. Pisano
Professor of Business Administration, Harvard Business School

The next ten years will witness an acceleration of the upheaval in the pharmaceutical industry. Profound changes in drug research and development, competition, government policies and markets will continue to challenge existing business models and strategies. Many established players will not make the transition. Some venerable companies have already disappeared through acquisitions.

The industry will bifurcate into firms that pursue a long-term commitment to creating novel drugs and those that focus on marketing. The latter may do better in the short term but are doomed to failure eventually. The development of effective treatments is the only sustainable source of value for the pharmaceutical industry. Given the paucity of therapies for many serious diseases and the mediocre efficacy of many existing drugs, the opportunities are huge. There are risks in trying to discover new drugs, but the risks of backing away from that commitment are higher.

I do not envisage one dominant model in terms of size or organizational structure. We will see an ecosystem of few large global players with deep scientific resources and many more specialized companies. Globalization of drug innovation will continue. No one should be surprised to see the emergence of a major Chinese multinational drug company with strong innovation capabilities.

"No one should be surprised to see the emergence of a major Chinese multinational drug company."

Demographics

Joshua R. Goldstein
Executive director, Max Planck Institute for Demographic Research

As population growth marked the twentieth century, population ageing will mark the twenty-first. By 2020, the average European will have fewer years of life expectancy remaining than years he or she has already lived. East Asians will soon follow. Humankind will spend much of the coming decade grappling with questions about how to organize and pay for the care of an increasing elderly population and about who will produce what the elderly consume.

In the longer term, a return to moderate fertility rates in those countries with very low fertility, and increases in immigration can do much to moderate population ageing. Sweden and Japan face quite different demographic futures, because fertility in Sweden is closer to replacement and a small but steady stream of immigrants will make up the difference. In Japan — the world's leader in longevity — fertility remains low, and immigration a major social challenge.

We need demographic research on four fronts addressing population ageing. Low birth rates can perhaps be increased by measures that reconcile work and family, enabling people to have the children they say they want. Fostering the social and economic integration of immigrants is another priority. Health research, helping people to stay younger longer, is already a priority of ageing societies; indeed, so far, the healthy period of life has been lengthening as fast or faster than life expectancy itself. But now — as the first 65-year-old baby-boomers prepare to blow out their birthday candles — we must address the larger question of rescheduling life's turning points, so that people can remain active and productive. The societies that respond to ageing successfully will be those that take advantage of longer life.

Chemistry

Paul Anastas
Center for Green Chemistry and Green Engineering, Yale University

The future of chemistry should look very different from the past. Traditional, reductionist, highly specialized academic chemistry has transformed food, energy, health, transportation, communications and the quality of modern life. It has also — accidentally — depleted finite and rare resources, endangered workers and contaminated ecosystems. Green chemistry is the way forwards: it combines expertise from synthetic, physical and biological chemists, together with that of toxicologists, environmental health and life scientists, to deliver sustainable chemical design.

Making chemical products and processes that reduce or eliminate the use and generation of hazardous substances is an inherently systems approach. The 'twelve principles of green chemistry' unite all aspects of the molecular life cycle, from obtaining the feedstock and starting materials, through the synthetic and manufacturing process, to the end of commercial life and ultimate disposal of products. These principles are based on the latest fundamental discoveries on the interaction between anthropogenic substances and the natural world.

Scant research funding, and hence insufficient effort, is devoted to sustainable innovation in chemistry. As a first step, chemistry needs to adopt a clearly stated research imperative that researchers in molecular science must maintain their creativity while not doing harm to people and the planet. We need to turn all of chemistry green.

NIH

Richard Klausner
Column Group
David Baltimore
California Institute of Technology

The National Institutes of Health (NIH) serves the US biomedical community by providing resources for experimentation, but it does so in ways that bias the enterprise towards short-range and unimaginative thinking. Our recommendations for the NIH of 2020 call for a profound change in its culture and in its decision-making processes.

First, funding criteria will put more weight on judgements about the individual who is applying, not the details of the proposed project. It is creative minds that we want to foster, and when the NIH identifies someone who has been innovative and productive, that person should be adequately supported so they can express their creativity in their own way.

Notably, the current system of hyper-specialized study sections for reviewing research proposals discourages risk-taking. They should be replaced or augmented by broad, institute-based interdisciplinary review teams, which assess greatly simplified applications that focus on the goals of the research, the importance of the problem and the quality of the investigators. The technical part of the review will shift from assessing the feasibility of the plan to the capabilities of the investigators.

At the same time, we should be encouraging new generations of independent scientists to begin productive careers by aiding their development outside the usual academic routes. So, instead of all trainees being graduate students and postdoctoral fellows under supervision by elders, there should be alternative pathways for independent or self-guided study.

By 2020, the NIH should see some of the fruits of its programme to revitalize clinical research. The clinical trials it supports (some 15% of the agency's overall budget) should be asking questions that enhance the scientific basis of medicine. For instance, NIH-sponsored trials should focus on streamlining trial execution and should pioneer new technologies for patient subtyping, testing biomarkers and determining biologically meaningful surrogates for clinical responses. That might mean fewer trials than now, but each should be designed to extend clinical science as a whole.

The intramural programme of the NIH represents some 10% of its funding and should remain strong. However, it lacks a defined mission and has deteriorated in quality. It does have a powerful and unique instrument in its new but underutilized clinical centre, which needs to move to the forefront of the NIH's translation efforts.

Individuals are also key to progress in clinical research. In the extramural community, we need an expanded cadre of clinical research scholars to pursue cross-disciplinary studies of human disease physiology, and to challenge the current one-way route from bench to bedside.

If the NIH carries out these reforms by 2020 (even better, by 2015), the United States' preeminence in biomedical research will be ensured.

Soil

David R. Montgomery
Author of *Dirt: The Erosion of Civilizations*, University of Washington

To avoid the mistakes of past societies, as 2020 approaches, the world must address global soil degradation, one of this century's most insidious and under-acknowledged challenges. Humanity has already degraded or eroded the topsoil off more than a third of all arable land. We continue to lose farmland at about 0.5% a year — yet expect to feed more than 9 billion people later this century.

> "Business as usual is not an option when it comes to soil. It's time for a greener revolution."

During the twentieth century, the Haber–Bosch process (allowing the mass production of nitrogen-based fertilizers) and the Green Revolution effectively divorced agriculture from soil stewardship. Increased yields were supported by intensive fertilizer inputs and mechanization that simplified and devastated soil life, reducing native soil fertility. For example, research in some conventional agricultural settings shows that other species such as bacteria have virtually replaced mycorrhizal fungi, which deliver soil nutrients to most plants. In a post-petroleum world, as the era of cheap fossil-fuel-produced fertilizers comes to an end, conventional, high-input agriculture is neither sustainable nor resilient. Ensuring future food security and environmental protection will require thoughtfully tailoring farming practices to the soils of individual landscapes and farms, rather than continuing to rely on erosive practices and fertilizer from a bag.

Towards these ends, governments should aggressively fund research on and promote the adoption of agricultural practices and technologies that cultivate beneficial soil life and sustain soil ecosystems. Over the next few decades, approaches such as low-till and organic methods could restore native soil fertility and store enough soil organic matter to offset global fossil-fuel emissions by 5–15%. Offsets, and soil fertility, could be further increased through adding biochar — charcoal made by heating organic wastes.

The thin layer of minerals, living microorganisms, dead plants and animals blanketing

the planet is the mother of all terrestrial life and every nation's most strategic resource. Yet we treat it like dirt. Business as usual is not an option when it comes to soil, food and people. It's time for a greener revolution.

Lasers

Thomas M. Baer
Stanford Photonics Research Center
Nicholas P. Bigelow
Department of Physics and Astronomy, University of Rochester

Those who conceived and invented the laser 50 years ago this year could not have predicted the roles that it has had over the past half-century: from communications to environmental monitoring, from manufacturing to medicine, from entertainment to scientific research.

By 2020, lasers will probably emit beams with spot sizes of the order of 1 nanometre — the size of a small molecule. Objects with dimensions less than a wavelength cannot usually be resolved using lasers or microscopy unless the photons are emitted from an aperture smaller than the object. Microscopes that incorporate laser sources with apertures the size of a single molecule will be useful in fast, direct sequencing of biomolecules such as DNA and RNA. These miniature beams will also provide hard-disk storage at densities 100 times greater than those available today — petabytes of storage in a personal computer.

Ultraprecise, laser-based clocks will measure the drift in fundamental constants as the Universe expands, challenging our theories describing the origin and evolution of the cosmos.

Next-generation lasers will allow the creation of new states of matter, compressing and heating materials to temperatures found only in the centres of massive stars, and at pressures that can squeeze hydrogen atoms together to a density 50 times greater than that of lead. The resulting fusion reactions may one day be harnessed to provide almost limitless carbon-free energy. Enough fusion fuel is present in the oceans to supply the current energy needs of the entire world for longer than the age of the Universe.

By 2020, lasers will generate ultrashort bursts of photons — with pulse widths shorter than the time it takes for light to traverse an atom. These attosecond pulses will allow strobe pictures to be taken of chemical reactions — stop-action pictures of electrons in motion. When amplified to ultrahigh intensities, these lasers will be used as engines to accelerate electrons and protons to velocities close to the speed of light. This will mean that table-top accelerators can be created to generate particles with kinetic energies that rival those in today's biggest particle accelerators at a fraction of the size and cost.

What are the challenges to achieving these remarkable goals? Developing new laser materials, sources, optics that can survive enormously high intensities and new nanofabrication technologies. Will all of this happen in the next decade? We believe so. Like the inventors of 1960, we are probably still underestimating the full potential and impact of lasers.

Ecology

Robert D. Holt
Department of Biology, University of Florida

The greatest practical challenge facing ecologists over the next decade is that much of what we wish to study may vanish before we can really fathom it. The planet is increasingly dominated by ersatz ecosystems — human-sculpted landscapes occupied by haphazard assemblies of introduced species and tolerant natives. These are legitimate objects of study, but there are considerable practical, aesthetic and moral costs of losing natural ecosystems before we can even fully document and understand them.

"Ecology will be viewed increasingly as an essential dimension of the Earth sciences."

A key task will be to predict and mitigate this loss of biodiversity and the degradation of ecosystem function. One step is to gauge the resilience of ecological networks such as food webs — in particular, their capacity to withstand disturbance and species loss. This will require insights from many disciplines. Stable isotope analysis and genetic bar-coding should provide a clearer picture of who eats whom in a community.

Change takes place at multiple levels, from individuals to populations, to spatially linked ecosystems. I predict that by 2020, ecological theory will be increasingly concerned with the often subtle biological details of organisms, as well as the implications of evolutionary dynamics. Microbial ecology will become mainstream. At the same time, it will be essential to look at how species and communities fit into Earth's history. In a decade's time, ecology will be viewed both as a core part of biology, and increasingly as an essential dimension of the Earth sciences.

Metabolomics

Jeremy K. Nicholson
Head, Department of Surgery and Cancer, Imperial College London

The analysis of the chemical fingerprints left by metabolic processes has already started to play a crucial part in personalized medicine, particularly cancer therapy. This stems from the understanding that humans are metabolic superorganisms carrying the genomes of many symbiotic organisms, all of which can affect an individual's physiology. Human metabolism is heavily influenced by interactions between our own genes and the activities of gut microbes, as well as by diet and environmental stressors. The products of this metabolic interplay have a direct influence on susceptibility to disease.

Determining how the body's metabolic processes interact with those of gut microbes is a priority in the coming years, because many conditions, including ulcerative colitis, Crohn's disease, obesity, diabetes and autoimmune disorders, are linked to poor gut health and microbial imbalances. By 2020, personalized health care could involve doctors monitoring the metabolic activities of a patient's gut microbes and, possibly, modulating them therapeutically. The use of mathematical models to interpret metabolic data obtained using nuclear magnetic resonance spectroscopy and mass spectrometry will help us to understand the changing patterns of human disease on a global scale, and generate new targets for drug or nutritional interventions. ■

*Authors declare competing financial interests: details accompany the article online at go.nature.com/rSQXyc.

TOMORROW'S GIANTS

On 1 July 2010, *Nature* and the UK Royal Society are organizing a meeting called Tomorrow's Giants (see go.nature.com/PWNbfI). It will ask what is required to enable academic achievement of the highest quality in the coming decades. The conference will look across themes such as measuring and assessment, science organization, data and interdisciplinary work. In the second half of last year, the Royal Society hosted a series of regional workshop meetings for researchers to exchange views on these topics, in particular identifying the impact of web 2.0 on how scientists share data; communication between industry/services and academia; and issues affecting careers and research. We invite Nature readers to participate, in the first instance by joining the Nature Network forum (at go.nature.com/b1tvCA).

The conference takes place at the Queen Elizabeth Hall, the Southbank Centre, Belvedere Road, London SE1 8XX, UK. It forms part of a week of celebrations for the Royal Society's 350th anniversary (see go.nature.com/VLSTMT).

Physics mystery peppered with profanity

The latest thesis on the disappearance of physicist Ettore Majorana adds little, but reminds us of the Nobel-prizewinning quality of the discoveries he made during his brief career, explains **Frank Close**.

A Brilliant Darkness: The Extraordinary Life and Mysterious Disappearance of Ettore Majorana, the Troubled Genius of the Nuclear Age
by João Magueijo
Basic Books: 2009 (US); 2010 (UK).
304 pp. $27.50, £15.99

A Brilliant Darkness tells the tale of Italian theoretical physicist Ettore Majorana, who worked with Enrico Fermi in Rome before mysteriously disappearing in 1938. In his all-too-brief career, Majorana came up with the concept of the neutron but did not publish it, regarding it as trivia. He also discovered the Majorana neutrino — a fermion that can be its own antiparticle, a concept that is currently centre-stage in particle physics owing to the recent discovery that neutrinos can have mass. It is Majorana's disappearance, however, that has excited speculation throughout the past 70 years.

Theoretical physicist João Magueijo offers a coarse version of the story. But *caveat lector*: reader beware. It is at least the eighth book on the subject — and a contender for being the worst. Magueijo revels in profanities and distasteful similes that are far too offensive to relate here. And the way in which the physics is rendered is at times shaky too. The book does mention Wolfgang Pauli's famous critique of an idea being "not even wrong"; if only that were the case here.

In popular writing, there is always tension between trying to simplify while remaining faithful to the facts. Magueijo crosses this line on more than one occasion, as with his primer on the absorption of α-, β- and γ-radiation. "Alpha radiation was the least energetic and could easily be stopped with a sheet of paper," he writes. This contrasts, he continues, with the "very energetic [β-] radiation that could be stopped only by sheets of aluminium", with γ-radiation being "by far the most powerful", stopped only by layers of lead.

This explanation is bizarre: α-particles tend to have much higher energy than β-rays, and the relative ease with which α-particles are absorbed is due to their large mass, lower speed and greater electric charge — all of which enable them to ionize atoms in material easily. Energy loss by ionization is proportional to mass and to the square of the electric charge, and inversely proportional to the energy. The 'power' of γ-rays, whatever that means, is irrelevant.

Theories about the disappearance of Ettore Majorana (far left) overshadow his contributions to physics.

E. RECAMI & F. MAJORANA, COURTESY AIP EMILIO SEGRE VISUAL ARCHIVES, E. RECAMI & E. MAJORANA JR COLLECTION

The book also bandies around assertions that have to be taken on trust. Fermi, whom Magueijo describes as "intellectually *a bit limited*" (the italics are the author's), is on two occasions accused of plagiarism, about which I would have liked to have known more. First, he mentions a "clear example of scientific robbery", and then later condemns both Fermi and fellow physicist Edoardo Amaldi together of "happily publishing Ettore's 1928 work [in 1933] without even acknowledging him".

Majorana was certainly a great mind, but his oeuvre was limited: first by a refusal to publish ideas that he felt were footling, and then most dramatically by his sudden disappearance. This is usually attributed to suicide, en route by boat from Palermo to Naples. No body, passport or money have been found — although he had withdrawn half a year's salary from the bank immediately before his disappearance — leading to many conspiracy theories. Perhaps the most plausible is that he went to Buenos Aires, Argentina. A few months before his disappearance, he met Italian physicist Giuseppe Occhialini, who had just arrived from South America, and there is some circumstantial evidence that this influenced Majorana's thinking. Majorana had a postcard of the ship *Oceania*, and this very ship sailed to Argentina from Naples on the night of Majorana's disappearance. There were inconclusive sightings of Majorana in Buenos Aires into the 1950s, but frustratingly, none was followed up.

The story of Majorana's life and disappearance is told in a recent article by Barry Holstein entitled 'The Mysterious Disappearance of Ettore Majorana' (B. R. Holstein *J. Phys. Conf. Ser.* **173,** 012019; 2009), an excellent example that deals with both the disappearance and the science. Italian physicist Salvatore Esposito, whose work Holstein cites, has also investigated the mystery. His analytical examination of Majorana's disappearance, due to be published in the journal *Contemporary Physics*, shows that there are several flaws in some earlier superficial studies, and argues for the Argentinian thesis. There have also been films and documentaries, some serious, others full of conspiracy. *A Brilliant Darkness* dips into these, although often in a confusing manner — it is not always clear whether one is reading a quote or something that the author has culled from the literature or his own enquiries.

For anyone interested in Majorana's work, life and possible death, I recommend starting with Holstein, Esposito and the classic books listed within those works. *A Brilliant Darkness* adds little, and removes a lot, from the memory of Majorana and his contemporaries.

That said, there is one thing on which I agree with Magueijo: that Majorana's work has Nobel-prize quality. Nobel prizes cannot be given posthumously, but they can be given in absentia, and in his prologue the author poses the critical question: "But is Ettore dead? We simply don't know." After all, if Majorana is alive, he would only be 103. ■

Frank Close is professor of physics and a fellow at Exeter College, University of Oxford, Oxford, UK.
e-mail: f.close1@physics.ox.ac.uk

HIGH-ENERGY READS

From mini black holes to the origin of mass, the science behind the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, is explored in two recent books. In *Collider* (Wiley, 2009), physicist Paul Halpern describes how the extreme energies of the LHC will unveil the basic building blocks of the Universe. He explains how the particle accelerator works, how and why it was built and what exciting results it may generate. And he reassures readers that media reports of the machine's ability to create dangerous miniature black holes are unfounded.

In *A Zeptospace Odyssey* (Oxford University Press, 2009), theoretical physicist Gian Francesco Giudice also relates the history and physics of the LHC, placing it into a broad context. As well as highlighting the theories that may be challenged by the groundbreaking experiment, including supersymmetry and string theory, he anticipates the intellectual revolution that it may trigger.

A more widely used type of accelerator is the subject of *Velvet Revolution at the Synchrotron* (MIT Press, 2009). Sociologist Park Doing examines how biological experiments, such as protein crystallography, have gradually taken over from physical-science studies at synchrotron facilities. Although the machine is a product of the particle-physics work of the Second World War, as an intense X-ray source it is frequently used to analyse the structures of proteins and complex molecules. Doing argues that this research switch came about through a series of alternating periods of assertion and resistance, rather than being driven from the top or bottom.

Body image in Ancient Egypt

Body Parts: Ancient Egyptian Fragments and Amulets
Brooklyn Museum, New York
Until 2 October 2011

Queen Nefertiti had flat feet. The chief wife of King Akhenaten, who ruled Egypt in the mid-fourteenth century BC, she is known as a long-necked beauty from her bust in Berlin's Neues Museum. But it is her life-sized, long-toed feet that are exhibited at New York's Brooklyn Museum, depicted on a chunk of pink limestone found in the ancient Egyptian city of Hermopolis. The relief, dated to about 1352 BC, is one of the first examples of accurate foot anatomy in Egyptian art.

Other representations of body fragments displayed in this small yet delightful exhibition, *Body Parts*, include arms, hands, eyes, an ear, a beard, a backbone and a heart. The pieces have been assembled by curator Yekaterina Barbash from artefacts stashed in the museum's store rooms, never previously put on show because many are broken.

"When I looked at these fragments, I became aware of details I don't usually notice when viewing a complete sculpture," Barbash says. "I decided that it would be interesting to focus the attention of viewers on just one body part, so that they notice details of modelling, symbolism and workmanship."

For example, a slick of embalming resin is evident between the two fingers of an amulet fashioned from obsidian, a black volcanic glass. Dating from 332 BC, it would have been placed on a mummy as a protective charm and was used to close the incision through which internal organs had been removed before mummification.

A tiny wooden arm, the size of a Barbie doll's, ends in a clenched fist that once held a sceptre, suggesting that it belonged to a statuette of a man, most probably from the third millennium BC. A braided and curved bronze beard from the mid-seventh century BC is associated with Osiris, the god of the underworld, and is inlaid with dark-blue glass to represent lapis lazuli, a semi-precious stone from which divine hair was said to have been made. A small obsidian amulet from the same period, shaped like a set of lungs attached to a windpipe, is also a hieroglyph used in the words 'to unite' — a testament to the importance of a unified corpse on its journey in the afterlife.

Egyptian art and anatomical accuracy went hand in hand.

CHARLES EDWIN WILBOUR FUND/BROOKLYN MUSEUM

The Egyptians "really strove to keep the body intact in its most ideal form", says Barbash. However, Egyptian mortuary texts associate different body parts with a specific deity: for example, the lips are associated with Anubis, the jackal-headed god of mummification, and the neck with Isis, patron of women.

Although Egyptian artists emphasized perfection — Nefertiti's flat feet are an ideal, rather than an accurate, model — they also expressed respect for atypical forms such as dwarfism, a syndrome that they linked with gods who protected women during pregnancy and childbirth. Dwarfism is common in Egyptian art, and on display is a granite statue of a short-legged scribe dated to between the first century BC and the first century AD. Nearby is a small wooden figure (1539–1075 BC) with a bowed head and severe curvature of the spine. The fact that sages taught people to be accepting of those with disabilities is shown by a wisdom text from the Ramesside period (1295–1069 BC, named after the 11 kings with the name Ramesses) placed alongside the wooden figure: "Do not laugh at a blind man, ridicule a dwarf, or impede the disabled." ■

Josie Glausiusz is a journalist based in New York City. She also has flat feet.
e-mail: jg@planetjosie.net

NEWS & VIEWS

SUPERNOVAE

A smashing success

D. Andrew Howell

The progenitors of type Ia supernovae, the standard candles that lit the way to dark energy, have been elusive. A largely dismissed scenario has now produced one, but the results aren't what anyone expected.

For half a century, the origin of type Ia supernovae (SNe Ia) has been mired in a frustrating state of half-explanation. Part of the story is clear: they are the thermonuclear explosions of carbon–oxygen white dwarfs[1], the dense, spent cores of stars that were once like the Sun, but are now incapable of nuclear fusion. These balls of 'ash' can explode only if they gain mass, which compresses and heats them until carbon fusion is ignited. But to grow they must gorge on an orbiting companion, either by accreting the outer hydrogen layer of a still-burning star[2] or by entirely devouring another white dwarf by merging with it[3] (Fig. 1).

Unfortunately, there are serious problems with both of these ideas. In the first, it seems difficult to drizzle on enough hydrogen at just the right rate to avoid causing a nova — a surface eruption that results in white dwarfs losing, rather than gaining, mass. And if there is so much hydrogen around, why is it never seen in the supernova[4]? The merger scenario circumvents these shortcomings, but has its own issues — simulations show that the smaller white dwarf is shredded into a disk that falls rapidly onto the survivor. This should cause carbon ignition near the surface, where densities are too low to trigger a supernova[5] — a cosmic dud instead of celestial fireworks. However, on page 61 of this issue, Pakmor *et al.*[6] demonstrate that the oft-maligned merger proposal is viable after all, at least under certain conditions.

Pakmor and colleagues show that the trick is to start with two white dwarfs of almost equal mass. When they violently merge, carbon ignition occurs deeper down, where densities are high enough to achieve a supernova. Most simulations of the merger would have stopped there, but here the authors use a series of different codes tailored to probe the physics inherent in each step of the process: the three-dimensional distortions in the merger, the expanding supernova ejecta and the photons that emerge out of the cataclysm.

The result is surprising. Even with the large amount of bomb fuel provided by two white dwarfs (in contrast to one in the alternative scenario), the explosion is fairly wimpy. In their simulations, Pakmor *et al.*[6] don't get a normal SN Ia but rather a sub-luminous one — a special class with distinct properties that has so far been even more resistant to explanation than normal SNe Ia (ref. 7). Their result may explain old mysteries, such as the inability of previous models to produce sub-luminous SNe Ia and the curiosity that the only sub-luminous SN Ia whose shape could be studied was found to be weirdly aspherical[8].

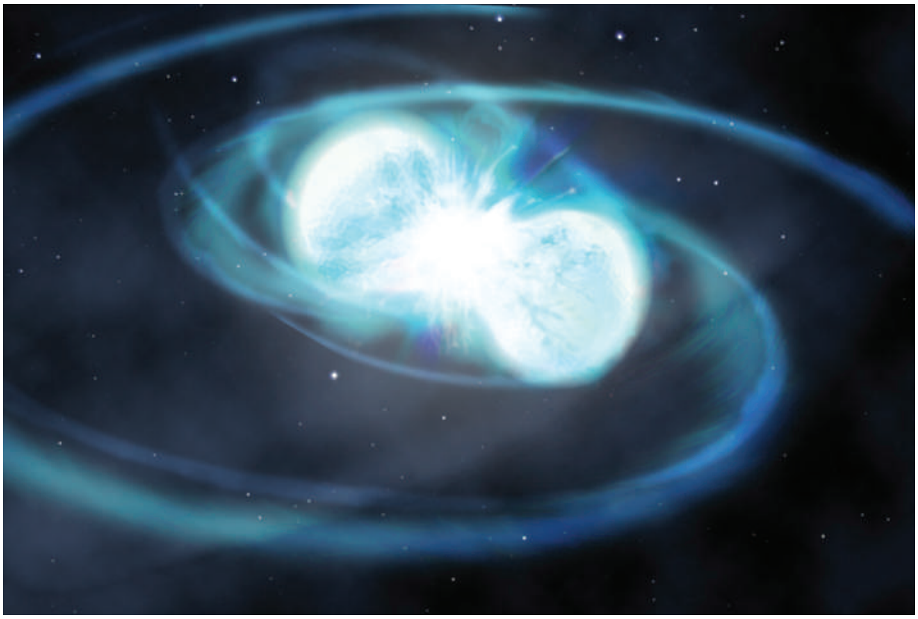


Figure 1 | Merging white dwarfs. **If two white dwarf stars are close enough together in a binary system, the emission of gravity waves will cause them to spiral together and merge. This idea was proposed as one explanation for the production of type Ia supernovae, but was largely dismissed because simulations showed that the merger would result in a neutron star rather than a supernova. Pakmor *et al.*[6] find that, if the white dwarfs are equal in mass, a type Ia supernova does result, but contrary to expectation it is a sub-luminous Ia — one that is dimmer than normal and that has so far defied explanation.**

Can these mergers explain more normal or even over-luminous SNe Ia? In particular, some explosions seem to be so bright that they require two massive white dwarfs' worth of material to explain their properties[9]. Pakmor *et al.* considered only white dwarfs near 0.9 solar masses, although their scenario ought to be viable for a more massive pair, which would almost certainly produce a brighter explosion. But the requirement of nearly equal-mass stars, surely a rare occurrence, may mean that this scenario is better for explaining uncommon oddities than the larger Ia class.

There are caveats to be sure. The authors' predicted supernova light curves and spectra are close — but not perfect — matches to sub-luminous SNe Ia. And, as with any novel code, the findings must be checked by other groups using different approximations. Furthermore, a merger may be one way to get a sub-luminous SN Ia, but it may not be a unique solution. The most serious concern is that Pakmor and colleagues' model[6] predicts a wide range of time delays, some shorter than a hundred million years, between the birth of the progenitor stars and their ultimate demise as supernovae. By contrast, sub-luminous SNe Ia generally don't show short time delays — they usually take a billion years or more to explode[10]. Reproducing this behaviour with the model is possible, but requires some fine-tuning.

Do these findings have implications for the use of SNe Ia as standard candles to measure the dark energy thought to be driving the accelerating expansion of the Universe? Yes and no. These sub-luminous SNe Ia are too dim to be seen at great distances, so are not useful in cosmological studies. However, one of the great worries about the use of SNe Ia, especially given their murky origins, is how their average properties may change with cosmic time[11]. Therefore, any understanding of their progenitors is progress. The dream is to one day understand what causes each subclass of SN Ia, so that we can model any change in supernova demographics as we look back in time through the Universe. Better yet, a separation of SNe Ia into different categories, arising from physically distinct processes, may make each subclass better standard candles.

Pakmor and colleagues' study is a big step forwards: after decades of modelling, it finally seems that white-dwarf mergers can make some supernovae. But it is an early step down a long path exploring where this scenario might take us. And if it can't explain all SNe Ia, what are the rest? ■

D. Andrew Howell is at the Las Cumbres Observatory Global Telescope Network and the Department of Physics, University of California, Santa Barbara, California 93117, USA.
e-mail: ahowell@lcogt.net

1. Hoyle, F. & Fowler, W. A. *Astrophys. J.* **132,** 565–590 (1960).
2. Whelan, J. & Iben, I. *Astrophys. J.* **186,** 1007–1014 (1973).
3. Iben, I. & Tutukov, A. V. *Astrophys. J. Suppl.* **54,** 335–372 (1984).
4. Leonard, D. C. *Astrophys. J.* **670,** 1275–1282 (2007).
5. Saio, H. & Nomoto, K. *Astron. Astrophys.* **150,** L21–L23 (1985).
6. Pakmor, R. *et al. Nature* **463,** 61–64 (2010).
7. Taubenberger, S. *et al. Mon. Not. R. Astron. Soc.* **385,** 75–96 (2008).
8. Howell, D. A. *et al. Astrophys. J.* **556,** 302–321 (2001).
9. Howell, D. A. *et al. Nature* **443,** 308–311 (2006).
10. Gallagher, J. S. *et al. Astrophys. J.* **685,** 752–766 (2008).
11. Sullivan, M. *et al. Astrophys. J.* **693,** L76–L80 (2009).

NEUROSCIENCE

Editing out fear

Gregory J. Quirk and Mohammed R. Milad

Retrieving a memory initiates a window of vulnerability for that memory. Simple behavioural methods can modify distressing memories during this window, eliminating fear reactions to traumatic reminders.

We all have memories that we would rather forget, some embarrassing and some painful. Most of us learn to cope with these memories, but this can be difficult for those suffering from post-traumatic stress disorder (PTSD), in which reminders of traumatic events can trigger dread and fear. One way to counteract such associations is to extinguish them with repeated exposures to traumatic reminders within a safe environment. However, because this extinction process does not eliminate the fear memory, fear reactions often return, especially during stress.

An alternative approach to treating PTSD was suggested by studies[1] in rodents. These studies showed that the mere retrieval of a memory triggers a reconsolidation process, during which the memory briefly becomes labile before being re-stored. Drugs that block reconsolidation can degrade the original fear memory in animals, but it has been difficult to apply such a strategy to humans because most of these drugs are toxic. On page 49 of this issue, Schiller *et al.*[2] report that giving extinction training to humans during the reconsolidation window effectively redefines fearful memories as safe*.

It has been known for a century that memories must undergo a consolidation process in order to be stored in the long term[3]. An exciting and more recent discovery was that retrieving a memory triggers a reconsolidation process that employs many of the molecular mechanisms used in the original consolidation. Reconsolidation has been observed for several types of memory across different species[4], but these studies beg the question "what is the advantage of remaking a memory multiple times?" One possibility is that reconsolidation allows memories to be updated in the light of events that have occurred since the last retrieval[5].

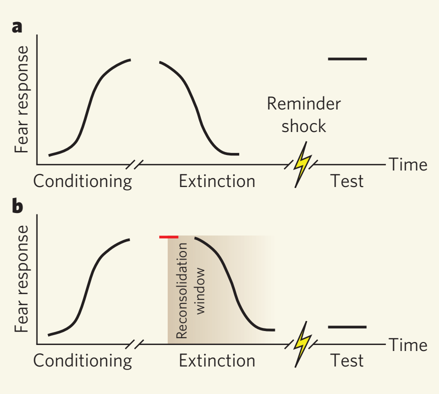


Figure 1 | Preventing fear's return. a, Schiller *et al.*[2] conditioned human volunteers to fear a visual cue paired with a mild electric shock. They then extinguished the fear using extinction training, in which the cue was repeatedly presented in the absence of the shock. Fear responses extinguished in this way returned the following day when the participants were given a reminder shock. The graph shows the magnitude of a participant's fear response during the three parts of the experiment. **b,** Volunteers who received extinction training shortly after a retrieval trial (red line), which causes the subject to recall the fear memory, exhibited no fear response after being given a reminder shock the next day. This is because the retrieval of a memory triggers a reconsolidation window during which the memory can be updated by extinction training.

This hypothesis was recently tested[6] by some of the authors of the present study. In rats conditioned to fear a tone paired with an electric shock, the researchers observed that extinction training conducted within 10 minutes of memory retrieval eliminated fear of the tone, and prevented the fear from returning under a variety of circumstances, even if a reminder shock was administered. Furthermore, rats in which fear memories had been extinguished within this critical window were resistant to re-learning the tone–shock association. So, instead of forming a new memory of safety, extinction training given during the updating window apparently converted the existing memory of fear into one of safety.

These findings[6] in rodents raised questions about human memory. Does the retrieval of fear memories in humans trigger similar updating windows in which the memories could be modified by extinction training? If so, then how specific and long lasting are the memory-editing effects of reconsolidation–extinction procedures? To find out, Schiller *et al.*[2] used a well established fear-conditioning protocol, in which human volunteers learned that the appearance of a visual cue (a blue square) predicts a shock to the wrist (Fig. 1). The authors used the participants' skin conductance as a measure of their fear — skin conductance increases with sweating, a phenomenon exploited by lie detectors.

As in rodents, humans who recalled their fear memory 10 minutes before extinction training showed no fear response when tested 24 hours later. Furthermore, their fear responses did not return even if they were given a reminder shock at the start of the test day — they reacted to the blue square as if it had never been associated with a shock. By contrast, those participants whose memories were extinguished outside the critical window exhibited high fear responses when tested 24 hours later.

So are the memory-editing effects of extinction specific to the reactivated memory, or do they generalize to other memories? To address this issue, Schiller *et al.* conditioned volunteers to fear two stimuli, a blue square and an orange square, but only reactivated the fear memory of the blue square in a reminder trial. The participants underwent extinction training for both stimuli and were then administered

*This article and the paper under discussion were published online on 9 December 2009.

a shock. This caused a return of their fear only in response to the orange square, confirming that the memory-editing effect of the reconsolidation–extinction procedure was specific to the reactivated memory.

Finally, to investigate the longevity of the effect, the authors brought a sample of their volunteers back to the lab one year later, and gave them a reminder shock. Remarkably, those who had received extinction training within 10 minutes of the reminder trial the year before continued to be immune to the shock. Taken together, Schiller and colleagues' results thus show that updating windows exist in humans, that the effects of extinction training during this window are stimulus-specific, and that the effects last for an extended period not commonly observed for other experiments in this field.

Schiller *et al.* studied healthy volunteers, but an exciting possibility is that their findings might be useful for the treatment of anxiety disorders such as PTSD. Current therapies involve extinction-based exposure to memory cues, but because extinction training is less effective in PTSD[7], pharmacological methods are being explored to augment fear extinction[8], or to block fear reconsolidation[9,10]. The obvious advantage of Schiller and colleagues' reconsolidation–extinction method is that no drugs are required, only a modification of the timing of standard exposure therapy.

There are, however, several issues that need to be carefully examined with regard to the potential clinical efficacy of this approach[2]. The aversive stimulus used in the study was a mild electric shock, which might have quite distinct effects from the kind of life-threatening events that lead to PTSD. Furthermore, it is not clear whether Schiller and colleagues' method would be effective for modifying fear memories acquired months or years before extinction training, rather than in the 24-hour period of their experiments. Finally, PTSD is a complex disorder that involves symptoms such as avoidance of traumatic reminders, emotional numbing, nightmares, flashbacks and sleep disturbances. The extent to which all these symptoms depend on aversive associations that are susceptible to editing remains to be determined. Nevertheless, Schiller and colleagues' findings are an exciting development that paves the way for mechanistic studies, in both rodents and humans, to discover how memory retrieval prepares fear circuits for updating by extinction training. ■

Gregory J. Quirk is in the Department of Psychiatry, University of Puerto Rico School of Medicine, San Juan, Puerto Rico 00936-5067, USA. Mohammed R. Milad is in the Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129, USA.
e-mails: gjquirk@yahoo.com; milad@nmr.mgh.harvard.edu

1. Nader, K. M., Schafe, G. E. & Le Doux, J. E. *Nature* **406,** 722–726 (2000).
2. Schiller, D. *et al. Nature* **463,** 49–53 (2010).
3. McGaugh, J. L. *Science* **287,** 248–251 (2000).
4. Nader, K. & Hardt, O. *Nature Rev. Neurosci.* **10,** 224–234 (2009).
5. Alberini, C. M. *Trends Neurosci.* **28,** 51–56 (2005).
6. Monfils, M.-H., Cowansage, K. K., Klann, E. & LeDoux, J. E. *Science* **324,** 951–955 (2009).
7. Milad, M. R. *et al. Biol. Psychiatry* **66,** 1075–1082 (2009).
8. Davis, M., Ressler, K., Rothbaum, B. O. & Richardson, R. *Biol. Psychiatry* **60,** 369–375 (2006).
9. Brunet, A. *et al. J. Psychiatr. Res.* **42,** 503–506 (2008).
10. Kindt, M., Soeter, M. & Vervliet, B. *Nature Neurosci.* **12,** 256–258 (2009).

QUANTUM PHYSICS

Trapped ion set to quiver

Christof Wunderlich

The peculiar ultra-fast trembling motion of a free electron — the Zitterbewegung predicted by Erwin Schrödinger in 1930 when he scrutinized the Dirac equation — has been simulated using a single trapped ion.

In seeking to investigate the properties of a particular system, natural scientists often encounter situations in which the difficulty in accessing and tuning the system of interest experimentally prevents such investigation being made. An effective, and widely used, remedy for these unfortunate instances is the numerical simulation of the system's properties and behaviour on a computer. However, in many cases, faithful digital replication of the system is not possible because of limitations in computing power and memory. These limitations become prohibitive for systems governed by the laws of quantum mechanics, not least for many-body quantum systems, because the range of possible system states grows exponentially with the number of system constituents.

In such cases, new insight may be provided by a quantum simulation, which simulates a quantum system using a different, experimentally accessible and controllable quantum system. On page 68 of this issue, Gerritsma *et al.*[1] report their use of a single atomic ion trapped in an electrodynamic cage to simulate a free particle (for instance, an electron) in an extremely fast quivering motion superimposed on a slow drift — the Zitterbewegung, as it is known, which was first predicted by Erwin Schrödinger in 1930 but has so far not been directly accessible to experiments.

In the late 1920s, Paul Dirac succeeded in devising an equation — the Dirac equation — that married two descriptions of the physical world, each of which had already revolutionized our view of it: quantum mechanics and special relativity. This equation describes the quantum-mechanical behaviour of half-integer-spin particles, taking into account the fundamental principles of special relativity — for example, that the speed of light in a vacuum is the ultimate speed limit at which information can be transferred across distances in the Universe.

Non-relativistic quantum mechanics predicts phenomena that are difficult to reconcile with our classical perception of the world. For example, quantum-mechanical superposition states, in which a particle simultaneously occupies separate regions of space, are hard to envisage, but cleverly designed wave-interference experiments reveal that such unexpected behaviour is possible. Adding special relativity to the mix results in even more perplexing phenomena. In interpreting the solutions of his relativistic quantum-mechanical equation, Dirac postulated the existence of an anti-particle to the electron — the positron. Although initially seen as a daring prediction, positrons were observed shortly thereafter, and today are routinely used for medical imaging.

Other predictions of the Dirac equation have remained elusive, particularly Schrödinger's Zitterbewegung, which arises from the interference of particle states that are interpreted to have positive and negative energies. This is a prediction of the Dirac equation that describes a 'free' particle — that is, one that is not subject to external forces and yet changes its velocity, in blatant conflict with Isaac Newton's second law of motion in classical mechanics.

The 'art' of a quantum simulation lies in the faithful reproduction of the Hamiltonian (a mathematical entity from which the system's static and dynamic properties can be derived) of the quantum system we want to learn about using a system we can experiment with[2–4]. The experiment performed by Gerritsma *et al.*[1] was designed such that each quantity appearing in the Hamiltonian of a trapped ion mirrors a quantity in the Hamiltonian of a free relativistic quantum particle (a free Dirac particle, for instance an electron). Two of the ion's internal energy states represent positive- and negative-energy states of a free Dirac particle; and the position and momentum of the trapped ion simulate the position and momentum of the free Dirac particle. To reproduce the (one-dimensional) Dirac Hamiltonian, the authors irradiate the ion with laser light, which allows the ion's motion in one dimension to be coupled to the two internal energy states.

By adjusting the intensity and frequency of the laser, Gerritsma and colleagues could vary at will the effective mass of the simulated free Dirac particle and the effective speed of light, which appears in the Dirac equation and constrains the particle's motion. They first observed the Zitterbewegung for an ion with zero average momentum, the internal states of which would be in a superposition (corresponding to the superposition of the positive- and negative-energy states of a free Dirac particle) with equal relative strengths. The frequency of this quasi-periodic motion extends from about 10 kHz to 80 kHz — a range that was accessible in the authors' experiments.

Next, the researchers created another superposition state of positive- and negative-energy states, but one in which these two components moved in opposite directions. They observed that the Zitterbewegung disappears as soon as these parts leave the space they had initially jointly occupied. Furthermore, they showed that a pure negative-energy state results in no Zitterbewegung. These results, obtained by controlling the ion's initial state, confirm that it is indeed the interference between positive- and negative-energy states that gives rise to the Zitterbewegung. When the authors changed the particle's effective mass and kept its momentum constant, both in the non-relativistic limit (large effective mass) and in the highly relativistic case (small effective mass), the Zitterbewegung disappeared, whereas this quivering motion was clearly present in the regime in between.

The measurement of the ion's average position as it evolves in time requires exacting experimental control, because it needs to be carried out with a precision of a few nanometres to be able to resolve the Zitterbewegung. Gerritsma *et al.* achieve this precision by mapping, using a sequence of laser pulses, the ion's motion onto its internal states, which can in turn be measured through the detection of scattered laser light.

Gerritsma and colleagues' experiment[1] not only demonstrates a much-sought-after effect in a real system, but also marks important progress in bringing quantum simulations closer to yielding new insight even in scientific fields that lie beyond the realm of quantum-information science. Trapped ions[5], neutral atoms[6], superfluids[7] or optical fields[8] may be used to further our understanding of relativistic quantum mechanics and astrophysical processes. Furthermore, simulating many-body physical phenomena with neutral atoms may help in deciphering hitherto unsolved problems in condensed-matter physics — for instance, the nature of high-temperature superconductivity[4,9]. Similarly, internal states of a collection of trapped ions can be made to interact as particle spins do, with the interaction strength designed by the experimenter[10]. Thus, trapped ions could be used to investigate phenomena such as quantum magnetism[11,12]. Finally, coupled cavity arrays in the solid state, although still in their infancy in terms of experimental work, offer promise for quantum simulations[13,14].

Although quantum-information research progresses in unforeseen, and sometimes spectacular, steps, building a universal quantum computer — one that would be able to simulate other quantum systems and thus solve problems that are intractable on a classical computer — still poses formidable challenges. Specialized quantum simulations, such as that performed by Gerritsma and colleagues[1], promise to be a versatile and, at the same time, more amenable scientific tool. ■

Christof Wunderlich is in the Department of Physics, University of Siegen, Siegen 57068, Germany.
e-mail: wunderlich@physik.uni-siegen.de

1. Gerritsma, R. *et al. Nature* **463,** 68–71 (2010).
2. Feynman, R. P. *Int. J. Theor. Phys.* **21,** 467–488 (1982).
3. Lloyd, S. *Science* **273,** 1073–1078 (1996).
4. Jané, E., Vidal, G., Dür, W., Zoller, P. & Cirac, J. I. *Quant. Inform. Comput.* **3,** 15–37 (2003).
5. Menicucci, N. C. & Milburn, G. J. *Phys. Rev. A* **76,** 052105 (2007).
6. Retzker, A., Cirac, J. I., Plenio, M. B. & Reznik, B. *Phys. Rev. Lett.* **101,** 110402 (2008).
7. Unruh, W. G. & Schützhold, R. *Quantum Analogues: From Phase Transitions to Black Holes and Cosmology* (Springer, 2007).
8. Philbin, T. G. *et al. Science* **319,** 1367–1370 (2008).
9. Bloch, I., Dalibard, J. & Zwerger, W. *Rev. Mod. Phys.* **80,** 885–964 (2008).
10. Wunderlich, C. in *Laser Physics at the Limit* (eds Figger, H., Meschede, D. & Zimmermann, C.) 261–271 (Springer, 2002).
11. Porras, D. & Cirac, J. I. *Phys. Rev. Lett.* **92,** 207901 (2004).
12. Friedenauer, A., Schmitz, H., Glueckert, J. T., Porras, D. & Schaetz, T. *Nature Phys.* **4,** 757–761 (2008).
13. Hartmann, M. J., Brandão, F. G. S. L. & Plenio, M. B. *Nature Phys.* **2,** 849–855 (2006).
14. Angelakis, D. G., Santos, M. F. & Bose, S. *Phys. Rev. A* **76,** 031805(R) (2007).

VIROLOGY

Bornavirus enters the genome

Cédric Feschotte

A survey of mammalian genomes has unexpectedly unearthed DNA derived from bornaviruses, leading to speculation about the role of these viruses in causing mutations with evolutionary and medical consequences.

Some people might find it disquieting that a hefty 8% of human genetic material originates not from our vertebrate ancestors but from viruses. The assimilation of viral sequences into the host genome is a process referred to as endogenization. It occurs when viral DNA integrates into a chromosome of reproductive germline cells and is subsequently passed from parent to offspring. Until now, retroviruses were the only viruses known to generate such endogenous copies in vertebrates. But on page 84 of this issue, Horie *et al.*[1] report that non-retroviral viruses called bornaviruses have been endogenized repeatedly during mammalian evolution. The finding unveils bornaviruses as a potential cause of mutation and also as an unforeseen source of genomic innovation (Fig. 1).

Borna disease virus (BDV) owes its name to the town of Borna, Germany, the site of a dreadful virus epidemic that decimated a regiment of cavalry horses in 1885. However, it is only recently that BDV has been characterized genetically: it belongs to the order Mononegavirales, and is a negative-sense RNA virus (in which the single-stranded RNA genome has the opposite sequence to messenger RNA). BDV infects a range of birds and mammals, including humans, and is unique among RNA viruses in that it naturally infects only neurons, establishing a persistent infection in its host's brain. In addition, the entire life cycle of BDV takes place in the nucleus of the infected cells, and does not require chromosomal integration[2]. This intimate association of BDV with the cell nucleus prompted Horie *et al.* to investigate whether bornaviruses may have left behind a record of past infection in the form of endogenous elements.

Horie *et al.* searched the 234 currently available eukaryotic genomes for sequences that are similar to that of BDV, and unearthed a plethora of endogenous Borna-like N (EBLN) elements in diverse mammals. The sequences of these elements resemble the nucleoprotein (*N*) gene of BDV, which encodes a structural protein involved in packaging the viral RNA into a nucleocapsid[2]. The authors show[1] that bornavirus endogenization has occurred in multiple mammalian lineages and at different times, ranging from more than 40 million years ago in anthropoid primates to less than 10 million years ago in squirrels. These molecular fossils add to the growing evidence[3–6] for the long-term coevolution of RNA viruses and their mammalian hosts.

All instances of endogenization described by Horie *et al.*[1] correspond to the *N* gene, and although most EBLN sequences are fragmentary and seem to be non-functional (they have decayed into pseudogenes), surprisingly, two EBLNs in the human genome are annotated as protein-coding genes. They retain long open reading frames (sequences that seem to encode proteins) and are transcribed from

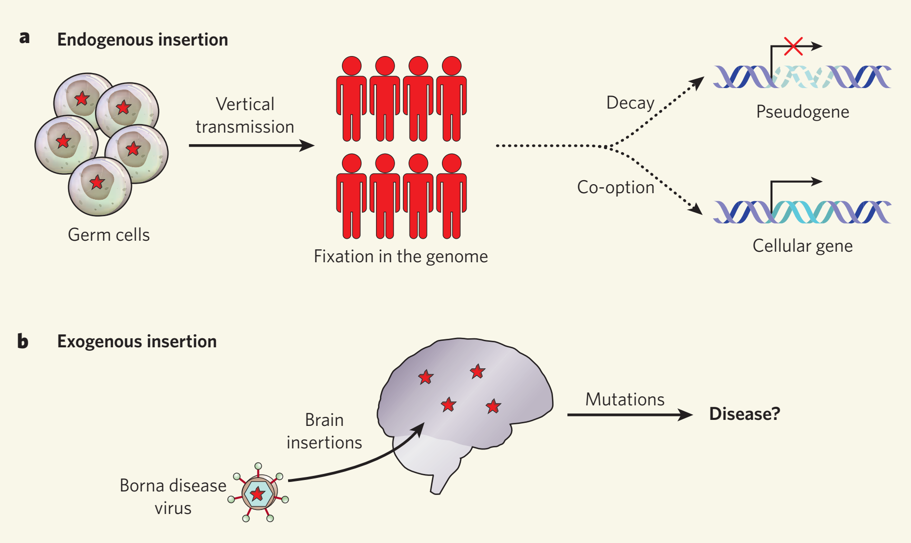


Figure 1 | Bornavirus in the genome, for better or worse. a, Horie *et al.*[1] report that bornavirus gene sequences (red stars) became integrated into the germline of our ancestors, and through vertical transmission (by conventional inheritance) have become 'fixed' in the genome, thereby becoming endogenous viral insertions. A fixed viral insertion can follow one of two evolutionary fates: it can either decay into a pseudogene or be co-opted to form a new gene whose product has a cellular function. **b**, Circulating bornavirus sequences can become integrated into the genome of brain cells (the current target of Borna disease virus) after infection (exogenous insertion). These sequences are not heritable, but might cause mutations that interfere with brain function and may contribute to the development of psychiatric disorders.

DNA into mRNAs in the various tissues and cell lines examined[1]. Also, one of the two proteins (LOC340900) has been reported[7] to interact with several well-known cellular proteins. Thus, the discovery of EBLNs uncovers two cases of viral DNA that has apparently been co-opted to form cellular genes (Fig. 1a). Although it is not known whether EBLN-derived proteins are functional in human cells, these proteins may have been usurped by the host, at least initially, to serve an antiviral function. There are precedents for this in mice and sheep[8,9], in which endogenous retroviral capsid proteins offer protection against exogenous retroviral infections.

How are EBLN elements generated? Unlike retroviruses, BDV does not need to integrate into the host DNA to replicate, and therefore the virus genome does not encode the machinery for reverse transcription of its RNA into DNA. Despite this, using the polymerase chain reaction, Horie *et al.*[1] were able to detect BDV DNA in various infected cell lines and in the brain of persistently infected mice. Furthermore, after infecting human cells for 30 days, the authors could isolate chromosomally integrated BDV DNA along with flanking host genomic sequences. These BDV insertions resemble EBLN elements in that they are derived from the *N* gene and exhibit the hallmarks of retroposition (the process by which RNA is integrated into a DNA genome), including a stretch of adenine nucleotides at the 3′ end of the inserted element and a short duplication of the target site.

In mammalian genomes, retroposition is primarily driven by the activity of L1 long interspersed nucleotide elements — pieces of mobile DNA that make copies of themselves and reinsert into the genome[10]. L1 has colonized the genome of mammals for more than 100 million years and continues to replicate actively in several species, including humans. The L1 enzymatic machinery can reverse transcribe its own RNA into DNA, but can also act on non-L1 RNA templates — throughout evolution, this promiscuity has caused the bombardment of mammalian genomes with millions of DNA inserts[10]. The abundance of BDV RNA in the nucleus of persistently infected cells, coupled with some peculiar properties of the *N* gene RNA, might have promoted their fortuitous recognition by the L1 machinery.

The fact that Horie and colleagues[1] could readily detect BDV DNA and chromosomal insertions in human cells suggests that BDV retroposition might occur at an appreciable frequency during BDV infection, creating a source of mutation in infected individuals (Fig. 1b). This yields a tantalizing and testable hypothesis for the alleged, but still controversial, causative association of BDV infection with certain psychiatric disorders, such as schizophrenia and mood disorders[2,11]. This possibility becomes even more intriguing when considering the recent demonstration of L1 hyperactivity in the human brain[12], the primary site of BDV infection. ■

Cédric Feschotte is in the Department of Biology, University of Texas, Arlington, Texas 76016, USA.
e-mail: cedric@uta.edu

1. Horie, M. *et al. Nature* **463,** 84–87 (2010).
2. Tomonaga, K., Kobayashi, T. & Ikuta, K. *Microbes Infect.* **4,** 491–500 (2002).
3. Katzourakis, A. *et al. Science* **325,** 1512 (2009).
4. Gilbert, C., Maxfield, D. G., Goodman, S. M. & Feschotte, C. *PLoS Genet.* **5,** e1000425 (2009).
5. Gifford, R. J. *et al. Proc. Natl Acad. Sci. USA* **105,** 20362–20367 (2008).
6. Holmes, E. C. *Annu. Rev. Microbiol.* **62,** 307–328 (2008).
7. Ewing, R. M. *et al. Mol. Syst. Biol.* **3,** 89 (2007).
8. Best, S., Le Tissier, P., Towers, G. & Stoye, J. P. *Nature* **382,** 826–829 (1996).
9. Arnaud, F., Murcia, P. R. & Palmarini, M. *J. Virol.* **81,** 11441–11451 (2007).
10. Cordaux, R. & Batzer, M. A. *Nature Rev. Genet.* **10,** 691–703 (2009).
11. Rott, R. *et al. Science* **228,** 755–756 (1985).
12. Coufal, N. G. *et al. Nature* **460,** 1127–1131 (2009).

PALAEONTOLOGY

Muddy tetrapod origins

Philippe Janvier and Gaël Clément

The tracks left by organisms are among the most difficult of fossils to interpret. But just such evidence puts debate about the origins of four-limbed vertebrates (which include ourselves) on a changed footing.

The term 'tetrapodomorph fishes' scarcely rolls off the tongue, but these are fossil animals that have a special place in the evolutionary history of vertebrates. It was through the stepwise transformation of paired fins in this lineage of lobe-finned fishes that paired limbs with digits arose, marking the advent of the four-limbed vertebrates, or tetrapods. This event occurred sometime during the Devonian period, between 416 million and 359 million years (Myr) ago. On page 43 of this issue, Niedźwiedzki *et al.*[1] describe fossil tracks that were clearly made by a four-limbed animal possessing digits (see image on the cover of this issue). But they date to a time well before tetrapods were thought to have existed.

The temporal and taxonomic context for this discovery is outlined in Figure 1. The earliest complete evidence for limbs with distinct digits is provided by the articulated skeletons of the iconic early tetrapods *Ichthyostega* and *Acanthostega*, which date to the Famennian stage (374–359 Myr ago)[2]. But the record extends much further back, to the Frasnian and possibly the late Givetian (385 Myr ago), thanks to the identification of several skeletal 'signatures' that allow the characterization of isolated tetrapod bones, even in the absence of actual limb bones. Meanwhile, palaeontologists provided strong evidence that the elpistostegalians, a group of large lobe-finned fishes dating to the Givetian/Frasnian boundary, were the closest fossil piscine relatives of tetrapods[2].

The fish–tetrapod transition was thus seemingly quite well documented. There was a consensus that the divergence between some elpistostegalians (such as *Tiktaalik* or *Panderichthys*)[3] and tetrapods might have occurred during the Givetian, 391–385 Myr ago. Coeval with the earliest fossil tetrapods, trackways dating to the Late Devonian were evidence for their ability to walk or crawl on shores[2].

Now, however, Niedźwiedzki *et al.*[1] lob a grenade into that picture. They report the stunning discovery of tetrapod trackways with distinct digit imprints from Zachełmie, Poland, that are unambiguously dated to the lowermost Eifelian (397 Myr ago). This site (an old quarry) has yielded a dozen trackways made by several individuals that ranged from about 0.5 to 2.5 metres in total length, and numerous isolated footprints found on fragments of scree. The tracks predate the oldest tetrapod skeletal remains by 18 Myr and, more surprisingly, the earliest elpistostegalian fishes by about 10 Myr. The implication is that both groups have a very long 'ghost range' — that is, a period of time during which members of the groups should have been present but for which no body fossils have yet been found.

Trace fossils — footprints, trackways or trails — are fascinating but often frustrating sources of information. Body fossils of the track makers almost never occur in the same rock beds, so complicating interpretation. Moreover, tracks are often blurred because they were made on a muddy substrate, or are deformed by artefacts such as gas bubbles or traces left by floating objects. It is especially important to rule out such factors when, as in the case of the Zachełmie tracks, the trace fossils are the only evidence of an animal group at a time when that group is presumed to have not yet existed.

Niedźwiedzki *et al.* provide convincing evidence that the tracks are indeed the traces left by an animal with four limbs bearing digits. They were made in the mud of a shallow marine lagoon, but no known coeval animal other than an elusive tetrapod could have left such imprints. Although large, ancient arthropods (crustaceans or extinct sea scorpions) can leave impressive tracks[4], with traces of claws, they do not display the diagonal pattern seen here. The lack of a body drag in the Zachełmie trackways is explained by the fact that the track makers were still floating in the water while walking on the muddy bottom. The match between these tracks and the limb anatomy of *Ichthyostega* and *Acanthostega*[2] is impressively close: were similar tracks to be found in Famennian/Frasnian rocks, they would be readily attributed to an *Ichthyostega*-like animal, as were the previously reported Late Devonian trackways (Fig. 1).

But we are left with a puzzling mismatch with the currently accepted timing for the elpistostegalian–tetrapod divergence[2,3], as depicted in Figure 1. The 397-Myr (early Eifelian) age of the Zachełmie tracks implies that limbs with digits arose earlier, during the preceding Emsian, a rather long stage (about 10 Myr) that has yielded very few lobe-finned fishes, none of which could be readily regarded as potential tetrapod or elpistostegalian ancestors. The earliest and most primitive tetrapodomorph fish, *Kenichthys*[5], is only late Emsian in age. The earlier Pragian and Lochkovian stages (416–407 Myr ago) yield only lobe-finned fishes that are anatomically different from tetrapodomorphs. That said, a possible tetrapodomorph has been recorded from the Pragian of China, and awaits description[6].

Perhaps the earlier stages of the elpistostegalian and tetrapod lineages include elusive pre-Eifelian tetrapodomorph fishes that we may not be able to recognize as such, because the known elpistostegalians and early tetrapods are too 'derived' — that is, maybe they are too different from their common ancestor as we imagine that ancestor to have been. Possibly, the latter looked like an 'osteolepidid' (Fig. 1), a tetrapodomorph fish with diamond-shaped scales and head bones covered with cosmine (a shiny tissue made of enamel and dentine), which is very different in aspect from the crocodile-like, vermiculate ornamentation of elpistostegalians and early tetrapods. Such a view would be broadly in accord with a recently proposed tetrapodomorph evolutionary tree[7].

Niedźwiedzki and colleagues' apparently anachronistic Eifelian tetrapod trackways[1] will thus shake up thinking about tetrapod origins. They show that the first tetrapods thrived in the sea, trampling the mud of coral-reef lagoons; this is at odds with the long-held view that river deltas and lakes were the necessary environments for the transition from water to land during vertebrate evolution. And in guiding the search for a gradual timing of the fin–limb transition during the Middle Devonian, they are likely to trigger a burst of field investigations into potential tetrapodomorph fish sites of Emsian or earlier age. ■

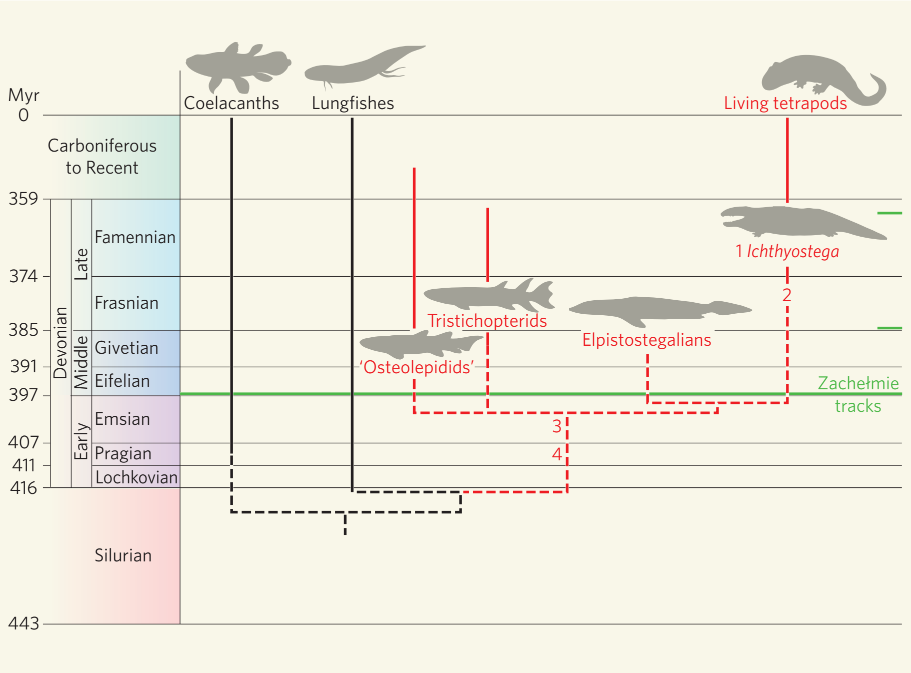


Figure 1 | Simplified evolutionary tree of the living and fossil lobe-finned fishes and tetrapods. The age of the previously identified Devonian tetrapod trackways (short green bars) contrasts with the 397-Myr-old Zachełmie tracks identified by Niedźwiedzki and colleagues[1]. These newly discovered tracks generate a mismatch between the currently accepted tree of tetrapodomorph fishes (lobe-finned fishes with internal nostrils) and its timing based on the body-fossil record (shown by solid red lines). The temporal mismatch implies the existence of long 'ghost ranges' (dashed red lines) among Devonian tetrapodomorphs. The divergence between elpistostegalian fishes and tetrapods with limbs and digits must have occurred much earlier than previously thought, perhaps during the 10-Myr-long Emsian stage, from which only few tetrapodomorph fishes are recorded. 1, Earliest articulated tetrapod skeletons with limbs and digits (*Ichthyostega*, *Acanthostega*)[2]; 2, earliest isolated tetrapod bones; 3, earliest tetrapodomorph fish (*Kenichthys*)[5]; 4, possible earlier tetrapodomorph fish[6].

Philippe Janvier and Gaël Clément are at the Muséum National d'Histoire Naturelle (CNRS UMR7207), 8 rue Buffon, 75231 Paris Cedex 05, France.
e-mails: janvier@mnhn.fr; gclement@mnhn.fr

1. Niedźwiedzki, G., Szrek, P., Narkiewicz, K., Narkiewicz, M. & Ahlberg, P. E. *Nature* **463,** 43–48 (2010).
2. Clack, J. A. *Gaining Ground* (Indiana Univ. Press, 2002).
3. Schubin, N. H., Daeschler, E. B. & Jenkins, F. A. Jr *Nature* **440,** 764–771 (2006).
4. Whyte, M. A. *Nature* **438,** 576 (2005).
5. Zhu, M. & Ahlberg, P. E. *Nature* **432,** 94–97 (2004).
6. Lu, J. & Zhu, M. *J. Vert. Paleont.* **29,** suppl. to no. 3, 134A (2009).
7. Long, J. A. *et al. Nature* **444,** 199–202 (2006).

ARTICLES

Tetrapod trackways from the early Middle Devonian period of Poland

Grzegorz Niedźwiedzki[1], Piotr Szrek[2,3], Katarzyna Narkiewicz[3], Marek Narkiewicz[3] & Per E. Ahlberg[4]

The fossil record of the earliest tetrapods (vertebrates with limbs rather than paired fins) consists of body fossils and trackways. The earliest body fossils of tetrapods date to the Late Devonian period (late Frasnian stage) and are preceded by transitional elpistostegids such as *Panderichthys* and *Tiktaalik* that still have paired fins. Claims of tetrapod trackways predating these body fossils have remained controversial with regard to both age and the identity of the track makers. Here we present well-preserved and securely dated tetrapod tracks from Polish marine tidal flat sediments of early Middle Devonian (Eifelian stage) age that are approximately 18 million years older than the earliest tetrapod body fossils and 10 million years earlier than the oldest elpistostegids. They force a radical reassessment of the timing, ecology and environmental setting of the fish–tetrapod transition, as well as the completeness of the body fossil record.

The last quarter-century has seen a dramatic expansion in the known body fossil record of Devonian tetrapods, the earliest known limbed vertebrates[1–21]. Equally importantly, the discovery of articulated specimens of elpistostegids, the animals that fall immediately below them in the tetrapod stem group, has greatly enhanced our understanding of the origin of tetrapod morphology[22–31]. Elpistostegids such as *Panderichthys* and *Tiktaalik* show a tetrapod-like head and body shape combined with the retention of 'fish' characters such as paired fins[23,28,29] and the absence of a sacrum[28]. Their close similarity to Devonian tetrapods and stable phylogenetic position below the latter in the tetrapod stem group[23,29,32] provide a morphological outline of the fish–tetrapod transition.

In parallel with this expansion of the morphological data set, the environmental, ecological and temporal contexts of the transition have been reassessed. It has become clear that many of the earliest tetrapods and elpistostegids derive from brackish to marginal marine deposits, and their wide geographical distribution also points to marine tolerance[13,16,19]. Temporally, the earliest record of tetrapod morphology has been pushed back from the late Famennian (about 360 million years ago) to the late Frasnian (about 375 million years ago)[3,6,9,33]. Known elpistostegids range from late Givetian to mid-Frasnian (approximately 386 to 380 million years ago), and the Frasnian *Elpistostege* and *Tiktaalik* appear more derived than the Givetian *Panderichthys*[12,29], suggesting a good fit between stratigraphy and phylogeny, with tetrapods originating sometime during the mid–late Frasnian. Many recent publications argue that tetrapods evolved from and rapidly replaced the elpistostegids, probably in brackish to freshwater environments, in response to the modification of the terrestrial and water's edge environment caused by the development of extensive tree-sized land vegetation[21]. However, a few data points have clashed with this consensus picture. Notably, the fragmentary genus *Livoniana*, although Givetian and thus contemporary with *Panderichthys*, is more derived than *Tiktaalik*, judging from its limited preserved anatomy[12].

Supposed trackways of very early tetrapods have been recorded from a number of localities in Europe and Australia[34–39]. The most securely identified of these, the Genoa River trackways from Australia, are Late Devonian (probably Famennian) in age[34,37]. Two large trackways from Valentia Island, Ireland[36], have been dated radiometrically to 385 million years ago. At the time of publication this was taken to imply an Eifelian (early Middle Devonian) age[39], which clashed with the occurrence of the Late Devonian index fossil (for Laurussia) *Bothriolepis* in the same strata. However, subsequent recalibration of the timescale indicates that 385 million years ago corresponds to the Givetian–Frasnian boundary[33]. This is consonant with the biostratigraphy but nevertheless suggests an earlier origin for tetrapods than indicated by the body fossil data.

Our discovery of diagnostic and securely dated tetrapod tracks from the marine Eifelian (early Middle Devonian) of Poland shows that the current consensus based on body fossils is substantially mistaken in both the timescale and, probably, the environmental setting of the fish–tetrapod transition.

The locality

The northern Łysogóry region of the Holy Cross Mountains (Góry Świętokrzyskie) in south-eastern Poland contains an extensive and well-dated sequence of marine Middle Devonian strata (Fig. 1)[40–45]. In the disused Zachełmie Quarry, the lower part of the Kowala Formation and the upper part of the Wojciechowice Formation are exposed. The trackway horizon lies within the Wojciechowice Formation, some 20 m below the level where a conodont sample showing a characteristic *costatus* Zone assemblage (Eifelian) was taken[42]. The Eifelian age of the formation is also indirectly confirmed by previous biostratigraphic data obtained from the underlying and overlying strata exposed in other sections[43–45]. It can be securely assigned to the lower–middle Eifelian, corresponding to an age of approximately 395 million years (see Supplementary Information).

The Wojciechowice Formation represents a unique episode of restricted, extremely shallow-water carbonate sedimentation within the generally open marine marly carbonate and, subordinately, siliciclastic deposition that prevailed during the Middle Devonian in the northern Holy Cross Mountains. The lower, trackway-containing part of this formation, almost devoid of other fossils, contains abundant laminites with desiccation cracks and raindrop impressions and seems to represent an extremely shallow marine tidal, perhaps lagoonal, environment. The tetrapod trackway assemblage is not only the earliest but by far the richest from the Devonian. What follows is a preliminary

[1]Department of Paleobiology and Evolution, Faculty of Biology, Warsaw University, 2S. Banacha Street, 02-097 Warsaw, Poland. [2]Department of Paleontology, Faculty of Geology, Warsaw University, 93 Żwirki i Wigury Street, 02-089 Warsaw, Poland. [3]Polish Geological Institute, 4 Rakowiecka Street, 00-975 Warsaw, Poland. [4]Subdepartment of Evolutionary Organismal Biology, Department of Physiology and Developmental Biology, Uppsala University, Norbyvägen 18A, 752 36 Uppsala, Sweden.

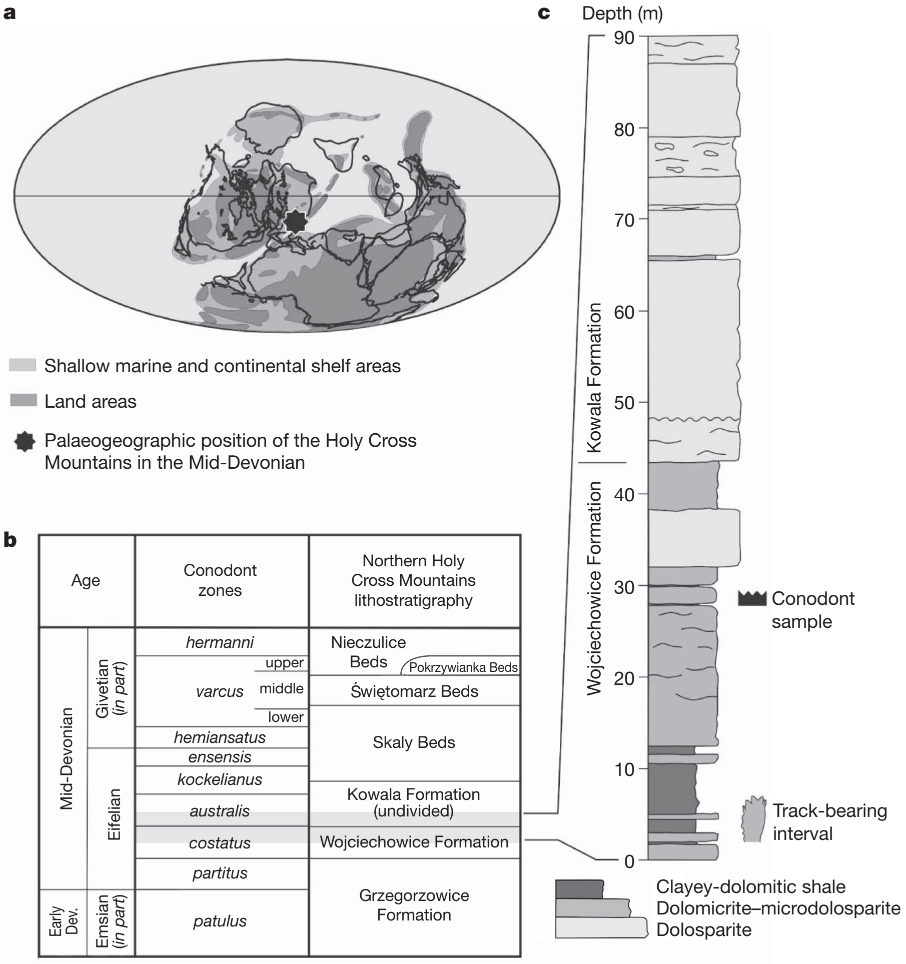


Figure 1 | The locality. a, Palaeogeographic position of the Holy Cross Mountains during Middle Devonian (map from ref. 50). **b**, Middle Devonian lithostratigraphy of the northern Holy Cross Mountains. **c**, Zachełmie quarry section showing locations of trackway horizon and *costatus* Zone conodont sample.

description: further study, including ichnotaxonomic description of the trackways, is in progress.

Trackways

The footprint fossils comprise numerous trackways of different sizes and characteristics, as well as a large number of isolated prints and a densely trampled surface. The appearance of the prints varies greatly depending on the size of the animal, the condition of the sediment and the pattern of movement. Digit marks are variably present or absent, probably depending on the cohesiveness of the sediment. Some very crisp and detailed prints are almost certainly true prints, whereas others may represent underprints. For a detailed account of the range of print preservations at Zachełmie see Supplementary Information.

Muz. PGI 1728.II.16 (Fig. 2a) has distinct manus ('hand') and pes ('foot') prints of somewhat different size arranged in diagonal stride sequence. The animal is moving in a straight line and is not leaving a body drag. The prints are circular without digit impressions or displacement rims. A single, slightly larger print on the same slab (Fig. 2a) shows a strong posterior displacement rim with digit marks. We interpret this type of isolated print, which is seen frequently at Zachełmie, as an aquatic print where a swimming tetrapod has used a single limb to kick against the substrate. It should not be confused with the larger single prints discussed below, which have been found on small blocks in the quarry scree and may well derive from longer trackways.

Stride length, relative spacing of the prints, and the absence of a body drag demonstrate that Muz. PGI 1728.II.16 is a tetrapod trackway. Elpistostegids and other tetrapodomorph fishes all have straight 'knees' and 'elbows', and shoulder and hip joints that face posteriorly[23–26,28]. In early tetrapods, by contrast, the knees and (in particular) elbows of the short, sprawling limbs allow greater flexion, and the shoulder and hip joints face laterally. The relative stride length shows that both fore- and hindlimbs were oriented anterolaterally at the anterior extremity of the movement arc, with the manus and pes placed on the ground well anterior to the respective shoulder and hip joint (Fig. 2b). This would be impossible for an animal with the girdle morphologies documented in *Tiktaalik* and *Panderichthys*[23–26,28], and in any case the absence of a sacrum would prevent the fish from lifting its tail clear off the substrate (Fig. 2b). Assuming standard early tetrapod proportions[8,18,21], the total length of the track maker was probably in the region of 40–50 cm.

Muz. PGI 1728.II.15 (Fig. 2c), a track made by a slightly smaller animal, is an example of a second trackway type. Its stride pattern is partly ladder-like, suggesting that for a few strides the limbs were moved symmetrically rather than alternately. The strides are short and no distinction between manus and pes prints can be observed, suggesting that only one pair of appendages is represented. If this is a true track, preserving the actual sediment surface impressed by the feet, the animal was pushing itself along using only one pair of limbs. If, on the other hand, it is an undertrack, we cannot rule out the possibility that both pairs of limbs contacted the sediment, but if so the second pair must have carried less weight and made shallower impressions. (An undertrack is the indentation that a foot leaves in sub-surface sediment layers.) It may have been made subaquatically, by an animal using one pair of limbs, but a confident interpretation of this trackway type must await detailed examination of multiple examples.

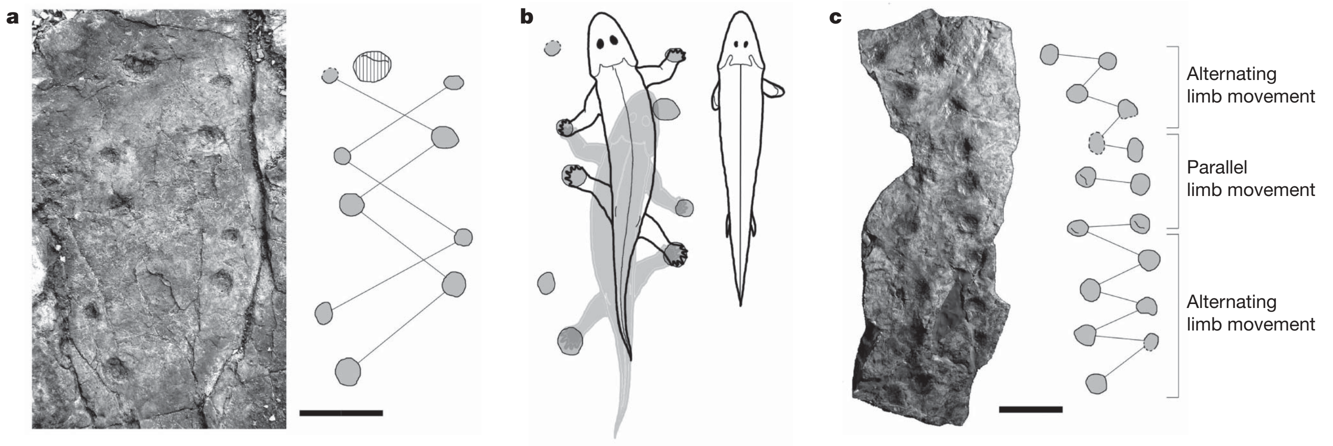


Figure 2 | Trackways. **a**, Muz. PGI 1728.II.16. (Geological Museum of the Polish Geological Institute). Trackway showing manus and pes prints in diagonal stride pattern, presumed direction of travel from bottom to top. A larger print (vertical hatching) may represent a swimming animal moving from top to bottom. **b**, On the left is a generic Devonian tetrapod based on *Ichthyostega* and *Acanthostega* (from ref. 18) fitted to the trackway. On the right, *Tiktaalik* (from ref. 29 with tail reconstructed from *Panderichthys*[23]) is drawn to the same shoulder–hip length. Positions of pectoral fins show approximate maximum 'stride length'. **c**, Muz. PGI 1728.II.15. Trackway showing alternating diagonal and parallel stride patterns. In **a** and **c**, photographs are on the left, interpretative drawings are on the right. Thin lines linking prints indicate stride pattern. Dotted outlines indicate indistinct margins and wavy lines show the edge of the displacement rim. Scale bars, 10 cm.

Individual prints

A number of large prints, collected from the quarry scree, provide information about the foot morphology in the largest Zachełmie tetrapods. In most instances, the foot is approximately 15 cm wide measured across the junction between sole and toe prints, more than twice the linear dimensions of the best-preserved *Ichthyostega* foot[2] and suggesting an animal about 2.5 m in length, but the largest print (Muz. PGI 1728.II.5) is 26 cm wide. Three prints, all representing the left pes, will be considered here (Figs 3 and 4); for others see Supplementary Information.

Muz. PGI 1728.II.3 (Fig. 3a) shows a large proximal displacement rim and long curved toe prints. Muz. PGI 1728.II.1 (Fig. 4a) and Muz. PGI 1728.II.2a,b (Fig. 3b) on the other hand have short, triangular toe impressions, as does Muz. PGI 1728.II.5. Their outlines are crisp. Muz. PGI 1728.II.1 has a moderate-sized, low, anterior displacement rim, whereas Muz. PGI 1728.II.2 has only a very small rim along the anterior margin of one of the toe impressions. We infer that the shared morphological features of Muz. PGI 1728.II.1, Muz. PGI 1728.II.2 and Muz. PGI 1728.II.5 reflect the morphology of the foot. By contrast, the large displacement rim of Muz. PGI 1728.II.3 suggests that the foot slipped during the formation of the print; the differences in toe shape between this and the other large prints may thus be an artefact, though a real morphological difference cannot be ruled out.

Muz. PGI 1728.II.1 and Muz. PGI 1728.II.2 both show three large triangular toe impressions, the anteriormost somewhat divergent from the other two. Posterior to these, Muz. PGI 1728.II.1 shows a slender toe that has not left an impression in Muz. PGI 1728.II.2. Anterior to the triangular toes, Muz. PGI 1728.II.1 shows a single slender anteriorly divergent toe whereas Muz. PGI 1728.II.2 shows two similar toes side by side. There are no claws. The tip of each triangular toe shows a small distinct cushion or pad, but there are no separate phalangeal pads, whereas such pads can be discerned faintly on the slender digits.

Comparison of Muz. PGI 1728.II.1 with known early tetrapod limb skeletons[2,7,8,21], all of which have short broad feet, indicates that the print includes the ventral surface of the lower leg and knee (Fig. 4a–c). It seems that the ankle was almost flat, as has been argued for *Ichthyostega* and *Acanthostega* on morphological grounds[8]. Supplementary Information 3 and 4 show a three-dimensional surface scan of Muz. PGI 1728.II.1.

Comparisons

Trackway Muz. PGI 1728.II.16 from Zachełmie is in many ways similar to previously described Devonian tetrapod tracks[34,36,37]. The trackways from Valentia Island (Ireland) and Tarbat Ness (Scotland), and one of the Genoa River tracks (Australia), all show similar

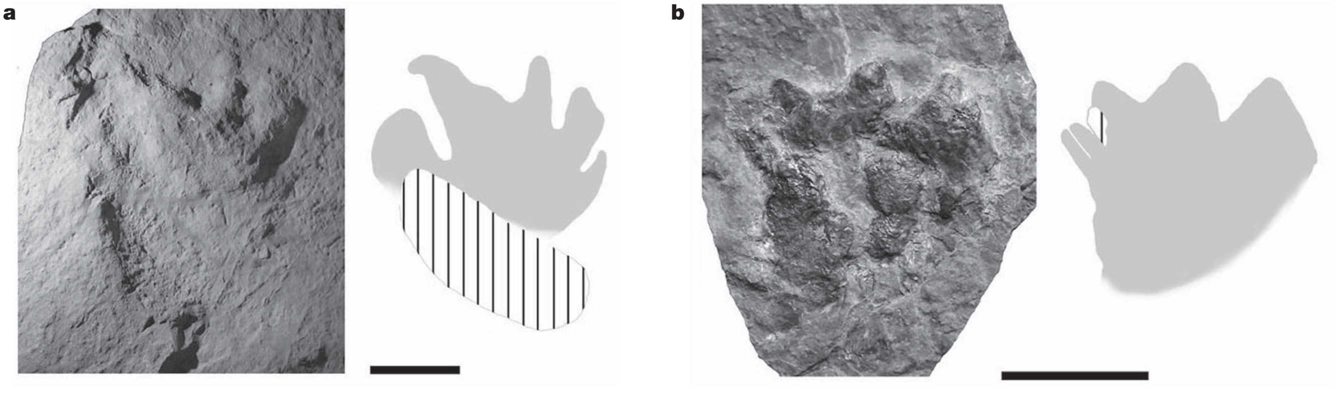


Figure 3 | Footprints. **a**, Muz. PGI 1728.II.3. Probable pes, preserved as natural cast (that is, mirror-imaged). Print with long digit impressions and large displacement rim, probably indicating slippage plus anticlockwise rotation of the foot. **b**, Muz. PGI 1728.II.2. Left pes, preserved as natural cast (that is, mirror-imaged). Photographs are on the left, interpretative drawings are on the right. In the drawings, grey indicates footprint, and vertical hatching indicates displacement rim. Scale bars, 10 cm.

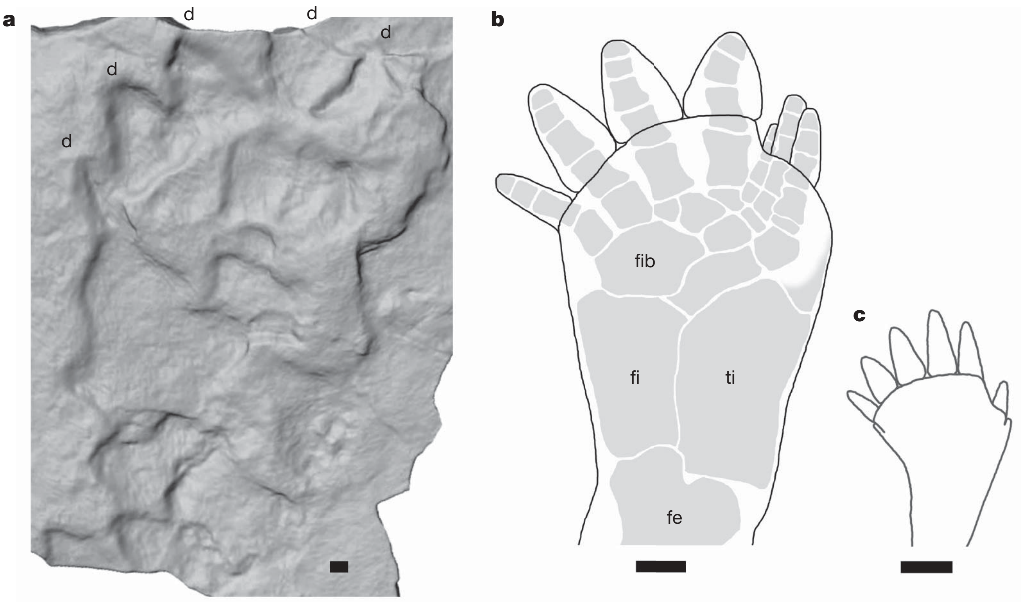


Figure 4 | Foot morphologies. a, Laser surface scan of Muz. PGI 1728.II.1, left pes. **b**, Complete articulated left hind limb skeleton of *Ichthyostega*, MGUH f.n. 1349, with reconstructed soft tissue outline. **c**, Left hind limb of *Acanthostega*, reconstructed soft tissue outline based on skeletal reconstruction in ref. 8. We note the large size of the print compared to the limbs of *Ichthyostega* and *Acanthostega*, and that the print appears to represent not just the foot but the whole limb as far as the knee. d, digit; fe, femur; ti, tibia; fi, fibula; fib, fibulare. Scale bars, 10 mm.

diagonal patterns of manus and pes prints without a body drag. These trackways are all demonstrably younger than Zachełmie. More problematic is the Glenisla trackway from Australia, which appears to be no later than Late Silurian[38]. The known body fossil record of this period includes stem sarcopterygians[46], but no tetrapodomorph lobe-fins or tetrapods. The trackway is ladder-like, a characteristic that has been used to argue against a tetrapod identity[37], but which is shared with some Zachełmie trackways. The best-preserved Zachełmie footprints are quite similar to the pes morphology of *Acanthostega* and, in particular, *Ichthyostega* (Fig. 4b, c). It is possible to reconstruct approximate footprint morphologies for the two latter genera, though with lower precision for *Acanthostega* because the pes skeleton is partly reconstructed.

Implications

The Zachełmie trackways show that very large stem-group tetrapods, exceeding 2 m in length, lived in fully marine intertidal to lagoonal environments along the south coast of Laurussia during the early Eifelian, some 18 million years before the earliest-known tetrapod body fossils were deposited. This forces us to infer much longer ghost lineages for tetrapods and elpistostegids than the body fossil record suggests (Fig. 5a, b). (Ghost lineages are those that must have existed at a particular time, according to the phylogeny, but which are not represented by fossils at that time.) Until now, the replacement of elpistostegids by tetrapods in the body-fossil record during the mid–late Frasnian has appeared to reflect an evolutionary event, with the elpistostegids as a short-lived 'transitional grade' between fish and tetrapod morphotypes (Fig. 5a). In fact, tetrapods and elpistostegids coexisted for at least 10 million years (Fig. 5b). This implies that the elpistostegid morphology was not a brief transitional stage, but a stable adaptive position in its own right. It is reminiscent of the lengthy coexistence of non-volant but feathered and 'winged' theropod dinosaurs with volant stem-group birds during the Mesozoic.

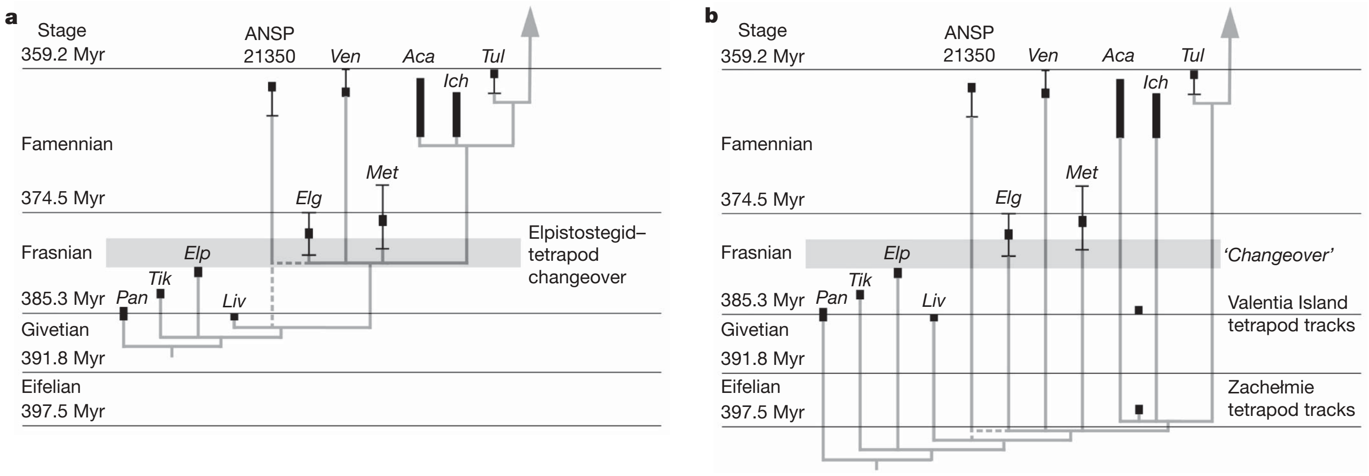


Figure 5 | Phylogenetic implications of tracks. a, Phylogeny of selected elpistostegids and stem tetrapods, based on refs 10, 12, 19 and 20, fitted to Devonian stratigraphy. The grey bar indicates replacement of elpistostegids by tetrapods in body fossil record. **b**, Effect of adding the Zachełmie tracks to the phylogeny: the ghost ranges of tetrapods and elpistostegids are greatly extended and the 'changeover' is revealed to be an artefact. *Pan, Panderichthys*; *Tik, Tiktaalik*; *Elp, Elpistostege*; *Liv, Livoniana*; *Elg, Elginerpeton*; *Ven, Ventastega*; *Met, Metaxygnathus*; *Aca, Acanthostega*; *Ich, Ichthyostega*; *Tul, Tulerpeton.* ANSP 21350 is an unnamed humerus described in ref 17. The bars are approximate measures of the uncertainty of dating. These are not statistical error bars but an attempt to reflect ongoing debate.

The Wojciechowice Formation represents a tidal flat environment or a lagoon in a broad shallow carbonate basin with little terrigenous input. This suggests that the origin of tetrapods occurred, not in the vegetated margins or surrounding seasonal 'flooded forest' environments of rivers, as has frequently been argued, but in the marine intertidal and/or lagoonal zone. Such a scenario has considerable explanatory power. The intertidal environment provides a ready food source of stranded marine animals on a twice-daily basis, in the immediate vicinity of the sea, and would thus have allowed marine ancestors of tetrapods gradually to acquire terrestrial competence while accessing a new and essentially untouched resource. The mid-Devonian riverine environment by contrast would probably not have provided any such reliable and easily captured terrestrial food source. Recent hypotheses about non-marine tetrapod origins have accordingly focused on other perceived benefits of limb-driven locomotion, either on land or in shallow water, relegating terrestrial feeding to a later stage of tetrapod evolution even though both dentition and sutural morphology indicate changes in feeding mechanics at the origin of tetrapods[10,21,47]. Under the intertidal hypothesis, these somewhat counterintuitive arguments become unnecessary.

The replacement of elpistostegids by tetrapods in the body fossil record of the mid–late Frasnian raises two questions: why do both groups have such long initial ghost ranges (a minimum of 10 and 18 million years, respectively), and why do the elpistostegids appear before tetrapods in the body fossil record in a manner that neatly simulates a stratophylogenetic fit (Fig. 5)? The first question is answered in part by the observation that the Wojciechowice Formation is almost devoid of body fossils; this environment was evidently not conducive to the preservation of skeletons. Contemporary vertebrate body fossil faunas, known mainly from the Baltic States and Scotland, come from rather different environments. The Baltic shallow marine strata are dominated by sandy terrigenous sediments[48] and the Scottish fossil assemblages derive from a succession of lakes within the Orcadian basin[49]. The absence of tetrapods in these deposits may simply be a matter of environmental preference. The false stratophylogenetic succession from elpistostegids to tetrapods is more of a puzzle. If their first appearance as body fossils reflects the time when they first colonized environments with preservation potential, as seems likely, the elpistostegids evidently arrived in advance of the tetrapods. The reason was presumably ecological but cannot be determined at present.

The discovery of the Zachełmie footprints substantially changes the context for future research on the origin of tetrapods. Intertidal laminites of Middle and Early Devonian age should be examined systematically for tetrapod tracks, and we should search for tetrapod and elpistostegid body fossils in associated marginal marine strata. For the present the timing of the fish–tetrapod transition is best regarded as uncertain, though it clearly pre-dates the early Eifelian; an Early Devonian date seems most likely, but even earlier potential tetrapod ichnofossils such as the Silurian Glenisla track should not be dismissed out of hand.

METHODS SUMMARY

The tracks were photographed in low-angle light to bring out the details, and were sometimes also highlighted with pigments. Tracks that could not be collected from the quarry were cast *in situ* using silicone rubber, and these silicone peels were then used as moulds to produce Jesmonite plaster replicas of the original surfaces. Surface scanning of Muz. PGI 1728.II.1 was performed in the Museum and Institute of Zoology of Polish Academy of Sciences using a three-dimensional Minolta VI–9i laser scanner. The movies of this scan, presented in Supplementary Information 7 and 8, were rendered using the rendering software package Rhino with the animation plug-in Bongo, both published by McNeel.

Received 21 July; accepted 29 October 2009.

1. Lebedev, O. A. The first find of a Devonian tetrapod in USSR. *Dokl. Akad. Nauk SSSR* [in Russian] **278**, 1407–1413 (1984).
2. Coates, M. I. & Clack, J. A. Polydactyly in the earliest known tetrapod limbs. *Nature* **347**, 66–67 (1990).
3. Ahlberg, P. E. Tetrapod or near-tetrapod fossils from the Upper Devonian of Scotland. *Nature* **354**, 298–301 (1991).
4. Ahlberg, P. E., Luksevics, E. & Lebedev, O. The first tetrapod finds from the Devonian (Upper Famennian) of Latvia. *Phil. Trans. R. Soc. B* **343**, 303–328 (1994).
5. Daeschler, E. B., Shubin, N. H., Thomson, K. S. & Amaral, W. W. A Devonian tetrapod from North America. *Science* **265**, 639–642 (1994).
6. Ahlberg, P. E. *Elginerpeton pancheni* and the earliest tetrapod clade. *Nature* **373**, 420–425 (1995).
7. Lebedev, O. A. & Coates, M. I. The postcranial skeleton of the Devonian tetrapod *Tulerpeton curtum*. *Zool. J. Linn. Soc.* **113**, 307–348 (1995).
8. Coates, M. I. The Devonian tetrapod *Acanthostega gunnari* Jarvik: postcranial anatomy, basal tetrapod interrelationships and patterns of skeletal evolution. *Trans. R. Soc. Edinb. Earth Sci.* **87**, 363–421 (1996).
9. Ahlberg, P. E. Postcranial stem tetrapod remains from the Devonian of Scat Craig, Morayshire, Scotland. *Zool. J. Linn. Soc.* **122**, 99–141 (1998).
10. Ahlberg, P. E. & Clack, J. A. Lower jaws, lower tetrapods—a review based on the Devonian genus *Acanthostega*. *Trans. R. Soc. Edinb. Earth Sci.* **89**, 11–46 (1998).
11. Daeschler, E. B. Early tetrapod jaws from the Late Devonian of Pennsylvania, USA. *J. Paleontol.* **74**, 301–308 (2000).
12. Ahlberg, P. E., Luksevics, E. & Mark-Kurik, E. A near-tetrapod from the Baltic Middle Devonian. *Palaeontology* **43**, 533–548 (2000).
13. Zhu, M., Ahlberg, P. E., Zhao, W. & Jia, L. First Devonian tetrapod from Asia. *Nature* **420**, 760–761 (2002).
14. Clack, J. A. *et al.* A uniquely specialized ear in a very early tetrapod. *Nature* **425**, 66–69 (2003).
15. Clement, G. *et al.* Devonian tetrapod from Western Europe. *Nature* **427**, 412–413 (2004).
16. Lebedev, O. A. A new tetrapod *Jakubsonia livnensis* from the Early Famennian (Devonian) of Russia and palaeoecological remarks on the Late Devonian tetrapod habitats. *Acta Univ. Latviensis* **679**, 79–98 (2004).
17. Shubin, N. H., Daeschler, E. B. & Coates, M. I. The early evolution of the tetrapod humerus. *Science* **304**, 90–93 (2004).
18. Ahlberg, P. E., Clack, J. A. & Blom, H. The axial skeleton of the Devonian tetrapod *Ichthyostega*. *Nature* **437**, 137–140 (2005).
19. Ahlberg, P. E., Clack, J. A., Luksevics, E., Blom, H. & Zupins, I. *Ventastega curonica* and the origin of tetrapod morphology. *Nature* **453**, 1199–1204 (2008).
20. Callier, V., Clack, J. A. & Ahlberg, P. E. Contrasting developmental trajectories in the earliest known tetrapod forelimbs. *Science* **324**, 364–367 (2009).
21. Clack, J. A. *Gaining Ground: The Origin and Early Evolution of Tetrapods* (Indiana Univ. Press, 2002).
22. Schultze, H. P. & Arsenault, M. The panderichthyid fish *Elpistostege*: a close relative of tetrapods? *Palaeontology* **28**, 293–309 (1985).
23. Vorobyeva, E. I. & Schultze, H. P. in *Origins of the Higher Groups of Tetrapods* (eds Schultze, H. P. & Trueb, L.) 68–109 (Cornell, 1991).
24. Vorobyeva, E. & Kuznetsov, A. in *Fossil Fishes as Living Animals* (ed. Mark-Kurik, E.) 131–140 (Academy of Sciences of Estonia, 1992).
25. Vorobyeva, E. I. The shoulder girdle of *Panderichthys rhombolepis* (Gross) (Crossopterygii), Upper Devonian, Latvia. *Geobios* **19**, 285–288 (1995).
26. Boisvert, C. A. The pelvic fin and girdle of *Panderichthys* and the origin of tetrapod locomotion. *Nature* **438**, 1145–1147 (2005).
27. Brazeau, M. D. & Ahlberg, P. E. Tetrapod-like middle ear architecture in a Devonian fish. *Nature* **439**, 318–321 (2006).
28. Boisvert, C. A., Mark-Kurik, E. & Ahlberg, P. E. The pectoral fin of *Panderichthys* and the origin of digits. *Nature* **456**, 636–638 (2008).
29. Daeschler, E. B., Shubin, N. H. & Jenkins, F. A. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature* **440**, 757–763 (2006).
30. Shubin, N. H., Daeschler, E. B. & Jenkins, F. A. The pectoral fin of *Tiktaalik rosae* and the origin of the tetrapod limb. *Nature* **440**, 764–771 (2006).
31. Downs, J. P., Daeschler, E. B., Jenkins, F. A. & Shubin, N. H. The cranial endoskeleton of *Tiktaalik roseae*. *Nature* **455**, 925–929 (2008).
32. Ahlberg, P. E. & Johanson, Z. Osteolepiforms and the ancestry of tetrapods. *Nature* **395**, 792–794 (1998).
33. Gradstein, F., Ogg, J. & Smith, A. *A Geologic Time Scale 2004*. (Cambridge Univ. Press, 2004).
34. Warren, J. W. & Wakefield, N. A. Trackways of tetrapod vertebrates from the Upper Devonian of Victoria, Australia. *Nature* **238**, 469–470 (1972).
35. Warren, A. A., Jupp, R. & Bolton, B. Earliest tetrapod trackway. *Alcheringa* **10**, 183–186 (1986).
36. Stössel, I. The discovery of a new Devonian tetrapod trackway in SW Ireland. *J. Geol. Soc.* **152**, 407–413 (1995).
37. Clack, J. A. Devonian tetrapod trackways and trackmakers; a review of the fossils and footprints. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **130**, 227–250 (1997).
38. Gourmanis, C., Webb, J. A. & Warren, A. A. Fluviodeltaic sedimentology and ichnology of part of the Silurian Grampians Group, western Victoria. *Aust. J. Earth Sci.* **50**, 811–825 (2003).
39. Williams, E. A., Sergeev, S. A., Stössel, I. & Ford, M. An Eifelian U-Pb zircon date for the Enagh Tuff Bed from the Old Red Sandstone of the Munster Basin in NW Iveragh, SW Ireland. *J. Geol. Soc.* **154**, 189–193 (1997).
40. Turnau, E. & Racki, G. Givetian palynostratigraphy and palynofacies: new data from the Bodzentyn Syncline (Holy Cross Mountains, central Poland). *Rev. Palaeobot. Palynol.* **106**, 237–271 (1999).

41. Szulczewski, M. Depositional evolution of the Holy Cross Mts. (Poland) in the Devonian and Carboniferous—a review. *Geol. Q.* **39,** 471–488 (1995).
42. Narkiewicz, K. & Narkiewicz, M. Mid Devonian carbonate platform development in the Holy Cross Mts. area (central Poland): new constraints from the conodont *Bipennatus* fauna. *N. J. Geol. Paläontol.* (in the press).
43. Pajchlowa, M. The Devonian in the Grzegorzowice-Skały Profile. *Biuletyn Instytutu Geologicznego* [in Polish] **122,** 145–254 (1957).
44. Adamczak, F. Middle Devonian Podocopida (Ostracoda) from Poland; their morphology, systematics and occurrence. *Senckenbergiana lethaea* **57,** 265–467 (1976).
45. Malec, J. & Turnau, E. Middle Devonian conodont, ostracod and miospore stratigraphy of the Grzegorzowice–Skały section, Holy Cross Mountains, Poland. *Bull. Pol. Acad. Sci. Earth Sci.* **45,** 67–86 (1997).
46. Zhu, M. *et al.* The oldest articulated osteichthyan reveals mosaic gnathostome characters. *Nature* **458,** 469–474 (2009).
47. Markey, M. J. & Marshall, C. R. Terrestrial-style feeding in a very early aquatic tetrapod is supported by evidence from experimental analysis of suture morphology. *Proc. Natl Acad. Sci. USA* **104,** 7134–7138 (2007).
48. Kurss, V. in *Fossil Fishes as Living Animals* (ed. Mark-Kurik, E.) 251–260 (Academy of Sciences of Estonia, 1992).
49. Trewin, N. H. Palaeoecology and sedimentology of the Achanarras fish bed of the Middle Old Red Sandstone, Scotland. *Trans. R. Soc. Edinb. Earth Sci.* **77,** 21–46 (1986).
50. Scotese, C. R. *PALEOMAP* ⟨http://www.scotese.com⟩ (2002).

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

Acknowledgements We thank G. Pieńkowski, Z. Złonkiewicz, S. Salwa (Polish Geological Institute, Warsaw and Kielce) and J. Dzik (Polish Academy of Sciences and Warsaw University) for their help in fieldwork. This research was supported, in part, by funding from the Faculty of Biology, Warsaw University and Polish Geological Institute. We thank J. O. Ebbestad (Museum of Evolution, Uppsala University) for help with preparing plaster casts of the trackways and P. Ślipiński (Museum and Institute of Zoology, PAN) for help with scanning. I. Jackson (Uppsala University) corrected the English of the Supplementary Information text. P.E.A. acknowledges the support of the Swedish Research Council.

Author Contributions G.N., P.S., K.N. and M.N. discovered and collected the Zachełmie footprints, wrote the Supplementary Information text, and provided all photographs and all geological information about the locality. G.N. invited P.E.A. to participate in the study. P.E.A. identified the footprints as tetrapod and wrote the main text in consultation with G.N.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.E.A. (per.ahlberg@ebc.uu.se).

Preventing the return of fear in humans using reconsolidation update mechanisms

Daniela Schiller[1,2], Marie-H. Monfils[1,3], Candace M. Raio[2], David C. Johnson[2], Joseph E. LeDoux[1] & Elizabeth A. Phelps[1,2]

Recent research on changing fears has examined targeting reconsolidation. During reconsolidation, stored information is rendered labile after being retrieved. Pharmacological manipulations at this stage result in an inability to retrieve the memories at later times, suggesting that they are erased or persistently inhibited. Unfortunately, the use of these pharmacological manipulations in humans can be problematic. Here we introduce a non-invasive technique to target the reconsolidation of fear memories in humans. We provide evidence that old fear memories can be updated with non-fearful information provided during the reconsolidation window. As a consequence, fear responses are no longer expressed, an effect that lasted at least a year and was selective only to reactivated memories without affecting others. These findings demonstrate the adaptive role of reconsolidation as a window of opportunity to rewrite emotional memories, and suggest a non-invasive technique that can be used safely in humans to prevent the return of fear.

Learning about potential dangers in the environment is critical for adaptive function, but at times fear learning can be maladaptive, resulting in excessive fear and anxiety. Research on changing fears has highlighted several techniques, most of which rely on the inhibition of the learned fear response. An inherent problem with these inhibition techniques is that the fear may return, for example with stress[1]. Recent research on changing fears targeting the reconsolidation process overcomes this challenge to some extent. During reconsolidation, stored information is rendered labile after being retrieved, and pharmacological manipulations at this stage result in an inability to retrieve the memories at later times, suggesting that they are either erased or persistently inhibited[2–6]. Although these pharmacological manipulations are potentially useful for changing learned fears, their use in humans can be problematic. Here we show that invasive techniques are not necessary to alter fear by targeting reconsolidation. This is based on the premise that reconsolidation is an adaptive update mechanism by which new information is incorporated into old memories[3,7,8]. By introducing new information during the reconsolidation period, it may be possible to permanently change the fear memory. In the present study, we provide evidence in humans that old fear memories can be updated with non-fearful information provided during the reconsolidation window. As a consequence, fear responses are no longer expressed. Furthermore, this effect is specific to the targeted fear memory, and not others, and persists for at least a year. These findings demonstrate the adaptive role of reconsolidation as a window of opportunity to rewrite emotional memories, and suggest a non-invasive technique that can be used safely and flexibly in humans to prevent the return of fear.

Pharmacological blockade of reconsolidation

In contrast to the traditional view of memory formation as a one-time process of consolidation[9,10], the reconsolidation hypothesis suggests that memories are consolidated each time they are retrieved[2–6]. Evidence for reconsolidation of emotional memories comes from studies using pharmacological perturbation after retrieval[11–13]. The retrieval-induced plasticity allows the transition from a labile to a stable state after which memories are no longer prone to interference[14]. Why would such a recurrent window of vulnerability exist for old memories? From an evolutionary perspective, reconsolidation may serve as an adaptive update mechanism allowing for new information, available at the time of retrieval, to be integrated into the initial memory representation[3,7,8]. This view captures the fluidity of memory and suggests a dynamic process through which memories are formed, updated and maintained.

Using Pavlovian fear conditioning as a model paradigm, research in non-human animals has detailed the molecular processes involved in emotional memory reconsolidation by pharmacologically blocking various stages of this process, after which the memory was no longer expressed. Most of these studies use protein synthesis inhibitors, or other pharmacological agents, that are not safe for use in humans[3,4,6,11–14]. Because the ability to impair emotional memories has important implications for the treatment for anxiety disorders linked to traumatic memories, such as post-traumatic stress disorder (PTSD), identifying techniques to target reconsolidation that can be used flexibly and safely in humans is critical. One possibility is to capitalize on reconsolidation as an update mechanism. If an old fear memory could be restored while incorporating neutral or more positive information provided at the time of retrieval, it may be possible to permanently modify the fearful properties of this memory.

Although this approach captures the very essence of reconsolidation, it has been surprisingly neglected in emotion research in humans and other animals. Until now, there is only one demonstration of this approach in non-human animals using fear conditioning[8], and efforts to alter fear memories by introducing non-fearful information during initial consolidation have had mixed results[15–17]. In humans, studies of motor and declarative memory suggest new information presented during the reconsolidation window may interfere with the older memories by either impairing the memory[18] or modifying it to incorporate the new information[7,19]. However, there is robust evidence that motor, declarative and emotional memories rely on distinct memory systems in the brain[20], and the reconsolidation process and effect of new information presented during the reconsolidation window may differ depending on the type of memory being updated.

[1]Center for Neural Science, [2]Psychology Department, New York University, New York, New York 10003, USA. [3]Psychology Department, University of Texas, Austin, Texas 78712, USA.

Interference of reconsolidation using extinction

In the present study, we sought to capitalize on reconsolidation as an update mechanism and attempted to alter emotional memories with new information. We propose that updating a fear memory with non-fearful information, provided through extinction training, would rewrite the original fear response and prevent the return of fear. A recent study in rats[8] provides strong evidence in support of this hypothesis. In brief, 24 h after fear conditioning, rats were reminded of the conditioned stimulus using a single retrieval trial, and subsequently underwent extinction training. The extinction phase was conducted either within or outside the reconsolidation window, which lasts about 6 h[11,18]. It was found that fear responses returned only in rats that underwent extinction after reconsolidation was completed. In contrast, rats that had extinction training during the reconsolidation window did not show recovery of fear.

To test this hypothesis in humans, we designed two experiments examining whether extinction training conducted during the reconsolidation window would block the return of extinguished fear. In the first study, three groups of subjects underwent fear conditioning using a discrimination paradigm with partial reinforcement (Fig. 1a). Two coloured squares were used. One square (conditioned stimulus+, hereafter termed CS+) was paired with a mild shock to the wrist (unconditioned stimulus) on 38% of the trials, whereas the other square was never paired with shock (CS−). A day later, all three groups underwent extinction training in which the two conditioned stimuli were repeatedly presented without the unconditioned stimulus. In two groups the fear memory was reactivated before extinction using a single presentation of the CS+. One group ($n = 20$) received the reminder trial 10 min before extinction (within the reconsolidation window), whereas the second group ($n = 23$) was reminded 6 h before extinction (outside the reconsolidation window[11,18]). The third group ($n = 22$) was not reminded of the fear memory before extinction training. Twenty-four hours later, all three groups were presented again with the conditioned stimuli without the unconditioned stimulus (re-extinction) to assess spontaneous fear recovery. The measure of fear was the skin conductance response (SCR). At each stage, the differential fear response was calculated by subtracting responses to the CS− from responses to the CS+.

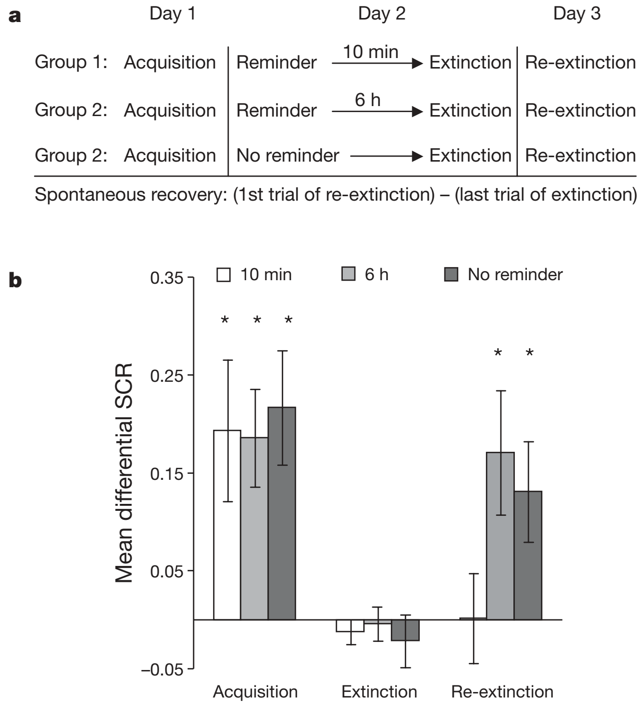


Figure 1 | Extinction during reconsolidation prevents spontaneous recovery of extinguished fear. a, Experimental design and timeline. **b**, Mean differential SCRs (CS+ minus CS−) during acquisition (late phase), extinction (last trial) and re-extinction (first trial) for each experimental group (10-min reminder, 6-h reminder and no reminder). The three groups showed equivalent fear acquisition and extinction. Spontaneous recovery (first trial of re-extinction versus the last trial of extinction) was found in the group that had not been reminded or that was reminded 6 h before extinction. In contrast, there was no spontaneous recovery in the group reminded 10 min before extinction. $*P < 0.05$ (between acquisition and extinction, or between extinction and re-extinction within group). Error bars represent standard errors.

The results of the spontaneous recovery experiment are presented in Fig. 1b (see also Supplementary Fig. 1). Subjects that showed successful levels of fear acquisition and extinction were included in the analysis. We verified that these levels were equivalent between the groups using two-way analysis of variance (ANOVA) with main effects of group (10 min, 6 h and no reminder) and time (early and late phase). For both acquisition and extinction there was a significant main effect of time ($F_{1,62} = 9.92$, $P < 0.05$; $F_{1,62} = 19.59$, $P < 0.01$, respectively) but no effect of group or interaction. Follow-up t-tests confirmed that subjects had significantly stronger responses to CS+ than to CS− during acquisition (late phase; 10-min group: $t = 2.68$, $P < 0.05$; 6-h group: $t = 3.72$, $P < 0.05$; no-reminder group: $t = 3.72$, $P < 0.05$), but by the last trial of extinction there was no difference (10-min group: $t = -0.94$; 6-h group: $t = -0.23$; no-reminder group: $t = -0.79$; all not significant).

The decrease in fear responses from acquisition (late phase) to extinction (last trial) for each group was assessed using a two-way ANOVA with main effects of group (10 min, 6 h and no reminder) and time (acquisition, extinction). This showed a significant main effect of time ($F_{1,62} = 29.9$, $P < 0.01$), but no effect of group or interaction. Follow-up t-tests confirmed the reduction of fear in all three groups (10-min group: $t = 2.70$, $P < 0.05$; 6-h group: $t = 4.06$, $P < 0.05$; no-reminder group: $t = 4.07$, $P < 0.05$), and there was no difference in the level of fear reduction between the groups ($P > 0.5$ for all three comparisons).

Spontaneous recovery was assessed using a two-way ANOVA with main effects of group (10 min, 6 h and no reminder) and time (early and late phase of re-extinction, defined by the mean first four responses versus the subsequent four, respectively) showing a significant main effect of time ($F_{1,62} = 6.26$, $P < 0.05$), and a group × time interaction ($F_{2,62} = 4.63$, $P < 0.05$). Follow-up t-tests compared the differential responses between the last trial of extinction and the first trial of re-extinction. Spontaneous recovery was found in subjects who did not receive a reactivation trial before extinction ($t = 2.69$, $P < 0.05$), or who underwent extinction 6 h after fear reactivation ($t = 2.66$, $P < 0.05$). In contrast, subjects that had extinction 10 min after reactivation showed no spontaneous recovery ($t = 0.28$, not significant). These results indicate that the spontaneous recovery of fear after extinction can be prevented if extinction training is conducted during the time window in which the fear memory is proposed to be undergoing reconsolidation.

Persistence of reconsolidation blockade

In this initial study, we used a 24 h interval to test for long-term memory, which, for practical reasons, is the standard in human fear recovery experiments[16,17,21–23]. However, if the fear memory is persistently altered, as would be predicted if we are affecting reconsolidation of the fear memory, we would expect this effect to last for much longer time intervals. In an attempt to examine whether the observed blockade of fear memory persists, we invited the participants for a follow-up test after approximately 1 year (10–14 months). Nineteen of the 65 original participants were located and included in the follow-up study (10-min group, $n = 8$; 6-h group, $n = 4$; no-reminder group, $n = 7$). We collapsed subjects from the two groups previously showing spontaneous recovery (that is, 6 h and no reminder) into one group. As mentioned earlier, after the spontaneous recovery test, subjects were re-extinguished using ten non-reinforced presentations of the stimuli ensuring that all subjects showed no evidence of conditioned fear at the conclusion of the initial experiment. This re-extinction allowed us to conduct a second test of fear recovery a year later. For this second recovery test, we used a more potent recovery assay, namely reinstatement, in which subjects were exposed to four unsignalled shocks, followed by non-reinforced presentations of the conditioned stimuli. The index of fear recovery (Fig. 2 and Supplementary Fig. 2) was the difference in the conditioned fear response at the end of re-extinction after the initial spontaneous recovery test and the conditioned fear response immediately after reinstatement 1 year later. The conditioned fear response at the end of re-extinction and post-reinstatement was calculated using a differential SCR score (CS+ minus CS−). A two-way ANOVA with main factors of group (10 min, 6 h/no-reminder) and stage (re-extinction, post-reinstatement) showed a significant main effect of group ($F_{1,17} = 5.89$, $P < 0.05$). The group × stage interaction was marginally significant ($F_{1,17} = 2.78$, $P < 0.07$, one-tail). Follow-up one-tail *t*-test comparisons showed that reinstatement was significant in the 6-h/no-reminder group ($t = 2.12$, $P < 0.03$), but not the 10-min group ($t = 0.22$, not significant). Moreover, the reinstatement index was significantly larger in the 6-h/no-reminder group than the 10-min group ($t = 1.75$, $P < 0.05$). Lastly, a comparison of post-reinstatement conditioned fear between the groups showed a significant difference ($t = 2.18$, $P < 0.03$).

These results indicate that reactivation of a fear memory renders it labile and extinction training during this lability period leads to a long lasting blockade of recovery of fear. In contrast, recovery of fear a year later was observed after regular extinction training. Fear recovery was also observed when extinction training was conducted with a sufficient temporal gap after reactivation, presumably allowing for reconsolidation to be complete.

Specificity of reconsolidation blockade

If interfering with reconsolidation using extinction is to be clinically useful, it is also important to determine whether it is specific. In real-life situations, a traumatic event can be associated with several cues, and each could potentially trigger the recollection of the event and elicit fear reactions. To assess the specificity of this fear blockade technique, we examined whether interfering with the reconsolidation of one fear predictive cue would affect the fate of another, associated cue.

In a second experiment, more than one stimulus was associated with the same aversive outcome (Fig. 3a). Specifically, using a within-subject design, subjects underwent fear conditioning using three coloured squares. Two squares (CSa+ and CSb+) were paired with the shock on 38% of the trials. The third square (CS−) was never paired with the shock. A day later, subjects received a single presentation of CSa+ and the CS−, but not CSb+. Ten minutes after the reminder trial, extinction training was conducted (within the reconsolidation window) using repeated presentations of all conditioned stimuli without the aversive outcome. Reinstatement of the fear memory was conducted 24 h later, when subjects returned to the experiment room and received four unsignalled presentations of the shock. Ten minutes later, the conditioned stimuli were presented without the aversive outcome (re-extinction).

The results of the experiment are presented in Fig. 3b (see also Supplementary Fig. 3). Subjects ($n = 18$) that showed successful fear acquisition and extinction were included. We verified that these levels were equivalent between the two conditioned stimuli (CSa+ and CSb+) using two-way ANOVAs with main effects of stimulus (CSa+, CSb+ and CS−) and time (early and late phase, defined by the mean response during the first and second half of each phase, respectively). In acquisition, there was a significant main effect of stimulus ($F_{2,51} = 3.51$, $P < 0.05$) and a stimulus × time interaction ($F_{2,51} = 3.27$, $P < 0.05$). In extinction, there was a significant main effect of time ($F_{1,51} = 48.74$, $P < 0.01$). Follow-up *t*-tests were used to further assess acquisition and extinction of fear. We compared the mean SCR to CSa+ or CSb+ with the CS− during the second half of the acquisition session. Subjects showed significantly stronger responses to CSa+ than to CS− ($t = 6.01$, $P < 0.05$), as well as to CSb+ compared to CS− ($t = 6.68$, $P < 0.05$). Moreover, the level of acquisition to CSa+ and CSb+ was equivalent ($t = 0.76$, not significant). To

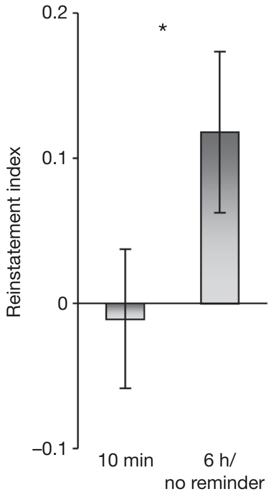


Figure 2 | Blockade of the return of fear persists one year later. The reinstatement index is the difference in the conditioned fear response (CS+ minus CS−) at the end of re-extinction after the initial spontaneous recovery test and the conditioned fear response immediately after reinstatement a year later. The magnitude of the reinstatement was significantly higher in the 6-h/no-reminder group than in the 10-min group, which showed no reinstatement. *$P < 0.05$; error bars represent standard errors.

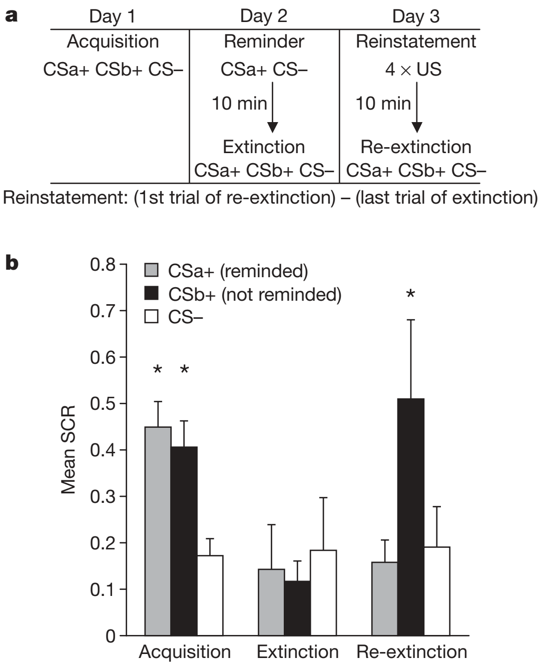


Figure 3 | Blockade of the return of fear is specific to reactivated memories. **a**, Experimental design and timeline. US, unconditioned stimulus. **b**, Mean SCRs (CSa+, CSb+ and CS−) during acquisition (late phase), extinction (last trial) and re-extinction (first trial). Subjects had equivalent levels of acquisition and extinction of conditioned fear to the two conditioned stimuli. The index of fear recovery was the first trial of re-extinction (after reinstatement) minus the last trial of extinction (before reinstatement). Fear reinstatement was found only to CSb+ (not reminded before extinction training), but not to CSa+ (reminded 10 min before extinction training). *$P < 0.05$ (between acquisition and extinction, or extinction and re-extinction for each stimulus). Error bars represent standard errors.

assess fear extinction, we compared the mean SCR to CSa+ or CSb+ with the CS− during the last trial of extinction. There were no significant differences in responses to CSa+ compared to CS− ($t = -0.26$, not significant), or to CSb+ compared to CS− ($t = -0.56$, not significant), and responses to CSa+ and CSb+ were equally extinguished ($t = 0.23$, not significant). Moreover, subjects had successful reduction of fear, as assessed by comparing the SCR during the second half of acquisition with the last trial of extinction, to both CSa+ ($t = 2.62$, $P < 0.05$) and CSb+ ($t = 4.08$, $P < 0.05$) but not to the CS− ($t = -0.09$, not significant), which was low to begin with.

To assess the recovery of fear, we used a two-way ANOVA with main effects of stimulus (CSa+, CSb+ and CS−) and time (early and late phase of re-extinction, defined by the mean first four responses versus the last four, respectively), which revealed a stimulus × time interaction ($F_{2,51} = 5.14$, $P < 0.01$). Using follow-up t-tests, we compared the SCR during the last trial of extinction (before reinstatement) with the first trial of re-extinction (after reinstatement). Subjects showed reinstated fear responses only to CSb+, which is the stimulus that was not reminded before extinction ($t = 2.16$, $P < 0.05$). In contrast, fear responses to CSa+, which was reminded 10 min before extinction training, did not recover ($t = 0.22$, not significant). As expected, there were also no fear responses to the CS− ($t = 0.16$, not significant). Thus, extinction during reconsolidation affected only the reactivated memory and no other trace associated with the original event.

Discussion

The present findings suggest a new technique to target specific fear memories and prevent the return of fear after extinction training. Using two recovery assays, we demonstrated that extinction conducted during the reconsolidation window of an old fear memory prevented the spontaneous recovery or the reinstatement of fear responses, an effect that was maintained a year later. Moreover, this manipulation selectively affected only the reactivated conditioned stimulus while leaving fear memory to the other non-reactivated conditioned stimulus intact.

It has been suggested that the adaptive function of reconsolidation is to allow old memories to be updated each time they are retrieved[3,7,8]. In other words, our memory reflects our last retrieval of it rather than an exact account of the original event. This notion has received support from interference paradigms targeting motor and declarative memories[7,18,19]. These studies demonstrate that new information provided during reconsolidation could affect old memories by modifying or interfering with them, but in contrast to the present study, they do not provide evidence for memory blockade. This difference in the effect of new information presented during reconsolidation on the subsequent qualities of different types of memory may be due to the diverse nature of the underlying memory systems. For instance, unlike the distributed cortical representation of declarative memories[20], conditioned fear has a more discrete neural representation localized in the amygdala[24]. Indeed, in the lateral amygdala, pharmacological blockade of the molecular cascade engaged by retrieval prevents the reconsolidation of fear memories in rats[4]. This raises the possibility that our behavioural manipulation, namely, extinction training during reconsolidation, targeted the same molecular mechanism.

Although the current behavioural study does not provide direct evidence that a process of reconsolidation mediates the effects of extinction training, support for this hypothesis comes from recent findings in rats[8]. After fear consolidation, a single isolated retrieval trial before extinction prevented the recovery of fear in rats. Interestingly, plasticity in the lateral amygdala induced by the conditioned stimulus retrieval was impaired by the presentation of a conditioned stimulus 1 h later, indicating possible interference with the reconsolidation process, similar to the interference caused to reconsolidation by pharmacological blockade in rats[4]. Together, these findings reveal cross-species similarities, which may reflect an evolutionarily preserved adaptive mechanism whereby the neural representation of fear memory can be significantly altered through time-dependent molecular mechanisms triggered by exposure to fear-eliciting stimuli.

The current results also suggest that timing may have a more important role in the control of fear than previously appreciated. Standard extinction training, without previous memory reactivation, also triggers the fear memory. Given this, one might expect mere extinction training to have similar effects. That is, the first trial of extinction might serve as the reminder cue triggering the reconsolidation cascade, which is immediately followed by extinction. However, there is abundant evidence that during standard extinction training the non-reinforced presentations of the fear-eliciting cue induce new inhibitory learning, which competes for expression with the initial fear learning, resulting in the recovery of fear responses in some circumstances[16,17,21–23,25,26]. Our findings indicate that the timing of extinction relative to the reactivation of the memory can capitalize on reconsolidation mechanisms. Two factors may be important determinants in this process: the timing of extinction training relative to retrieval, and/or the chunking of the conditioned stimulus presentations during extinction relative to reactivation (that is, the fact that they are massed relative to the single retrieval trial during the reconsolidation phase). Further studies are required to disentangle these possibilities.

In conclusion, the present study showed that updating fear memories with non-fearful information provided through extinction training led to the blockade of previously learned fear responses and a lasting change in the original fear memory. These results have significant implications for the treatment of anxiety disorders. Current forms of therapy rely heavily on extinction[27,28], but the fact that extinguished fear could recover under certain conditions dampens the resilience of anxiety patients after treatment. The discovery that certain pharmacological manipulation can potentially erase memories through effects on reconsolidation has been encouraging; however, most compounds showing such effects in various species are toxic to humans. Recently, there has been promising evidence using compounds that are testable on humans, namely β-adrenergic receptor blockers[29], which also show effects in trauma patients[30], but these effects are not observed in every case[31]. The present study proposes that such invasive techniques are not necessary. Using a more natural intervention that captures the adaptive purpose of reconsolidation allows a safe and easily implemented way to prevent the return of fear.

METHODS SUMMARY

Two experiments were designed to examine whether extinction training conducted during the reconsolidation window would block the return of extinguished fear. The measure of fear was the SCR. In the first study, three groups of subjects underwent a discrimination fear conditioning paradigm with partial reinforcement. Two coloured squares (CS+ and CS−) were used. The CS+ was paired with a mild shock to the wrist (unconditioned stimulus) on about one-third of the trials, and the CS− was never paired with the shock. A day later, all three groups underwent extinction training (repeated conditioned stimulus presentations without the unconditioned stimulus). In two groups the fear memory was reactivated before extinction using a single presentation of the CS+. One group received the reminder trial 10 min before extinction (within the reconsolidation window), whereas the second group was reminded 6 h before extinction (outside the reconsolidation window). The third group was not reminded of the fear memory before extinction training. To assess spontaneous fear recovery, a day later all three groups were presented with the conditioned stimuli without the unconditioned stimulus (re-extinction). About a year later, the return of fear was assessed again using a different recovery assay (reinstatement).

The second experiment used a within-subject design where subjects underwent fear conditioning using three coloured squares. Two squares (CSa+ and CSb+) were paired with the shock on about one-third of the trials. The third square (CS−) was never paired with the shock. A day later, subjects received a single presentation of CSa+ and the CS−, but not CSb+. Ten minutes after the reminder trial, extinction training was conducted (within the reconsolidation window) using repeated presentations of all conditioned stimuli without the unconditioned stimulus. Reinstatement of the fear memory was conducted

24 h later, when subjects returned to the experiment room and received four unsignalled presentations of the shock. Ten minutes later the conditioned stimuli were presented without the aversive outcome (re-extinction).

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

Received 24 August; accepted 5 November 2009.
Published online 9 December 2009.

1. Miracle, A. D., Brace, M. F., Huyck, K. D., Singler, S. A. & Wellman, C. L. Chronic stress impairs recall of extinction of conditioned fear. *Neurobiol. Learn. Mem.* **85,** 213–218 (2006).
2. Misanin, J. R., Miller, R. R. & Lewis, D. J. Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science* **160,** 554–555 (1968).
3. Alberini, C. M. Mechanisms of memory stabilization: are consolidation and reconsolidation similar or distinct processes? *Trends Neurosci.* **28,** 51–56 (2005).
4. Nader, K., Schafe, G. E. & LeDoux, J. E. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406,** 722–726 (2000).
5. Dudai, Y. Reconsolidation: the advantage of being refocused. *Curr. Opin. Neurobiol.* **16,** 174–178 (2006).
6. Sara, S. J. & Hars, B. In memory of consolidation. *Learn. Mem.* **13,** 515–521 (2006).
7. Hupbach, A., Gomez, L., Hardt, O. & Nadel, R. Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. *Learn. Mem.* **14,** 47–53 (2007).
8. Monfils, M.-H., Cowansage, K. K., Klann, E. & LeDoux, J. E. Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science* **324,** 951–955 (2009).
9. Squire, L. R. & Davis, H. P. The pharmacology of memory: a neurobiological perspective. *Annu. Rev. Pharmacol. Toxicol.* **21,** 323–356 (1981).
10. McGaugh, J. L. Memory—a century of consolidation. *Science* **287,** 248–251 (2000).
11. Duvarci, S. & Nader, K. Characterization of fear memory reconsolidation. *J. Neurosci.* **24,** 9269–9275 (2004).
12. Alberini, C. M., Milekic, M. H. & Tronel, S. Memory: mechanisms of memory stabilization and de-stabilization. *Cell. Mol. Life Sci.* **63,** 999–1008 (2006).
13. Lee, J. L., Milton, A. L. & Everitt, B. J. Reconsolidation and extinction of conditioned fear: inhibition and potentiation. *J. Neurosci.* **26,** 10051–10056 (2006).
14. Doyère, V., Debiec, J., Monfils, M. H., Schafe, G. E. & LeDoux, J. E. Synapse-specific reconsolidation of distinct fear memories in the lateral amygdala. *Nature Neurosci.* **10,** 414–416 (2007).
15. Myers, K. M., Ressler, K. J. & Davis, M. Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learn. Mem.* **13,** 216–223 (2006).
16. Alvarez, R. P., Johnson, L. & Grillon, C. Contextual-specificity of short-delay extinction in humans: renewal of fear-potentiated startle in a virtual environment. *Learn. Mem.* **14,** 247–253 (2007).
17. Schiller, D. *et al.* Evidence for recovery of fear following immediate extinction in rats and humans. *Learn. Mem.* **15,** 394–402 (2008).
18. Walker, M. P., Brakefield, T., Hobson, J. A. & Stickgold, R. Dissociable stages of human memory consolidation and reconsolidation. *Nature* **425,** 616–620 (2003).
19. Forcato, C. *et al.* Reconsolidation of declarative memory in humans. *Learn. Mem.* **14,** 295–303 (2007).
20. Squire, L. H. & Knowlton, B. J. in *The New Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 765–780 (MIT Press, 2000).
21. Phelps, E. A., Delgado, M. R., Nearing, K. I. & LeDoux, J. E. Extinction learning in humans: role of the amygdala and vmPFC. *Neuron* **43,** 897–905 (2004).
22. Kalisch, R. *et al.* Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *J. Neurosci.* **26,** 9503–9511 (2006).
23. Milad, M. R. *et al.* Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biol. Psychiatry* **62,** 446–454 (2007).
24. LeDoux, J. E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23,** 155–184 (2000).
25. Bouton, M. E. Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biol. Psychiatry* **52,** 976–986 (2002).
26. Quirk, G. J. & Mueller, D. Neural mechanisms of extinction learning and retrieval. *Neuropsychopharmacology* **33,** 56–72 (2008).
27. Foa, E. B., Franklin, M. E. & Moser, J. Context in the clinic: how well do cognitive-behavioral therapies and medications work in combination. *Biol. Psychiatry* **52,** 987–997 (2002).
28. Rauch, S. L., Shin, L. M. & Phelps, E. A. Neurocircuitry models of posttraumatic stress disorder and extinction: human neuroimaging research—past, present and future. *Biol. Psychiatry* **60,** 376–382 (2006).
29. Kindt, M., Soeter, M. & Vervliet, B. Beyond extinction: erasing human fear responses and preventing the return of fear. *Nature Neurosci.* **12,** 256–258 (2009).
30. Brunet, A. *et al.* Effect of post-retrieval propranolol on psychophysiologic responding during subsequent script-driven traumatic imagery in post-traumatic stress disorder. *J. Psychiatr. Res.* **42,** 503–506 (2008).
31. Tollenaar, M. S., Elzinga, B. M., Spinhoven, P. & Everaerd, W. Psychophysiological responding to emotional memories in healthy young men after cortisol and propranolol administration. *Psychopharmacology (Berl.)* **203,** 793–803 (2009).

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

Acknowledgement We thank K. Doelling for assistance with data collection and discussions on the revised version of the manuscript. We also thank Y. Niv and M. Milad for advice on the experimental protocols. This study was funded by the James S. McDonnell Foundation and National Institutes of Health (NIH) grant R21 MH072279 (E.A.P.), NIH grants R37 MH038774, P50 MH058911, RO1 MH046516 and K05 MH067048 (J.E.L.), Postdoctoral fellowships NSERC, CIHR and AHFMR (M.-H.M.), and a Fulbright award (D.S.).

Author Contributions D.S. designed the experiments, collected and analysed data, interpreted the data and wrote the first draft of the manuscript; C.M.R. and D.C.J. collected the data and contributed to experimental design, analysis, interpretation and the final version of the manuscript; M.-H.M., J.E.L. and E.A.P. contributed to experimental design, data interpretation, and the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.A.P. (liz.phelps@nyu.edu).

METHODS

Experiment 1. The study consisted of three consecutive stages conducted 24 h apart: day 1, acquisition; day 2, reactivation and extinction; and day 3, re-extinction (Fig. 1a). During acquisition, three randomly assigned groups of subjects underwent a Pavlovian discrimination fear-conditioning paradigm with partial reinforcement. The conditioned stimuli (CS+, CS−) were yellow and blue squares (4 s) and the unconditioned stimulus was a mild shock to the wrist (200 ms) co-terminating with the CS+. The inter-trial-interval (ITI) was 10–12 s. The CS+ was paired with the shock on a 38% partial reinforcement schedule and the CS− was never paired with shock (10 CS+, 10 CS−, 6 CS+ with shock). Subjects were instructed to pay attention to the computer screen and to try to figure out the relationship between the stimuli appearing on the screen and the shocks. A day later, all three groups underwent extinction training in which the CS+ and CS− were repeatedly presented without the unconditioned stimulus. In two groups, the fear memory was reactivated before extinction. During reactivation, the CS+ was presented once (unreinforced), followed by a 10-min break. One group ($n = 20$) underwent extinction after the 10-min break (10 CS+, 11 CS−; within the reconsolidation window). The second group ($n = 23$) underwent extinction 6 h after the reactivation (10 CS+, 11 CS−; outside of the reconsolidation window). In the third group ($n = 22$), the fear memory was not reactivated. After the break, extinction immediately followed for half of the subjects in this group, or was conducted 6 h later for the other half (11 CS+, 11 CS−). During the break, all participants watched a pre-selected television show episode. Day 3 consisted of re-extinction in which participants were presented with non-reinforced presentations of the stimuli (10 CS+, 11 CS−). During all sessions (acquisition, reminder, extinction and re-extinction), with the exception of the breaks, the participants were attached to the SCR and shock electrodes, and the shock stimulator was set to the 'on' position.

To examine how long the blockade of memory persists, we invited the participants of the experiment to come back to the laboratory after about a year (10–14 months). Twenty-three participants were located (10-min group, $n = 10$; 6-h group, $n = 5$; no-reminder group, $n = 8$). As mentioned earlier, after the spontaneous recovery test, subjects were re-extinguished using ten non-reinforced presentations of the stimuli, which allowed us to reassess their recovery of fear. We used a more potent recovery assay, namely, reinstatement, in which subjects were exposed to four unsignalled shocks, followed by non-reinforced presentations of the same conditioned stimuli that were used in the spontaneous recovery experiment (10 CS+, 10 CS−, using two randomized orders counterbalanced across subjects). The index of fear recovery was the difference in the conditioned fear response at the end of re-extinction after the initial spontaneous recovery test and the conditioned fear response immediately after reinstatement a year later. Specifically, a differential SCR score (CS+ minus CS−) was calculated for the end of re-extinction (mean of last two trials) and post-reinstatement (mean of first four trials). We collapsed subjects from the two groups previously showing spontaneous recovery (that is, 6 h and no reminder) into one group. Subjects that failed to re-extinguish after the spontaneous recovery test (differential SCR score > 0.2) or showed no measurable responses to the shocks during reinstatement were not included in the analysis (four subjects). The final analysis included 19 subjects (10-min group, $n = 8$; 6-h/no-reminder group, $n = 11$). Throughout the session, the participants were attached to the SCR and shock electrodes, and the shock stimulator was set to the 'on' position.

Experiment 2. The study consisted of three consecutive stages conducted 24 h apart: day 1, acquisition; day 2, reactivation and extinction; and day 3, reinstatement and re-extinction, using a within-subject design (Fig. 2a). During acquisition, subjects underwent fear conditioning using three coloured squares. Two squares (CSa+ and CSb+) were paired with the shock on a 38% partial reinforcement schedule. The third square (CS−) was never paired with the shock (eight non-reinforced presentations of CSa+, CSb+ and CS− each, intermixed with an extra 5 CSa+ and 5 CSb+ presentations that co-terminated with the shock). The stimuli were presented for 4 s each with a 10–12 s variable ITI. Subjects were instructed to pay attention to the computer screen and to try to figure out the relationship between the stimuli appearing on the screen and the shocks. Day 2 consisted of reactivation and extinction. During reactivation, the CSa+ and the CS− were each presented once (unreinforced), in a counterbalanced fashion. Participants were then given a 10-min break in which they watched a pre-selected television show episode. Extinction immediately followed and consisted of non-reinforced presentations of the three stimuli (10 CSa+, 11 CSb+ and 11 CS−). Day 3 consisted of reinstatement and re-extinction. During reinstatement, subjects were administered four unsignalled shocks. After a 10-min break, a re-extinction session began in which participants were presented with non-reinforced presentations of the three stimuli (10 CSa+, 10 CSb+ and 11 CS−). During all sessions (acquisition, reminder, extinction, reinstatement and re-extinction), with the exception of the breaks, the participants were attached to the SCR and shock electrodes, and the shock stimulator was set to the 'on' position.

Psychophysiological stimulation and assessment. Mild shocks were delivered through a stimulating bar electrode attached with a Velcro strap to the right inner wrist. A Grass Medical Instruments stimulator charged by a stabilized current was used. Subjects determined the level of the shock themselves, beginning at a very mild level of shock (10 V) and gradually increasing the level until the shock reached the maximum level that they determined was uncomfortable, but not painful (the maximum level was 60 V). All shocks were given for 200 ms, with a current of 50 pulses per second.

SCR was assessed using two Ag–AgCl electrodes, which were connected to a BioPac Systems skin conductance module. The electrodes were attached to the first and second fingers of the left hand, between the first and second phalanges. SCR waveforms were analysed offline, using AcqKnowledge 3.9 software (BIOPAC Systems Inc.). SCR amplitudes to the conditioned and unconditioned stimuli were the dependent measures of conditioned and unconditioned responses, respectively. The level of SCR response was determined by taking the base-to-peak difference for the first waveform (in microsiemens, μs) in the 0.5–4.5 s window after stimulus onset. The minimal response criterion was 0.02 μs. The raw SCR scores were square-root transformed to normalize distributions. These normalized scores were scaled according to each subject's unconditioned response by dividing each response by the mean square-root-transformed unconditioned stimulus response.

ARTICLES

Ubiquitin-like small archaeal modifier proteins (SAMPs) in *Haloferax volcanii*

Matthew A. Humbard[1]*, Hugo V. Miranda[1]*, Jae-Min Lim[3]*, David J. Krause[1], Jonathan R. Pritz[1], Guangyin Zhou[1], Sixue Chen[2], Lance Wells[3] & Julie A. Maupin-Furlow[1]

Archaea, one of three major evolutionary lineages of life, encode proteasomes highly related to those of eukaryotes. In contrast, archaeal ubiquitin-like proteins are less conserved and not known to function in protein conjugation. This has complicated our understanding of the origins of ubiquitination and its connection to proteasomes. Here we report two small archaeal modifier proteins, SAMP1 and SAMP2, with a β-grasp fold and carboxy-terminal diglycine motif similar to ubiquitin, that form protein conjugates in the archaeon *Haloferax volcanii*. The levels of SAMP-conjugates were altered by nitrogen-limitation and proteasomal gene knockout and spanned various functions including components of the Urm1 pathway. LC-MS/MS-based collision-induced dissociation demonstrated isopeptide bonds between the C-terminal glycine of SAMP2 and the ε-amino group of lysines from a number of protein targets and Lys 58 of SAMP2 itself, revealing poly-SAMP chains. The widespread distribution and diversity of pathways modified by SAMPylation suggest that this type of protein conjugation is central to the archaeal lineage.

In eukaryotic cells, the conjugation of ubiquitin (Ub) and ubiquitin-like (Ubl) proteins to protein targets plays an integral role in a wide variety of processes, including proteasome-mediated proteolysis, heterochromatin remodelling and protein trafficking[1,2]. Elaborate ATP-dependent systems mediate these covalent attachments, including the use of E1 Ub-activating, E2 Ub-conjugating and E3 Ub-protein ligase enzymes[1,2]. Of these, E1 catalyses the ATP-dependent adenylation of the Ub/Ubl C-terminal carboxylate and transfers this activated form of Ub/Ubl to a conserved cysteine on E1. This Ub/Ubl thioester intermediate is transferred to an E2 to form a second thioester linkage. The E2 Ub-conjugating enzyme then transfers the Ub/Ubl to an ε-amino group of a lysine residue either within a target protein or on a growing poly-Ub/Ubl chain[2,3]. Transfer to N^{α}-amino groups has also been observed[4]. Often Ub-transfer is with assistance from an E3 Ub-protein ligase either forming an E3-Ub/Ubl thioester intermediate or with E3 facilitating Ub/Ubl-transfer from E2 directly to the substrate protein.

Although universal in eukaryotes, the presence of Ub-like protein conjugation systems in prokaryotes is less clear. PUP, the first example of a protein covalently attached to target proteins in prokaryotes[5,6], appears restricted to *Actinobacteria* and *Nitrospira* and is distinct from ubiquitination in its use of deamidase and glutamine synthetase-like ligase[6,7] reactions for conjugation and its disordered structure[8,9]. The β-grasp fold of Ub/Ubl proteins, however, are common to a growing superfamily of proteins involved in diverse functions that span all three domains of life[10–12]. Of these β-grasp functions, the enzymology and mechanism of sulphur activation for the biosynthesis of thiamine, tungsten and molybdenum cofactors bears striking resemblance to the activation of Ub/Ubl[13]. Jab1/MPN domain metalloenzyme (JAMM) motifs common to deubiquitinating enzymes used for the recycling of Ub and removal of Ubl modifiers are also conserved in many prokaryotes[14–16]. On the basis of these features, it is unclear (1) whether eukaryotic Ub/Ubl-systems were derived from a combination of various prokaryotic β-grasp fold pathways that function in related yet distinct chemistry or (2) whether prokaryotes figured out how to conjugate Ub/Ubl-proteins to protein targets before the divergence of eukaryotes. Here we demonstrate that two small archaeal modifier proteins (SAMPs) of the β-grasp superfamily are differentially conjugated to protein targets in the archaeon *Haloferax volcanii*, thus providing an evolutionary link in Ub/Ubl-protein conjugation systems.

SAMP1 and SAMP2 form protein conjugates

Small proteins with a β-grasp fold and C-terminal diglycine motif similar to Ub are widespread among *Archaea*[10–12]. Although presumed to activate sulphur for the biosynthesis of cofactors such as thiamine, tungsten and molybdenum, the biological function of these proteins remains unknown. In this study, Ub-like β-grasp proteins were identified in the deduced proteome of *H. volcanii* (Fig. 1). The proteins were fused to an N-terminal Flag tag and synthesized in *H. volcanii* grown under various conditions including complex and minimal media, nitrogen-limitation and salt concentrations ranging from

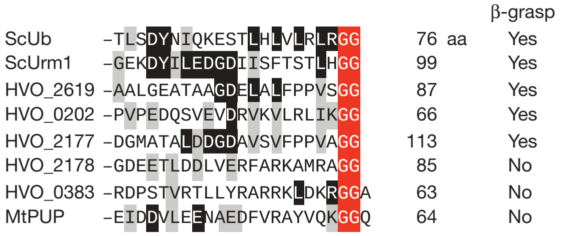

β-grasp
ScUb –TLSDYNIQKESTLHLVLRLRGG 76 aa Yes
ScUrm1 –GEKDYILEDGDIISFTSTLHGG 99 Yes
HVO_2619 –AALGEATAAGDELALFPPVSGG 87 Yes
HVO_0202 –PVPEDQSVEVDRVKVLRLIKGG 66 Yes
HVO_2177 –DGMATALDDGDAVSVFPPVAGG 113 Yes
HVO_2178 –GDEETLDDLVERFARKAMRAGG 85 No
HVO_0383 –RDPSTVRTLLYRARRKLDKRGGA 63 No
MtPUP –EIDDVLEENAEDFVRAYVQKGGQ 64 No

Figure 1 | Multiple amino acid sequence alignment of the C termini of Ub, Urm1 and PUP to select diglycine motif proteins of *H. volcanii*. C-terminal diglycine motifs are shaded in red. Identical and similar amino acids are shaded in black and grey, respectively. Amino acid length of protein (aa) and membership in the Ub/ThiS/MoaD β-grasp superfamily are indicated. HVO, *Haloferax volcanii*; Sc, *Saccharomyces cerevisiae*; Mt, *Mycobacterium tuberculosis*; HVO_2619, SAMP1; HVO_0202, SAMP2.

[1]Department of Microbiology and Cell Science, [2]Department of Biology and Interdisciplinary Center for Biotechnological Research, University of Florida, Gainesville, Florida 32611, USA. [3]Department of Biochemistry and Molecular Biology, Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602, USA.
*These authors contributed equally to this work.

suboptimal to optimal (1.0–2.5 M NaCl). The Flag-tagged proteins were analysed for conjugate formation by anti-Flag immunoblot (anti-Flag) of cell lysate separated by reducing SDS–polyacrylamide gel electrophoresis (PAGE).

Using this approach, two Ubl-proteins, HVO_2619 (SAMP1) and HVO_0202 (SAMP2) that share only 21% identity and 30% similarity in amino acid sequence, were found to form differential protein conjugates that were modulated by growth condition (Fig. 2). Protein conjugates were not detected for the remaining proteins examined (HVO_2177, HVO_2178 and HVO_0383) (Supplementary Fig. 1). Although the number of SAMP-conjugates detected was minimal when cells were grown under standard conditions in complex medium with only two discrete protein bands detected for each SAMP (58 and 14 kDa for SAMP1 and 18 and 16 kDa for SAMP2) (Fig. 2a), a dramatic increase in the number of SAMP-conjugates was observed when cells were transferred to glycerol-alanine minimal medium (Fig. 2b). Systematic supplementation of media with glycerol, alanine and ammonium chloride revealed low nitrogen was the signal for this prominent increase (Fig. 2b). Each of the SAMPs was associated with distinct patterns of protein-conjugates suggesting the presence of a relatively complex regulatory network of SAMPylation that not only senses environmental cues, but also discriminates and differentially conjugates the two SAMP proteins to their protein targets. Interestingly, the predominant SAMP2-conjugates detected migrated in regular intervals of $\sim$ 11–12 kDa by SDS–PAGE, suggesting SAMP2 formed free SAMP2 polymers.

Proteasomes alter SAMP conjugates

H. volcanii mutant strains with markerless deletions in proteasomal genes, including those encoding the subunits of the 20S proteasomal core particle and Rpt-like ATPase subtypes[17], were used to examine the influence of proteasome function on the levels of SAMP-conjugate formation. Site-2-type metalloprotease (S2P) knockout strains were also included in this analysis. Unlike some archaea that synthesize a single core particle of α- and β-type subunit composition and do not encode Rpt-like ATPases, *H. volcanii* synthesizes multiple proteasomal subtypes, including core particles with a β-type subunit that associates with α1 and/or α2 subunits as well as PAN-A and PAN-B proteins that are closely related to eukaryotic 26S proteasomal Rpt subunits[18,19]. Of these, α1 and PAN-A are highly abundant during all phases of growth[19], double knockout of the Rpt-like genes has little impact on standard growth and synthesis of core particles containing either α1 or α2 can be separately abolished[17]. However, conditional knockout of all core particle subtypes renders cells inviable[17].

Analysis of the Flag–SAMP fusions in the various proteasomal mutants revealed significant differences in SAMP-conjugate levels compared to wild type. A substantial increase in SAMP1-conjugate and decrease in SAMP2-conjugate levels was observed during nitrogen-limitation in Δ*panA* Δ*psmA* mutant strains (deficient in synthesis of PAN-A and α1), whereas deletion of S2P metalloprotease genes had no effect (Fig. 3). Consistent with this, Δ*panA* Δ*psmA* single and double knockouts have the most pronounced phenotypes of the viable proteasomal mutant strains of *H. volcanii*, with

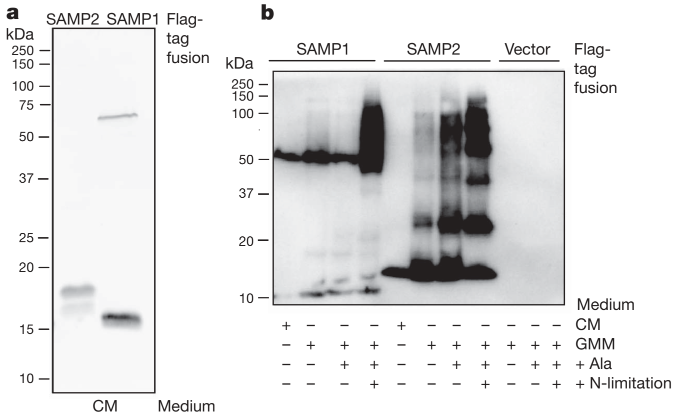

a
SAMP2 SAMP1 Flag-tag fusion
kDa
250
150
100
75
50
37
25
20
15
10
CM Medium
b
SAMP1 SAMP2 Vector Flag-tag fusion
kDa
250
150
100
50
37
20
10
Medium
+ – – – + – – – – – – CM
– + + + – + + + + + + GMM
– – + + – – + + – + + + Ala
– – – + – – – + – – – + N-limitation

Figure 2 | SAMP1 and SAMP2 are differentially conjugated to proteins and influenced by nitrogen-limitation. **a**, Anti-Flag immunoblot of SAMP1 and SAMP2 expressed as N-terminal Flag-tagged fusions in *H. volcanii* cells grown on complex medium (CM). **b**, Flag–SAMP fusions similarly expressed and analysed from cells grown on CM, glycerol minimal medium (GMM), GMM supplemented with alanine (+ Ala) and GMM + Ala devoid of NH$_4$Cl (+ N-limitation). All details on experimental procedures and strains are available as Supplementary Data.

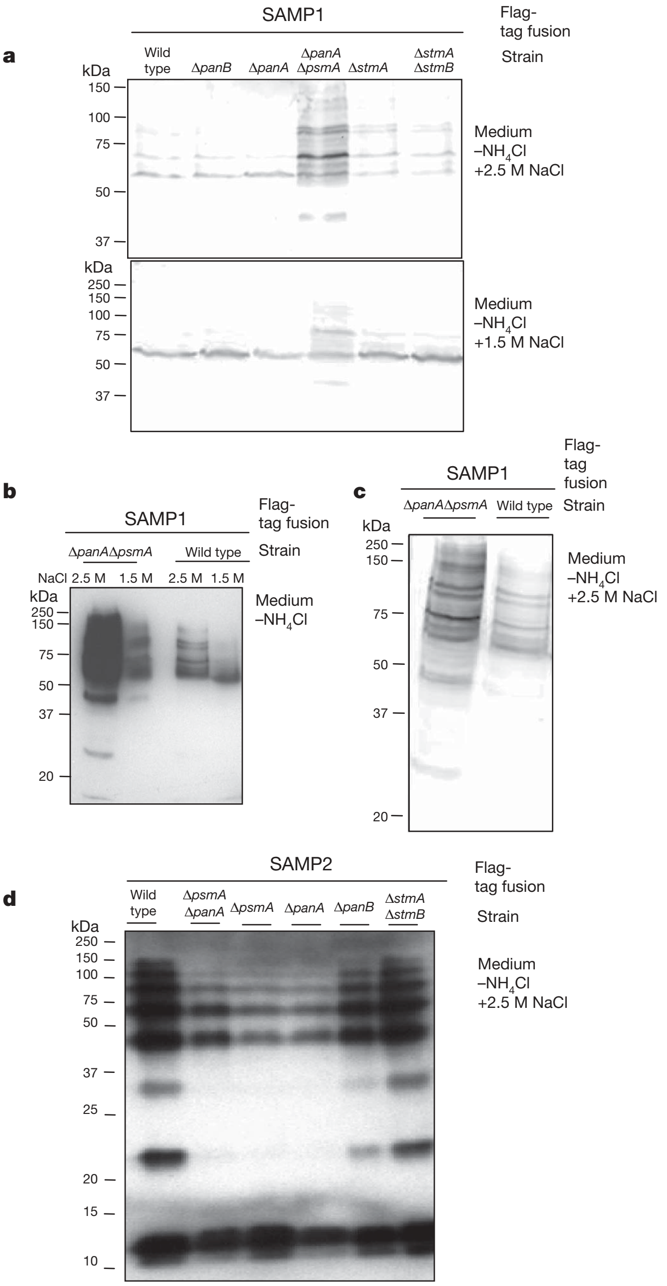

a
SAMP1 Flag-tag fusion
Wild type ΔpanB ΔpanA ΔpanA ΔpsmA ΔstmA ΔstmA ΔstmB Strain
kDa
150
100
75
50
37
Medium –NH4Cl +2.5 M NaCl
kDa
250
150
100
75
50
37
Medium –NH4Cl +1.5 M NaCl
b
SAMP1 Flag-tag fusion
ΔpanAΔpsmA Wild type Strain
NaCl 2.5 M 1.5 M 2.5 M 1.5 M
kDa
250
150
75
50
37
20
Medium –NH4Cl
c
SAMP1 Flag-tag fusion
ΔpanAΔpsmA Wild type Strain
kDa
250
150
75
50
37
20
Medium –NH4Cl +2.5 M NaCl
d
SAMP2 Flag-tag fusion
Wild type ΔpsmA ΔpanA ΔpsmA ΔpanA ΔpanB ΔstmA ΔstmB Strain
kDa
250
150
100
75
50
37
25
20
15
10
Medium –NH4Cl +2.5 M NaCl

Figure 3 | SAMP-conjugates are altered by proteasomal gene knockout. **a–c**, Anti-Flag immunoblot of SAMP1 expressed as an N-terminal Flag-tagged fusion in *H. volcanii* wild type and protease mutant strains grown under nitrogen-limiting conditions with 2.5 or 1.5 M NaCl as indicated. **d**, SAMP2 was similarly expressed and analysed in wild type and mutant strains. SAMP1-conjugate levels of Δ*psmA* and Δ*panA* Δ*panB* mutant strains were similar to wild type, and SAMP-conjugates were not detected in strains with vector alone (data not shown). *psmA* (core particle α1), *panA* and *panB* (Rpt-like AAA ATPases), Δ*stmA* and Δ*stmB* (site-2 type metalloprotease homologues HVO_1870 and HVO_1862, respectively).

enhanced sensitivity to nitrogen-limitation, hypo-osmotic shock and the amino acid analogue L-canavanine[17]. The enhanced levels of SAMP1-conjugates in the Δ*panA* Δ*psmA* mutant suggest SAMP1 targets proteins for destruction by proteasomes. Other functions of SAMPylation are also likely based on the decrease in SAMP2-conjugates observed in select proteasomal mutant strains.

Identification of SAMP conjugates

SAMP conjugates were purified from *H. volcanii* cells expressing the Flag–SAMP fusions by anti-Flag immunoprecipitation compared to cells expressing the Flag–SAMP fusions with deletions in their C-terminal diglycine motif (ΔGG) or vector alone (Fig. 4). Unlike most organisms, the vast majority of proteins from haloarchaea are highly acidic and require high salt (>1 M) for stability and activity[20]. Non-covalent protein complexes from these 'salt-loving' organisms typically dissociate in the low salt and detergent conditions required for immunoprecipitation. Consistent with this, SAMP conjugates were readily purified by immunoprecipitation from *H. volcanii* based on anti-Flag immunoblot and SYPRO Ruby stain of these fractions (Fig. 4). The purified SAMP conjugates were resistant to boiling in the presence of SDS and reducing reagents (Fig. 4a). The results also demonstrated that the C-terminal diglycine motif of SAMP1 and SAMP2 was required for their conjugation to proteins and that immunoprecipitation enhanced the ability to detect a notable diversity of SAMP conjugates present in cells grown under rich and nitrogen-limiting conditions. It should also be noted that the SAMP-conjugate banding patterns were not influenced by addition of reducing reagents. Thus, immunoprecipitation combined with boiling, separation by SDS–PAGE and staining with SYPRO Ruby proved ideal for the isolation of covalently-linked Flag–SAMP conjugates (Fig. 4b). Proteins specific for the Flag–SAMP-expressing strains were excised from the gels, digested with trypsin and identified by mass spectrometry (MS). Using this approach, 34 SAMP protein conjugates were identified, including those present in cells grown under nutrient-rich and nitrogen-limiting conditions (Table 1). Of the proteins identified, all were unique to the strains expressing the Flag–SAMP fusions compared to cells with vector alone, and three were common to both SAMP1 and SAMP2 (HVO_0558, HVO_0025 and HVO_A0230; Table 1). Consistent with their role as small archaeal modifier proteins, SAMP1 and SAMP2 were the only proteins identified in SDS–PAGE gel slices that spanned a wide-range of molecular masses (5–125 kDa, Supplementary Table 3).

Many of the SAMPylated proteins were homologues of enzymes associated with Ubl-conjugation and/or sulphur-activation (Table 1). These included homologues of Uba4p, Yor251c and Ncs6p/Ncs2p associated with the Urm1 pathway involved in thiolation of tRNA and protein conjugation[21,22] as well as MobB, MoaE, MoeA and SufB/D, all predicted to be involved in pathways associated with sulphur metabolism. Interestingly, homologues of the amino- and C-terminal domains of Uba4p are encoded as separate proteins in *H. volcanii* and other archaea. HVO_0558, identified as a SAMP-conjugate, is similar to the Uba4p N-terminal domain and Cys225 active site required for adenylyltransferase activity[21,23,24] (Fig. 5), whereas the divergently transcribed HVO_0559 is related to the Uba4p C terminus including the rhodanese domain (RHD) and Cys397 needed for persulphide formation in sulphurtransferase reactions[25]. Whether HVO_0558 functions as an E1 and activates the SAMPs for protein conjugation and/or sulphur transfer to tRNA or cofactors such as molybdopterin remains to be determined; however, its association with both SAMP proteins under all conditions examined and its relationship to the Urm1 pathway is consistent with this possibility.

A wide variety of proteins spanning functions from stress response to basic transcription, translation and DNA replication were also conjugated to the SAMPs (Table 1). Many of these proteins were previously found to accumulate in *H. volcanii* cells after chemical and/or genetic perturbation of proteasome function (as indicated in Table 1). Furthermore, many have been linked to Ubl/Ub-proteasome systems including the translation elongation factor EF-1α[26,27], predicted transcriptional regulator HVO_1577 associated with *H. volcanii* 20S core particles[28], Shwachman–Bodian–Diamond syndrome protein encoded in proteasomal operons in archaea[29] and HVO_1250 and HVO_1289 of similar antioxidant function to the Urm1 target Ahp1p[30].

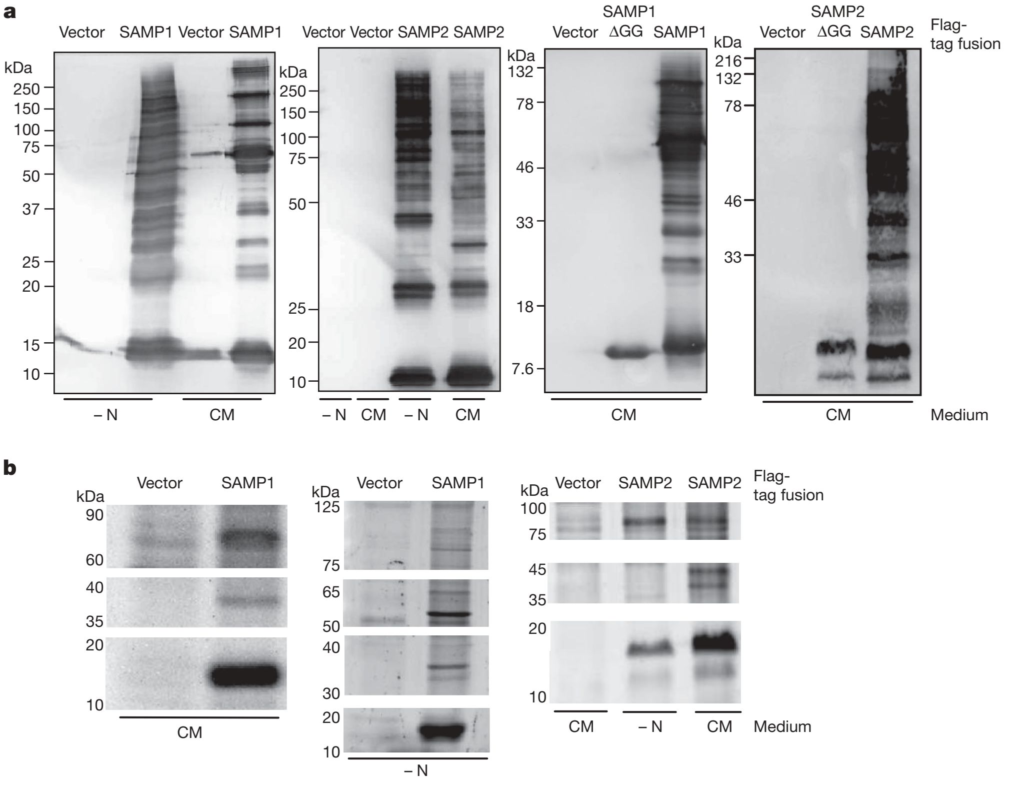


Figure 4 | SAMP-conjugates are isolated by immunoprecipitation. SAMP1 ± ΔGG and SAMP2 ± ΔGG were expressed as N-terminal Flag-tagged fusions in *H. volcanii* grown in complex medium (CM) and nitrogen-limiting conditions (– N). Proteins were immunoprecipitated with anti-Flag, boiled and separated by either: **a**, reducing 12% SDS–PAGE and analysis by anti-Flag immunoblot or **b**, non-reducing 12% SDS–PAGE and staining for total protein by SYPRO Ruby. Molecular mass standards and range of gel slices excised for MS-analysis are indicated on left. *H. volcanii* with vector alone served as a negative control in all experiments including MS-analysis of gel slices.

Table 1 | *H. volcanii* SAMPs and SAMP-conjugates identified by MS*

Protein	Homologue/description	CM	−N	CM	−N	Relation to Ub, sulphur and proteasomes
		Flag–SAMP1		Flag–SAMP2		
Ubl-/S-chemistry:						
HVO_2619	SAMP1	+	+	–	–	Ubl β-grasp
HVO_0202	SAMP2	–	–	+	+	Ubl β-grasp
HVO_0558	UBA/E1/MoeB, Ub- and sulphur-activating enzymes	+	+	+	+	Homologue of the N-terminal domain of Uba4p, the E1-enzyme of the Urm1 pathway
HVO_1864	N-terminal domain related to MobB P-loop NTPase; C-terminal domain related to MoaE sulphur-conjugating enzyme	+	+	–	–	S-conjugation
HVO_2305	MoeA, functions with MoaB in metal insertion into molybdopterin	–	–	+	–	Mo/W-insertion
HVO_0025	SseA/TssA, tandem RHD thiosulfate sulphurtransferase	–	+	–	+	Homologue of Urm1-associated Yor251cp[21,25]
HVO_0861	SufB/SufD, cysteine desulphurase activator subunit	–	–	–	+	Cysteine desulphurase activator; accumulates in HVO after cLβL treatment[36]
HVO_0580	N-type ATP pyrophosphatases and ATP sulphurylases	–	–	+	+	Homologue of Urm1-associated Ncs6p, functions in tRNA adenylation[21,25,32]
N-limitation/stress response:						
HVO_A0230	MsrA, methionine-*S*-sulfoxide reductase	–	+	+	+	
HVO_2402	Glycine cleavage P-protein, catalyses initial step of oxidative cleavage of glycine to NH_4^+, CO_2 and methylene group ($-CH_2-$)	–	+	–	–	
HVO_2900	FumC, ROS-resistant fumarase C	–	+	–	–	
HVO_1289	OsmC, osmotically inducible protein C peroxiredoxin	–	–	–	+	OsmC accumulates in HVO Δ*panA*[37], Ahp1p is a peroxiredoxin and the only known target of urmylation[30]
HVO_1250	Peroxiredoxin-/thioredoxin-like	–	–	+	–	
HVO_2682	Dodecin-flavoprotein, may prevent riboflavin degradation and trap phototoxic lumichrome waste	–	–	–	+	
Metabolism:						
HVO_2583	HmgA, 3-hydroxy-3-methylglutaryl CoA reductase	–	–	+	–	
HVO_2328	Isochorismatase	–	–	+	–	
HVO_1545	DhaL, dihydroxyacetone kinase (DHAK) subunit	–	–	+	–	Components of DHAK-PTS system accumulate in HVO after cLβL treatment and Δ*panA*[36,37]
HVO_1496	PtsI, PTS system EI	–	–	+	–	
HVO_0481	GAPDH, glyceraldehyde-3-P dehydrogenase	–	–	+	–	HVO_0480 (3-phospho-glycerate kinase) encoded within operon accumulates in HVO Δ*panA*[37]
HVO_1000	Acetyl-CoA synthetase	–	–	+	–	
HVO_0887	2-oxoglutarate oxidoreductases, β	–	–	+	–	Homologue HVO_1304, accumulates in HVO after cLβL treatment[36]
HVO_A0379	agaF, N-methyl-hydantoinase A	–	–	+	–	HVO_A0378 (oxoprolinase homologue) within operon accumulates in HVO after cLβL treatment[36]
HVO_0980	NdhG, NADH-quinone oxidoreductase, chain c/d	–	–	+	–	
DNA replication, transcription, translation and RNA processing:						
HVO_1727	TATA-box binding protein E	–	–	–	+	
HVO_1478	TFB, transcription initiation factor	–	–	+	–	
HVO_0359	EF-1α, translation elongation factor	–	+	–	–	Accumulates in HVO after cLβL treatment[36], putative isopeptidase[26,27]
HVO_0966	aIF2ba, archaeal translation initiation factor	–	–	+	+	
HVO_1921	SerS, seryl-tRNA synthetase	–	–	+	–	
HVO_0677	AspS, aspartyl-tRNA synthetase	–	–	+	–	
HVO_1572	GyrB, DNA gyrase B	–	–	+	–	
HVO_1344	Shwachman–Bodian–Diamond syndrome protein, putative role in RNA metabolism	–	–	+	–	Gene neighbour of archaeal α-type 20S proteasomal subunits[29]
HVO_1577	Putative winged-helix transcriptional regulator, C-terminal CBS domains	–	–	+	–	HVO 20S proteasome-associated protein[28]
Conserved:						
HVO_0736	DUF302	–	–	–	+	
HVO_B0053	C-terminal H-X_3-H motif protein	–	–	–	+	

* –, Undetectable; +, MS-identified protein conjugate unique to immunoprecipitation fractions of *H. volcanii* strains expressing the Flag-tagged β-grasp Ub-like protein SAMP1 or SAMP2 compared to vector alone. cLβL, *clasto*-lactacystin β-lactone proteasome inhibitor. DHAK-PTS, dihydroxyacetone kinase linked to phosphotransferase system. Cells were grown on complex medium (CM) or under nitrogen-limiting conditions (−N). Protein identities are reported according to the *H. volcanii* gene locus tag from the USCS Archaeal Genome Browser (April 2007 version). SAMP conjugates were reproducibly purified by immunoprecipitation with anti-Flag, boiled in SDS buffer, separated by SDS–PAGE and analysed by immunoblot with anti-Flag. Only anti-Flag reactive bands were further analysed by MS for protein identity and covalent linkages. Proteins were identified using a hybrid quadrupole-time of flight (ABI QSTAR XL) and hybrid quadrupole-linear ion trap (ABI 4000 QTRAP). All details on experimental procedures, MS data and FASTA files of identified protein sequences are available as Supplementary Data.

Mapping sites of SAMPylation

To enhance MS coverage and map the sites of SAMPylation, Flag–SAMP2 conjugates were purified by anti-Flag in liquid phase for analysis of trypsinized peptides by reversed phase liquid chromatography coupled with tandem mass spectrometry (RP-LC-MS/MS) using a data dependent MS/MS scan mode and parent mass list method. Unlike SAMP1, which has a limited number of C-terminal trypsin cleavage sites, SAMP2 has a lysine at position 64. Thus, if an isopeptide bond is formed between the C-terminal carboxylate of SAMP2 and an amino group of the substrate protein, SAMP2 will leave a 'GG-footprint' on the target site after trypsinization. Using this approach, eleven sites of SAMP2 modification were mapped by collision-induced dissociation (CID) based MS/MS (Table 2). The sites were based on the mass differences between the y and b ion series containing the SAMP2-derived GG-footprint on lysine residues (Fig. 6 and Supplementary Fig. 2). The SAMPylated peptides were detected from doubly to quadruply charged molecular ions and mapped by more than one peptide on the same protein.

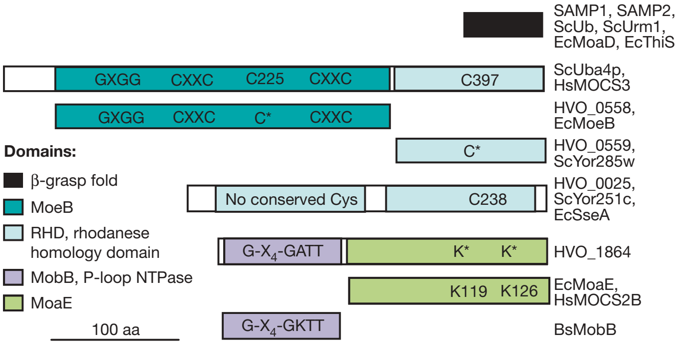


Figure 5 | SAMP and SAMP-conjugates are related to sulphur-activation and ubiquitination pathways. SAMP1 and SAMP2 cluster to the β-grasp superfamily. HVO_0558 is related to Uba4p of the Urm1 pathway and molybdopterin (MPT) synthase sulphurases (for example MoeB, MOCS3). Although the RHD common to Uba4p is not conserved in HVO_0558 it is found in the gene neighbour HVO_0559. HVO_1864 is related to MoaE proteins that associate with β-grasp proteins to form active MPT synthases[38,39] and MobB, a P-loop NTPase of MPT synthesis[40]. HVO_0025 is a dual RHD protein related to 3-mercaptopyruvate sulphurtransferases that form persulphide intermediates[41,42] and ScYor251c of the Urm1 pathway[21].

A number of fundamental insights were revealed by the CID-based MS/MS spectra concerning how SAMP2 modifies proteins. First and foremost, the C-terminal glycine of SAMP2 is covalently attached through an isopeptide bond to the ε-amino group of lysine residues of at least nine different substrate proteins. Second, SAMP2 can modify a single substrate protein at several sites based on the finding that TATA-binding protein E (HVO_1727) and SseA/Yor251cp (HVO_0025) homologues are modified at either of two lysine residues in close proximity. Third, although a thioester bond was not detected by MS/MS between SAMP2 and any of the cysteine residues of HVO_0558, SAMP2 did modify this Uba4p homologue through an isopeptide bond at K113, suggesting SAMP2 may regulate the adenylation of either itself or SAMP1. Furthermore, the MS/MS data revealed that SAMP2 forms polymeric chains with itself at lysine 58 similar to Ub and other Ubl proteins (such as SUMO2/3 and NEDD8)[31]. Whether the SAMP2 polymeric chains are free or covalently attached to substrate proteins and the full diversity of these SAMP2 chains (that is, homotypic, heterologous or mixed with SAMP1) remain to be determined. Likewise, it remains to be determined whether SAMP1 and SAMP2 compete for the same or different lysine residues on substrates and target these proteins for different fates, or whether they are mutually exclusive in their sites of protein targeting. Proteins with multiple SAMP sites occupied remain to be identified. Our results do show, however, that the same protein can be modified by either SAMP1 or SAMP2 (that is, Uba4p and MsrA homologues) and that the same protein can be modified on different lysine residues (that is, TATA-binding protein E and SseA/Yor251cp homologues).

Widespread distribution of SAMP homologues

Although SAMP1 and SAMP2, share limited primary sequence identity to each other, both proteins are members of a large superfamily that shares a common β-grasp fold and includes members from all archaea[10–12]. In addition to this common three dimensional fold, SAMP1 and SAMP2 are related in primary amino acid sequence to small proteins from other archaea including species of haloarchaea, methanogens and *Archaeoglobus* (30 to 80% identity) (Supplementary Figs 3 and 4). SAMP1 also shares a close relationship with the N termini of small proteins that have a C-terminal domain of unknown function (DUF1952) from thermophilic bacteria of the deep branching *Thermus* species (33 to 39% identity) (Supplementary Fig. 3). Interestingly, a number of these SAMP homologues (five from haloarchaea and two from *Thermus*) have 2 to 82 amino acid residues carboxyl to the diglycine motif and, thus, would probably require proteolytic cleavage before covalent attachment if functioning similar to the *H. volcanii* SAMPs.

The organization of the SAMP1 and SAMP2 genes on the *H. volcanii* genome is also revealing (Supplementary Fig. 5). Unlike eukaryotes that encode Ub as fusion proteins that are proteolytically processed to expose a functional C-terminal diglycine motif, SAMP1 and SAMP2 are encoded as single small proteins (of 87 and 66 amino acids, respectively) with the diglycines apparently exposed after translation. Comparison of the SAMP operons to other microbial genomes reveals a high conservation of immediate gene order between *H. volcanii* and other diverse haloarchaea. This includes the prediction that SAMP1 is co- and divergently transcribed with genes encoding proteins with regulatory of K^+ conductance (RCK) domains likely to form K^+ channels for cellular defence against osmotic stress. Likewise, haloarchaeal SAMP2 genes seem to be commonly co- and divergently transcribed with Gcn5-related N-acetyltransferase (GNAT) and AAA ATPase replication factor C small subunit homologues. This conservation in gene order suggests that SAMPylation is linked to osmotic stress, DNA replication and/or protein acetylation. Although SAMP-conjugates were not altered by low salt stress (Fig. 3a and data not shown), a strong and constitutive rRNA P2 promoter was used to drive expression of the Flag–SAMP genes for this analysis. Interestingly, we did detect an increase in the levels and change in the types of SAMP-conjugates formed during nitrogen-limitation suggesting stress and/or reduced growth rate may be associated with SAMP function.

Table 2 | SAMP2-conjugate sites mapped by MS/MS*

No	ORF no	Protein description	*z*	Mass accuracy (p.p.m.)	Xcorr	Sf	Residue modified	Peptides
1	HVO_0202	SAMP2	2	1.2	2.01	0.71	K58	(R)VK@VLR(L)
2	HVO_0966	eif2ba/aIF-2BII translation initiation factor	4	−0.1	4.90	0.94	K210	(R)YLNDVDHVLVGADAVAADGSVINK@IGTSGLAVNAR(E)
			3	−3.3	7.61	0.99		
3	HVO_1572	GyrB, DNA gyrase B subunit	2	−13.2	0.97	0.30	K624	(R)K@QFIK(D)
4	HVO_2328	Isochorismatase	4	0.4	4.01	0.94	K90	(R)SDGEGFAWKPEAEPVDGEPVFTK@R(V)
			3	0.3	5.52	0.97		
			2	−0.2	3.59	0.92		
5	HVO_0558	MoeB, molybdopterin biosynthesis protein	3	0.5	4.12	0.93	K113	(R)VDK@SNVHEVVAGSDVVVDASDNFPTR(Y)
6	HVO_0980	NdhG, NADH-quinone oxidoreductase chain c/d	2	28.5	0.73	0.47	K517	(R)FK@IR(S)
7	HVO_1289	OsmC-like protein superfamily	2	3.9	2.78	0.69	K59	(R)VGGQK@TGFDDLGK(V)
8	HVO_1727	TATA-box binding protein E	2	7.5	3.27	0.91	K63	(R)SGK@IVC#TGAK(S)
			2	−0.1	2.65	0.88	K53	(R)TQDPK@AAALIFR(S)
9	HVO_0025	SseA/TssA, tandem RHD thiosulfate sulphurtransferase	2	−2.2	3.14	0.81	K162	(R)AYRDDVEK@AVDK(G)
			2	0.6	2.00	0.52	K166	(K)AVDK@GLPLVDVR(S)

* *z*, charge state; Xcorr, cross correlation; Sf, final score; @, SAMP2-modification; #, alkylated cysteine.

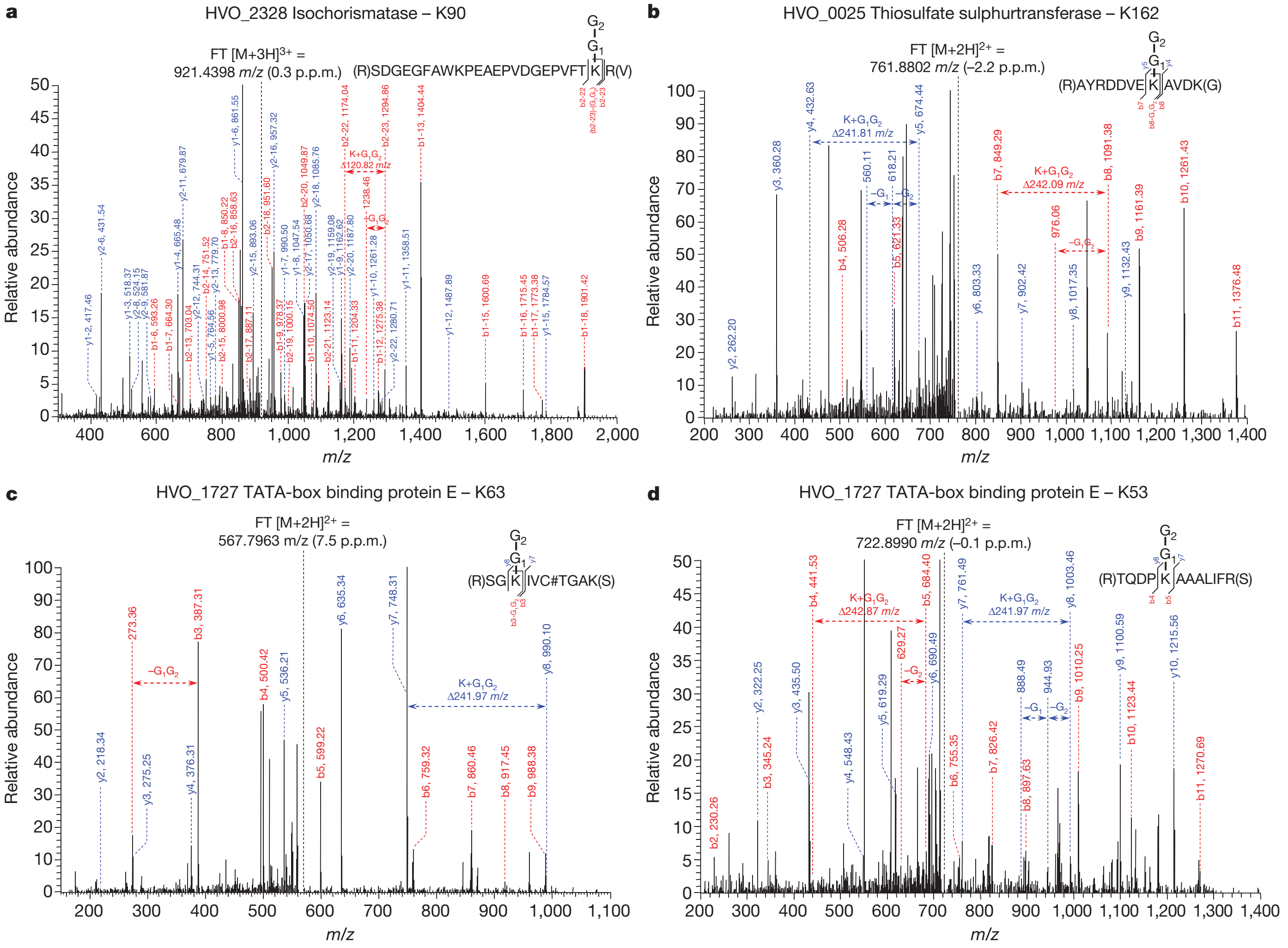


Figure 6 | MS/MS spectra of SAMP2-conjugate sites. a, SAMP2-modification of HVO_2328 K90 based on mass difference between b2-22 and b2-23 ions and loss of Gly$_1$-Gly$_2$ at 1238.46 *m*/*z* from the b2-23 ion derived from the triply charged precursor ion. **b**, SAMP2-modification of HVO_0025 K162 based on mass difference between both ion series derived from the doubly charged precursor ion: y4 and y5 ions and loss of Gly$_1$-Gly$_2$ at 618.21 and 560.11 *m*/*z*, b7 and b8 ions and loss of Gly$_1$-Gly$_2$ at 976.06 *m*/*z*. **c**, **d**, SAMP2-modification of HVO_1727 K63 and K53. SAMP2 C-terminal diglycine (−Gly$_1$−Gly$_2$). Other MS/MS spectra, Supplementary Fig. 2.

Discussion

H. volcanii forms a relatively elaborate network of protein conjugates, including the covalent linkage to target proteins of at least two different Ubl-proteins, SAMP1 and SAMP2, that are conserved among diverse archaea. These data indicate ubiquitin-like protein conjugation system origins reside in archaea. *H. volcanii* forms these differential SAMP-conjugates in the presence of only a single E1 and in the absence of any apparent E2 or E3 homologues suggesting a streamlined Ubl system for protein conjugation. In support of this possibility, (1) the related eukaryotic E1s can be relatively promiscuous and activate more than one type of structurally distinct Ubl protein[32], (2) E2-intermediates have yet to be identified for the ancient Urm1 pathway and (3) ubiquitination can occur in the absence of E3 Ub-ligases[33]. Thus, an archaeal E2- and E3-independent Ubl-conjugation mechanism is feasible. Their conjugation to an E1 (Uba4p N-terminal domain) homologue was common to SAMP1 and SAMP2 under all conditions examined, suggesting a close association of this E1-like protein with SAMPylation. The multiple RHD proteins that are related to the C terminus of Uba4p and encoded as separate proteins in most archaea, including *H. volcanii*, may add functional flexibility to the SAMPylation system. Small Zn-finger proteins such as Brz[34], prevalent in archaea and similar to the RING domains of E3 Ub ligases[35], may also assist in discerning the various interactions required for SAMPylation. Although the full extent of poly- versus mono-SAMPylation and whether the SAMPs are reused has yet to be determined, SAMP2-polymeric chains were detected in this study and archaea encode proteins with JAMM motifs similar to eukaryotic deubiquitinating enzymes[14–16], suggesting a SAMP-recycling mechanism is conserved.

METHODS SUMMARY

Small proteins were selected from the deduced proteome of *H. volcanii* based on the presence of a β-grasp fold and C-terminal diglycine motif. N-terminal Flag-tagged fusions of these proteins were expressed in *H. volcanii* (± proteasomal gene mutations) grown under rich and nitrogen-limiting conditions. Formation of SAMP-conjugates was monitored by anti-Flag immunoblot of cell lysate that was separated by reducing SDS–PAGE. SAMP conjugates were enriched from cell lysate by anti-Flag immunoprecipitation and further purified by SDS–PAGE before identification by MS (compared to cells with Flag–SAMPΔGG or vector alone). Sites of SAMPylation were mapped by LC-MS/MS-based CID of Flag–SAMP2 conjugates purified by anti-Flag chromatography.

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

Received 28 September; accepted 10 November 2009.

1. Hochstrasser, M. Origin and function of ubiquitin-like proteins. *Nature* 458, 422–429 (2009).
2. Pickart, C. M. & Fushman, D. Polyubiquitin chains: polymeric protein signals. *Curr. Opin. Chem. Biol.* 8, 610–616 (2004).

3. Xu, P. *et al.* Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell* **137,** 133–145 (2009).
4. Ciechanover, A. & Ben-Saadon, R. N-terminal ubiquitination: more protein substrates join in. *Trends Cell Biol.* **14,** 103–106 (2004).
5. Burns, K. E., Liu, W. T., Boshoff, H. I., Dorrestein, P. C. & Barry, C. E. III. Proteasomal protein degradation in *Mycobacteria* is dependent upon a prokaryotic ubiquitin-like protein. *J. Biol. Chem.* **284,** 3069–3075 (2009).
6. Pearce, M. J., Mintseris, J., Ferreyra, J., Gygi, S. P. & Darwin, K. H. Ubiquitin-like protein involved in the proteasome pathway of *Mycobacterium tuberculosis*. *Science* **322,** 1104–1107 (2008).
7. Striebel, F. *et al.* Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes. *Nature Struct. Mol. Biol.* **16,** 647–651 (2009).
8. Liao, S. *et al.* Pup, a prokaryotic ubiquitin-like protein, is an intrinsically disordered protein. *Biochem. J.* **422,** 207–215 (2009).
9. Chen, X. *et al.* Prokaryotic ubiquitin-like protein pup is intrinsically disordered. *J. Mol. Biol.* **392,** 208–217 (2009).
10. Iyer, L. M., Burroughs, A. M. & Aravind, L. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like β-grasp domains. *Genome Biol.* **7,** R60 (2006).
11. Burroughs, A. M., Balaji, S., Iyer, L. M. & Aravind, L. A novel superfamily containing the β-grasp fold involved in binding diverse soluble ligands. *Biol. Direct* **2,** 4 (2007).
12. Burroughs, A. M., Iyer, L. M. & Aravind, L. Natural history of the E1-like superfamily: Implication for adenylation, sulfur transfer, and ubiquitin conjugation. *Proteins* **75,** 895–910 (2009).
13. Kessler, D. Enzymatic activation of sulfur for incorporation into biomolecules in prokaryotes. *FEMS Microbiol. Rev.* **30,** 825–840 (2006).
14. Yao, T. & Cohen, R. E. A cryptic protease couples deubiquitination and degradation by the proteasome. *Nature* **419,** 403–407 (2002).
15. Verma, R. *et al.* Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome. *Science* **298,** 611–615 (2002).
16. Cope, G. A. *et al.* Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298,** 608–611 (2002).
17. Zhou, G., Kowalczyk, D., Humbard, M. A., Rohatgi, S. & Maupin-Furlow, J. A. Proteasomal components required for cell growth and stress responses in the haloarchaeon *Haloferax volcanii*. *J. Bacteriol.* **190,** 8096–8105 (2008).
18. Kaczowka, S. J. & Maupin-Furlow, J. A. Subunit topology of two 20S proteasomes from *Haloferax volcanii*. *J. Bacteriol.* **185,** 165–174 (2003).
19. Reuter, C. J., Kaczowka, S. J. & Maupin-Furlow, J. A. Differential regulation of the PanA and PanB proteasome-activating nucleotidase and 20S proteasomal proteins of the haloarchaeon *Haloferax volcanii*. *J. Bacteriol.* **186,** 7763–7772 (2004).
20. Albuquerque, C. P. *et al.* A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell. Proteomics* **7,** 1389–1396 (2008).
21. Leidel, S. *et al.* Ubiquitin-related modifier Urm1 acts as a sulphur carrier in thiolation of eukaryotic transfer RNA. *Nature* **458,** 228–232 (2009).
22. Schlieker, C. D., Van der Veen, A. G., Damon, J. R., Spooner, E. & Ploegh, H. L. A functional proteomics approach links the ubiquitin-related modifier Urm1 to a tRNA modification pathway. *Proc. Natl Acad. Sci. USA* **105,** 18255–18260 (2008).
23. Furukawa, K., Mizushima, N., Noda, T. & Ohsumi, Y. A protein conjugation system in yeast with homology to biosynthetic enzyme reaction of prokaryotes. *J. Biol. Chem.* **275,** 7462–7465 (2000).
24. Schmitz, J. *et al.* The sulfurtransferase activity of Uba4 presents a link between ubiquitin-like protein conjugation and activation of sulfur carrier proteins. *Biochemistry* **47,** 6479–6489 (2008).
25. Noma, A., Sakaguchi, Y. & Suzuki, T. Mechanistic characterization of the sulfur-relay system for eukaryotic 2-thiouridine biogenesis at tRNA wobble positions. *Nucleic Acids Res.* **37,** 1335–1352 (2009).
26. Gonen, H. *et al.* Protein synthesis elongation factor EF-1α is essential for ubiquitin-dependent degradation of certain Nα-acetylated proteins and may be substituted for by the bacterial elongation factor EF-Tu. *Proc. Natl Acad. Sci. USA* **91,** 7648–7652 (1994).
27. Gonen, H., Dickman, D., Schwartz, A. L. & Ciechanover, A. Protein synthesis elongation factor EF-1α is an isopeptidase essential for ubiquitin-dependent degradation of certain proteolytic substrates. *Adv. Exp. Med. Biol.* **389,** 209–219 (1996).
28. Humbard, M. A., Stevens, S. M. Jr & Maupin-Furlow, J. A. Posttranslational modification of the 20S proteasomal proteins of the archaeon *Haloferax volcanii*. *J. Bacteriol.* **188,** 7521–7530 (2006).
29. Maupin-Furlow, J. A., Wilson, H. L., Kaczowka, S. J. & Ou, M. S. Proteasomes in the archaea: from structure to function. *Front. Biosci.* **5,** d837–d865 (2000).
30. Goehring, A. S., Rivers, D. M. & Sprague, G. F. Jr. Attachment of the ubiquitin-related protein Urm1p to the antioxidant protein Ahp1p. *Eukaryot. Cell* **2,** 930–936 (2003).
31. Ikeda, F. & Dikic, I. Atypical ubiquitin chains: new molecular signals. 'Protein Modifications: Beyond the Usual Suspects' review series. *EMBO Rep.* **9,** 536–542 (2008).
32. Schulman, B. A. & Harper, J. W. Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nature Rev. Mol. Cell Biol.* **10,** 319–331 (2009).
33. Hoeller, D. *et al.* E3-independent monoubiquitination of ubiquitin-binding proteins. *Mol. Cell* **26,** 891–898 (2007).
34. Tarasov, V. Y. *et al.* A small protein from the *bop–brp* intergenic region of *Halobacterium salinarum* contains a zinc finger motif and regulates *bop* and *crtB1* transcription. *Mol. Microbiol.* **67,** 772–780 (2008).
35. Borden, K. L. RING fingers and B-boxes: zinc-binding protein-protein interaction domains. *Biochem. Cell Biol.* **76,** 351–358 (1998).
36. Kirkland, P. A., Reuter, C. J. & Maupin-Furlow, J. A. Effect of proteasome inhibitor *clasto*-lactacystin-β-lactone on the proteome of the haloarchaeon *Haloferax volcanii*. *Microbiology* **153,** 2271–2280 (2007).
37. Kirkland, P. A., Gil, M. A., Karadzic, I. M. & Maupin-Furlow, J. A. Genetic and proteomic analyses of a proteasome-activating nucleotidase a mutant of the haloarchaeon *Haloferax volcanii*. *J. Bacteriol.* **190,** 193–205 (2008).
38. Leimkuhler, S., Freuer, A., Araujo, J. A., Rajagopalan, K. V. & Mendel, R. R. Mechanistic studies of human molybdopterin synthase reaction and characterization of mutants identified in group B patients of molybdenum cofactor deficiency. *J. Biol. Chem.* **278,** 26127–26134 (2003).
39. Matthies, A., Rajagopalan, K. V., Mendel, R. R. & Leimkuhler, S. Evidence for the physiological role of a rhodanese-like protein for the biosynthesis of the molybdenum cofactor in humans. *Proc. Natl Acad. Sci. USA* **101,** 5946–5951 (2004).
40. McLuskey, K., Harrison, J. A., Schuttelkopf, A. W., Boxer, D. H. & Hunter, W. N. Insight into the role of *Escherichia coli* MobB in molybdenum cofactor biosynthesis based on the high resolution crystal structure. *J. Biol. Chem.* **278,** 23706–23713 (2003).
41. Colnaghi, R., Cassinelli, G., Drummond, M., Forlani, F. & Pagani, S. Properties of the *Escherichia coli* rhodanese-like protein SseA: contribution of the active-site residue Ser240 to sulfur donor recognition. *FEBS Lett.* **500,** 153–156 (2001).
42. Spallarossa, A. *et al.* The ''rhodanese'' fold and catalytic mechanism of 3-mercaptopyruvate sulfurtransferases: crystal structure of SseA from *Escherichia coli*. *J. Mol. Biol.* **335,** 583–593 (2004).

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

Acknowledgements We thank the staff at UF ICBR including C. Diaz and R. Zheng for MS and S. Shanker for DNA sequencing. Thanks to M. Terns and F. Aydemir for advice on purification of SAMP conjugates from anti-Flag agarose, N. Furlow for plasmid DNA preparations and J. Foster for other advice. We thank also T. Allers, M. Mevarech, M. Dyall-Smith and M. Danson labs for *H. volcanii* strains and plasmids. This work was funded in part by NIH 1S10 RR025418-01 to SC, Integrated Technology Resource for Biomedical Glycomics at UGA (supported by NIH/NCRR P41 RR018502, L.W. senior investigator) and NIH R01 GM057498 and DOE DE-FG02-05ER15650 to J.A.M.-F.

Author Contributions J.A.M.-F., M.A.H., D.J.K., J.R.P. and H.V.M. performed cloning and immunoblot experiments. M.A.H. and H.V.M. purified SAMP conjugates by anti-Flag immunoprecipitation and chromatography. G.Z. transformed *H. volcanii* and prepared media. S.C. directed the identification of SAMP conjugates by MS. J.-M.L. and L.W. mapped the SAMP-conjugate sites by CID-based MS/MS. J.A.M.-F., L.W., M.A.H. and J.-M.L. interpreted the data. J.A.M.-F.planned the studies and wrote the manuscript. All authors commented on the manuscript.

Author Information 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 J.A.M.-F. (jmaupin@ufl.edu).

METHODS

Materials. Biochemicals were purchased from Sigma-Aldrich. Other organic and inorganic analytical grade chemicals were from Fisher Scientific and Bio-Rad. Desalted oligonucleotides were from Integrated DNA Technologies. DNA modifying enzymes and polymerases were from New England Biolabs.

Strains, media and plasmids. *H. volcanii* and *Escherichia coli* strains, oligonucleotide primers used for cloning and plasmids are summarized in Supplementary Tables 1 and 2. *E. coli* DH5α was used for routine recombinant DNA experiments, and *E. coli* GM2163 was used for isolation of plasmid DNA for transformation of *H. volcanii* as described previously[43]. *H. volcanii* wild type and protease mutant strains expressing N-terminal Flag-tagged fusions were grown to stationary phase (D_{600} of 1.5–2.2) at 42 °C and 200 r.p.m. Media included ATCC 974, composed of 2.14 M NaCl, 246 mM $MgCl_2{\cdot}6H_2O$, 29 mM K_2SO_4, 0.91 mM $CaCl_2{\cdot}2H_2O$, 0.5% yeast extract (Difco) and 0.5% tryptone; YPC, composed of 0.5% yeast extract (Difco), 0.1% peptone (Oxoid), 0.1% casamino acids (Difco) with 18% salt water (2.5 M NaCl, 88 mM $MgCl_2{\cdot}6H_2O$, 85 mM $MgSO_4{\cdot}7H_2O$, 56 mM KCl, 3 mM $CaCl_2$) and 12 mM Tris-HCl buffer pH 7.5 according to ref. 43; and GMM, composed of 20 mM glycerol with 18% salt water, 5 mM NH_4Cl, trace minerals (1.8 μM $MnCl_2{\cdot}4H_2O$, 1.5 μM $ZnSO_4{\cdot}7H_2O$, 8.3 μM $FeSO_4{\cdot}7H_2O$, 0.2 μM $CuSO_4{\cdot}5H_2O$), cofactors (3 μM thiamine or vitamin B1 and 40 nM biotin or vitamin H) and buffers (42 mM Tris-HCl pH 7.5 and 1 mM KPO_4 pH 7.5). Media were supplemented with alanine (25 mM) (+Ala), devoid of ammonium chloride (–N) or reduced to 1.5 M NaCl as indicated. Media were also supplemented with novobiocin (0.1 μg ml^{-1}) and uracil (50 μg ml^{-1}) as needed. Uracil was solubilized in 100% dimethylsulphoxide (DMSO) at 50 mg ml^{-1} before addition to media.

DNA purification and analysis. The *H. volcanii* genes encoding HVO_2619 (SAMP1), HVO_0202 (SAMP2), HVO_2177, HVO_2178 and HVO_0383 were isolated from genomic DNA by PCR using primers listed in Supplementary Table 1, *H. volcanii* genomic DNA as template, Phusion DNA polymerase and 3% (v/v) DMSO according to supplier (New England Biolabs). PCR was performed with a thermal gradient for annealing at ± 5 °C primer T_m using an iCycler (Bio-Rad). PCR products were analysed on 2% (w/v) agarose gels in TAE buffer (40 mM Tris acetate, 2 mM EDTA, pH 8.5) using Hi-Lo DNA molecular weight marker (Minnesota Molecular) and ethidium bromide staining at 0.5 μg ml^{-1}. DNA fragments of appropriate molecular mass were purified by MinElute (Qiagen) or isolated from SeaKem GTG agarose (FMC Bioproducts) gels using the QIAquick gel extraction kit (Qiagen) as needed. DNA fragments were ligated into the NdeI and BlpI sites of pJAM202 or NdeI and KpnI sites of pJAM939 using appropriate restriction enzymes, Antarctic phosphatase and T4 ligase as indicated in Supplementary Tables 1 and 2. Plasmid DNA was isolated from *E. coli* strains using the QIAprep Spin Miniprep kit (Qiagen). Fidelity of all PCR amplified products was confirmed by sequencing the DNA of plasmid inserts by Sanger automated DNA sequencing using an Applied Biosystems Model 3130 Genetic analyser (UF ICBR Genomics Division).

Immunoblot. *H. volcanii* cells expressing N-terminal Flag-tagged fusions were harvested by centrifugation (10,000*g*, 10 min, 25 °C), boiled 20 to 30 min in SDS-loading buffer with reducing reagents (2.5% v/v β-mercaptoethanol or 10 mM dithiothreitol) and separated by SDS–PAGE (10 and 12%) at 0.065 D_{600} units per lane. Equivalent protein loading was confirmed by staining with Coomassie blue. Proteins were electroblotted onto Hybond-P polyvinylidene fluoride (PVDF) membranes (Amersham) (14.5 h at 20 V or 2.5 h at 90 V; 4 °C). Flag-tagged fusions were detected by immunoblot using (1) anti-Flag M2 antibody (Stratagene) and anti-mouse IgG-alkaline phosphatase antibody raised in goat (Sigma) and (2) alkaline phosphatase-linked anti-Flag M2 monoclonal antibody (Sigma). Alkaline phosphatase activity was detected colorimetrically using nitro blue tetrazolium chloride (NBT) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP) and by chemiluminescence using CDP-Star according to supplier's protocol (Applied Biosystems) with X-ray film (Hyperfilm; Amersham Biosciences).

Preparation of cell lysate for immunoprecipitation and Flag column elution. *H. volcanii* cells expressing Flag–SAMP fusions and vector alone (100 ml cultures) were harvested by centrifugation (6,000*g*, 20 min, 25 °C) and resuspended in 1 ml of lysis buffer (50 mM Tris-Cl buffer at pH 7.4 with 1% v/v Triton-X-100, 5 mM EDTA, 0.02% w/v sodium azide, 10 mM iodoacetamide, 1 mM PMSF, 300 mM NaCl, 1 U ml^{-1} DNase I). Debris was removed by centrifugation (14,000*g*, 20 min, 25 °C).

Immunoprecipitation. Anti-Flag M2 agarose (Sigma; product number A2220) was prepared for immunoprecipitation by washing twice with PBS and twice with wash buffer (50 mM Tris-Cl buffer at pH 7.4 with 0.1% v/v Triton-X-100, 300 mM NaCl, 5 mM EDTA, 0.02% w/v sodium azide, 0.1% w/v SDS, 0.1% w/v deoxycholate). Clarified cell lysate (1 ml) was added to washed agarose beads (100 μl) and incubated by rocking at 4 °C for 12–16 h. Protein-bound-beads were washed 10 times with wash buffer (1 ml per wash) and eluted with either SDS–PAGE or glycine buffer as described below.

For SDS–PAGE, proteins were eluted from beads by boiling for 10 min in 40 μl SDS–PAGE buffer (100 mM Tris-Cl buffer at pH 6.8 with 2% w/v SDS, 10% v/v glycerol, 0.6 mg ml^{-1} bromophenol blue). Sample (20 μl) was separated by 12% SDS–PAGE at 200 V for 40 to 50 min. Gels were stained with SYPRO Ruby according to manufacturer's protocol (BioRad) or developed with alkaline phosphatase-linked anti-Flag M2 monoclonal antibody as described above. Gels were imaged on a Bio-Rad XR imager and gel pieces were cut manually for mass spectrometry analysis by QSTAR and QTRAP (see below for details).

For glycine elution, 100 μl of 0.1 M glycine-HCl buffer at pH 2.5 was added to the protein-bound agarose beads and gently rocked (5 min at room temperature). The agarose beads were centrifuged (8,500*g*, 30 s at room temperature), and the supernatant was added to a sterile 1.5 ml microcentrifuge tube that contained 20 μl of 1 M Tris-HCl buffer at pH 8.0 supplemented with 1 M NaCl. The addition of 0.1 M glycine-HCl buffer at pH 2.5 was repeated twice to maximize elution from the beads, and eluted proteins were collected in the same 1.5 ml microcentrifuge tube.

Flag column elution. A polypropylene column (0.5 × 5 cm; Bio-Rad) was packed with anti-Flag M2 agarose to a total bed volume of 0.5 ml, as directed by the manufacturer (Sigma). After preparation of the resin, the column was equilibrated with 10 column volumes of TBS (50 mM Tris-HCl, 150 mM NaCl, pH 7.4). Clarified lysate (1 ml) (prepared as described above) was applied to the column (4 ×) and washed with 20 column volumes of TBS. Bound proteins were eluted with five column volumes of TBS containing 1× Flag peptide (Sigma) at 100 μg ml^{-1}. Eluted proteins were collected in nine fractions (~300 μl each). The column was regenerated immediately after use with three column volumes of 0.1 M glycine-HCl, pH 3.5, re-equilibrated with 13 column volumes of TBS, and stored in TBS with 50% v/v glycerol and 0.02% w/v sodium azide, as directed by the manufacturer (Sigma). All buffers were filtered with a 0.45 μm surfactant-free cellulose acetate (SFCA) filter (Nalgene Nunc) before use.

Mass spectrometry. SAMP-conjugates were identified from SYPRO-Ruby stained SDS–PAGE gels by mass spectrometry using a QTRAP triple quadrupole ion-trap mass spectrometer and a QSTAR quadrupole time-of-flight mass spectrometer with an inline capillary reverse-phase high-performance liquid chromatography (HPLC) separation of protein digests (UF ICBR Proteomics Division). A PepMap C18 column (75-μm inside diameter, 15-cm length; LC Packings) was used for reverse-phase separation in combination with an Ultimate capillary HPLC system (LC Packings) operated at a flow rate of 200 nl min^{-1} with a 60-min gradient from 5 to 50% v/v acetonitrile in 0.1% v/v acetic acid. In-gel proteins were extracted by successive washes of gel slices in acetonitrile to a final volume of 100 μl. Extracted proteins were dried under vacuum centrifugation. The resulting desiccant was suspended in 100 μl of 50 mM NH_4HCO_3 (pH 7.5). Samples were reduced by the addition of 5 μl of 200 mM dithiothreitol (DTT solution) for 1 h at 25 °C. Samples were alkylated by the addition of 4 μl of 1 M iodoacetamide for 1 h at 25 °C. Alkylation was stopped by the addition of 20 μl of DTT solution. Samples were digested with a 1:20 ratio of mg trypsin or AspN to mg protein for 18–24 h at 37 °C. Digested peptides were purified using 300 μl C18 spin columns and dried under vacuum centrifugation. The resulting dessicant was resuspended in 5–10 μl of 5% ACN (loading buffer for HPLC).

Mapping of SAMPylation sites was performed as follows. *H. volcanii* (pJAM949, Flag–SAMP2) and (pJAM202c, vector alone) cells grown on complex medium (ATCC 974) to stationary phase were used for generation of cell lysate. Clarified lysate (1 ml) was bound to the anti-Flag agarose beads and eluted by glycine buffer or 1× Flag peptide as described earlier. Eluted protein samples were diluted into 40 mM ammonium bicarbonate (NH_4HCO_3), reduced with 10 mM DTT for 1 h at 56 °C, carboxy-amidomethylated with 55 mM iodoacetamide for 45 min in the dark, and digested with 3 μg of trypsin (Promega) in 40 mM NH_4HCO_3 overnight at 37 °C. After digestion, the peptides were acidified with trifluoroacetic acid (TFA) at a final concentration of 0.1% TFA. Desalting was performed with C18 spin columns (Vydac Silica C18, The Nest Group) and the resulting peptides were dried down in a Speed Vac and stored at −20 °C until analysed. The peptides were resuspended with 19.5 μl of mobile phase A (0.1% formic acid in water) and 0.5 μl of mobile phase B (80% acetonitrile, ACN, and 0.1% formic acid in water) and filtered with 0.2 μm filters (Nanosep, PALL). The sample was loaded off-line onto a nanospray tapered capillary column/emitter (360 × 75 × 15 μm, PicoFrit, New Objective) self-packed with C18 reverse-phase resin (10.5 cm, Waters) in a nitrogen pressure injection cell for 10 min at 1,000 p.s.i. (~5 μl load) and separated using a 160 min linear gradient of increasing mobile phase B at a flow rate of ~200 nl min^{-1} directly into the mass spectrometer. LC-MS/MS analysis was performed on a LTQ Orbitrap XL ETD mass spectrometer (ThermoFisher) equipped with a nanospray ion source. A full FTMS (Fourier transform mass spectrometry) spectrum at 30,000 resolution was collected at 250–2,000 *m*/*z* followed by six data-dependent MS/MS spectra in ITMS (ion trap mass spectrometry) of the most intense ion peaks following collision-induced dissociation (CID) (36% normalized collision energy). For the parent mass list method, five data-dependent MS/MS spectra from the full FTMS were activated in the most intense ion peaks from parent mass list following 36% CID. The parent mass width was set up ± 20.0 p.p.m. To obtain the parent mass list, the identified protein sequences were theoretically digested by trypsin allowing to one internal miscleavage. The masses of theoretical tryptic peptides were allowed for dynamic modifications with the masses of oxidized methionine, alkylated cysteine, and two glycines on lysine (15.9949, 57.0215 and 114.0429 Da) respectively and then calculated with up to quintuply charge states. The masses were selected between 250–2,000 *m*/*z* at each charge state for the parent mass list.

MS data analysis. SAMP-conjugate peptides were identified from the MS data using MASCOT algorithms[44] that searched a custom, non-redundant database based on the hypothetical proteome of translated open-reading frames from the *H. volcanii* genome (April 2007 version, http://archaea.ucsc.edu/). Probability-based MOWSE scores were estimated by comparison of search results against estimated random match population and are reported as ~10log10(p), where p is the absolute probability. Individual ion scores greater than 22 indicates identity or extensive homology ($P < 0.05$). Carbamidomethylation was used as a fixed modification due to sample preparation. Variable modifications that were searched included deamidation of asparagines and glutamine, oxidation (single and double) of methionine, glycine-glycine addition on lysine, thiocarboxylation of C termini, and pyro-glutamine of N-terminal glutamine.

Data generated for site mapping of SAMP2-protein conjugates was searched against the *H. volcanii* sequence database containing the common contaminants database using the TurboSequest algorithm (BioWorks 3.3.1 SP1, Thermo Fisher Scientific). Spectra with a threshold of 15 ions, a TIC of 10^3, and a mass range of $[MH]^+ = 500$–5,000 *m*/*z* were searched. The SEQUEST parameters were set to allow 30.0 p.p.m. of precursor ion mass tolerance and 0.5 Da of fragment ion tolerance with monoisotopic mass. Only fully tryptic peptides were allowed with up to three missed internal cleavage sites. Dynamic mass increases of 15.9949, 57.0215, and 114.0429 Da were allowed for oxidized methionine, alkylated cysteine, and two glycines on lysine residue respectively. Proteins identified by more than two peptides were only considered to be statistically significant at $\leqslant 1\%$ false discovery rate (FDR) using the ProteoIQ software package (BioInquire). The fragmentations of all peptides containing an internal Gly-Gly modified lysine residue were subjected to manual validation.

Protein sequences. All *H. volcanii* protein sequences described in this study are included with gene locus tag numbers as Supplementary Information. The following protein sequences were also described: ScUb (P61864); ScUrm1 (P40554); EcMoaD (CAA49864); EcThiS (O32583); ScUba4p (P38820); HsMOCS3 (O95396); EcMoeB (P12282); ScYor285W (Q12305); ScYor251c (Q08686); EcSseA (P31142); EcMoaE; (P30749); HsMOCS2B (O96007); BsMobB (O31704) (GenBank or Swiss-Prot accession numbers in parenthesis; Sc, *Saccharomyces cerevisiae*; Ec, *E. coli*; Bs, *Bacillus subtilis*; Hs, *Homo sapiens*).

43. Dyall-Smith, M. *The Halohandbook: Protocols for Halobacterial Genetics.* (2008).
44. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567 (1999).

Sub-luminous type Ia supernovae from the mergers of equal-mass white dwarfs with mass $\sim 0.9M_\odot$

Rüdiger Pakmor[1], Markus Kromer[1], Friedrich K. Röpke[1], Stuart A. Sim[1], Ashley J. Ruiter[1] & Wolfgang Hillebrandt[1]

Type Ia supernovae are thought to result from thermonuclear explosions of carbon–oxygen white dwarf stars[1]. Existing models[2] generally explain the observed properties, with the exception of the sub-luminous 1991bg-like supernovae[3]. It has long been suspected that the merger of two white dwarfs could give rise to a type Ia event[4,5], but hitherto simulations have failed to produce an explosion[6,7]. Here we report a simulation of the merger of two equal-mass white dwarfs that leads to a sub-luminous explosion, although at the expense of requiring a single common-envelope phase, and component masses of $\sim 0.9M_\odot$. The light curve is too broad, but the synthesized spectra, red colour and low expansion velocities are all close to what is observed for sub-luminous 1991bg-like events. Although the mass ratios can be slightly less than one and still produce a sub-luminous event, the masses have to be in the range $0.83M_\odot$ to $0.9M_\odot$.

Thermonuclear burning in normal type Ia supernovae produces $0.4M_\odot$ to $0.9M_\odot$ (ref. 8) of ^{56}Ni, the radioactive decay of which powers the optical display of type Ia supernovae and makes them amongst the most luminous objects in the Universe. This range of ^{56}Ni masses can be reproduced by explosion models of white dwarfs which have masses at the Chandrasekhar limit[2,9]. However, 1991bg-like supernovae are characterized[3] by a B-band peak magnitude of about -17 magnitudes (mag), implying ^{56}Ni masses of only about $0.1M_\odot$ (ref. 8). This subclass must be understood in order to capture theoretically the full range of type Ia supernova diversity. A recent analysis of the spectral evolution of SN 2005bl (ref. 10), which is representative of the 1991bg-like type Ia supernovae, showed that both iron-group elements and silicon are present over a wide range of radii extending as close to the centre of the ejecta as is accessible observationally. It is difficult to imagine an explosion model in which hydrodynamic processes alone account for as strong a mixing of silicon and iron-group elements as is observed in SN 2005bl. In normal type Ia supernovae the core of the ejecta consists predominantly of iron-group elements, surrounded by a silicon-rich layer. This difference in the characteristic chemical structures indicates a different burning regime realized in the 1991bg-like objects.

Incomplete silicon burning is a natural way of producing a mixture of iron-group elements and silicon. It occurs in low-density carbon–oxygen fuel for a narrow window of ash temperatures[11]. Burning significant amounts of stellar material in this regime therefore requires a shallow density profile of the exploding object. Moreover, supersonic propagation of the burning front is required to avoid pre-expansion of the material. The model described below satisfies both of these conditions and yields a small ^{56}Ni mass.

In contrast to previous merger simulations[6,7] that considered white dwarfs of significantly different masses, our model assumes equal-mass white dwarfs. Both have a central density of $1.4 \times 10^7\,\mathrm{g\,cm^{-3}}$, a mass of $0.89M_\odot$, a composition of equal parts by mass of carbon and oxygen and an initial temperature of 5×10^5 K. The initial orbit is circular with a period of 28 s. This corresponds to the state when tidal forces deform the white dwarfs sufficiently to make the system unstable (prior evolution is driven by gravitational wave emission and is not simulated). This progenitor system is set up as the initial condition for a simulation using a smoothed-particle-hydrodynamics code[12].

In the initial condition, marginal asymmetries were deliberately introduced, because perfect symmetry is not expected in nature. In this first simulation, we follow the inspiral and subsequent merger of the binary system (Fig. 1). After two orbits one of the two white dwarfs is disrupted. This unequal evolution of the white dwarfs originates from the symmetry-breaking in the initial conditions. The disrupted white dwarf violently merges with the remaining white dwarf and material is heated by compression. In the hottest regions carbon burning begins and releases additional energy, which further heats the material. A hotspot, which is resolved by several smoothed-particle-hydrodynamics particles, forms with a temperature of 2.9×10^9 K in high-density material ($3.8 \times 10^6\,\mathrm{g\,cm^{-3}}$). High-resolution small-scale simulations[13] show that under such conditions a detonation ignites.

In the second step of our simulation sequence, we impose the triggering of a detonation at the hottest point and follow it with a grid-based hydrodynamics code[14,15] as it crosses the merged object. The energy release from the nuclear burning in the detonation disrupts the system (see Fig. 1). The asymptotic kinetic energy of the ejecta is 1.3×10^{51} erg. This is comparable to typical explosion energies of standard type Ia supernova models[16] arising from Chandrasekhar-mass ($1.38M_\odot$) white dwarfs. However, because the total mass of the ejecta is about 1.3 times larger in our model, it has lower velocities on average[17].

Using tracer particles that record the conditions during the explosion phase and a 384-isotope nuclear network[18], we reconstruct the detailed nucleosynthesis of the explosion in the third step. Because of the low densities in the merged object, the nucleosynthesis primarily proceeds in the regime of incomplete silicon burning. Thus, only $0.1M_\odot$ of ^{56}Ni are synthesized and the ejecta consist mostly of intermediate-mass elements ($1.1M_\odot$) and oxygen ($0.5M_\odot$). Less than $0.1M_\odot$ of carbon is left unburned. (For detailed nucleosynthesis results see the Supplementary Information.) Thus, our model fulfils the requirements necessary to reproduce the characteristics of the 1991bg-like class of supernovae.

Finally, we use the structure of the ejecta and the detailed chemical abundances to calculate synthetic light curves and spectra using a Monte Carlo radiative transfer code[19] as required to test this model quantitatively against observations. Owing to the small ^{56}Ni mass synthesized during the nuclear burning, the synthetic light curves (Fig. 2) are faint and decline rapidly compared to those of normal type Ia supernovae, despite the large total ejecta mass of our simulation ($1.8M_\odot$). Given that there has been no fine-tuning of the explosion model, the light curves

[1]Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Strasse 1, 85748 Garching, Germany.

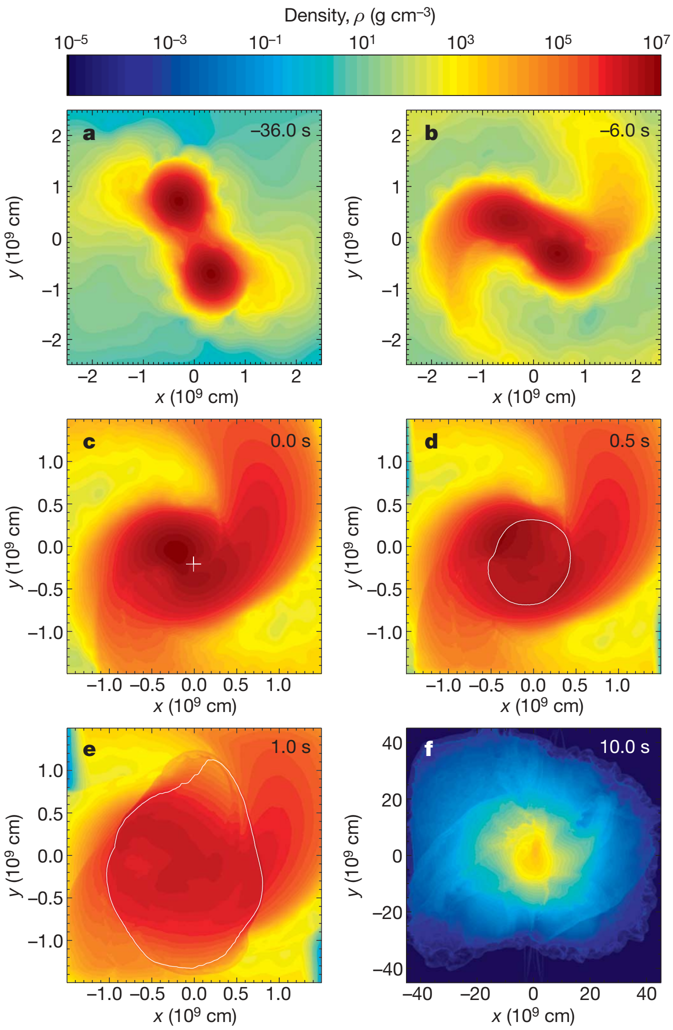


Figure 1 | Time evolution of the binary system. Slices in the $z = 0$ plane showing colour-coded density in logarithmic units. The panels illustrate the complete evolution. Starting from a nearly stable orbit (**a**), dynamic interactions lead to the disruption of one of the white dwarfs (**b**). In the resulting single object a detonation is triggered (**c**). The projected location where the detonation starts is marked with a cross. It then propagates through the object (**d**, **e**). The white contour shows the position of the detonation shock. Finally, the object becomes gravitationally unbound and reaches free expansion (**f**). The evolution before the detonation occurs is simulated using the smoothed-particle-hydrodynamics code GADGET2 (ref. 12) with two million particles. At the time the detonation starts (**c**), the current state of the simulation is mapped onto a uniform Cartesian 512^3 grid to follow the propagation of the detonation adequately[14,15]. Times are relative to the detonation. Further detailed information regarding the simulations and the codes can be found in the Supplementary Information.

agree remarkably well with those of the 1991bg-like type Ia supernovae, in both absolute magnitude and colour evolution. Moreover, our model naturally predicts the lack of secondary maxima in the near-infrared (J, H and K) light curves, which is a peculiarity of 1991bg-like objects compared to normal type Ia supernovae.

In detail, however, there are some discrepancies between our model light curves and the observational data. Comparing the difference in brightness at B band maximum and 15 days thereafter we find values between 1.4 and 1.7, depending on the line-of-sight. This is less than typically observed for 1991bg-like objects (1.9), but at worst it is similar to the fastest-declining normal type Ia supernovae and substantially faster than for objects which have previously been claimed as possible super-Chandrasekhar explosions (for example, 0.69 for SN 2006gz; ref. 20). We note that the exact light curve shapes are affected by details of both the nucleosynthesis and the radiative transfer and are thus very sensitive to any systematic shortcomings of the simulations. In particular, necessary approximations in the treatment of the ionization state of the ejecta can influence the decline of the light curves[19]. However, our simulation is only one particular realization of the model, so a closer agreement may be found by exploring the initial parameter space.

Figure 3 compares the spectrum of SN 2005bl (a well observed example of 1991bg-like objects) near maximum light to our angle-averaged model spectrum. This illustrates that both the overall flux distribution and the individual spectroscopic features agree remarkably well. Although the features show some variation for different lines of sight these are small and the angle-averaged spectrum is representative. For detailed points of comparison particularly relevant to 1991bg-like events, see the Supplementary Information.

The total mass of the system is essentially a free parameter, except that a mass ratio close to one is required. Our simulation predicts 1991bg-like events for white dwarf masses of about $0.9M_\odot$. Systems of somewhat heavier white dwarfs will synthesize considerably more ^{56}Ni owing to their higher densities and will therefore lead to much brighter explosions. Conversely, systems of significantly lower mass will not lead to events classified as type Ia supernovae. First, they probably fail to trigger detonations because the hot spots occur at densities that are too low. Moreover, even were a detonation to form, the density of the white dwarf material is too low to synthesize any ^{56}Ni.

A mass ratio of exactly one is artificial. To be realized in nature, our model must also permit somewhat smaller values. To verify this, we conducted three additional smoothed-particle-hydrodynamics simulations of the merger phase for binary systems with a primary white dwarf mass of $M_1 = 0.89M_\odot$ and secondary masses of $M_2 = 0.87M_\odot$, $0.85M_\odot$ and $0.83M_\odot$, respectively. All of these systems show evolution similar to our complete simulation. The secondary white dwarf is disrupted and merges violently with the primary and all simulations ignite carbon burning. Although the hotspots reach temperatures above 2.5×10^9 K, the most important difference is that the associated densities are slightly lower. The lowest value (around 3×10^6 g cm^{-3}) is found in the simulation with the least massive secondary. According to the detonation criteria[18], this is still sufficient to ignite a detonation. It is clear, however, that for substantially smaller mass ratios (M_2/M_1), the ignition of a detonation will fail because the densities at the hotspots are insufficiently high.

Population synthesis studies[21] predict that a range of total masses and mass ratios for merging white dwarf binaries occur naturally. For our model of violent mergers with mass ratios close to one, systems with total mass similar to that in our simulation are the most relevant case for type Ia supernovae. Although lower-mass systems will be more common, they will not lead to a detonation, as discussed above. Thus they will not result in type Ia supernovae, but may be observable as faint short-lived transients. In contrast, more massive mergers will be brighter but much rarer owing to the paucity of high-mass white dwarfs. We used the results of recent population synthesis studies[21] computed with the StarTrack code[22,23] to predict the expected rate of binary mergers suitable for our model relative to those of other possible type Ia supernova progenitor formation channels. We find that events from our model should occur with a rate of approximately 2–11% of the total type Ia supernova rate. This fraction is consistent with the observed rate of 1991bg-like supernovae.

Observations indicate that sub-luminous type Ia supernovae are expected to arise in old (more than a gigayear) stellar populations[24]. It may seem counterintuitive that the relatively massive (zero-age main-sequence masses of about $4M_\odot$–$6M_\odot$) binaries considered here would produce type Ia supernovae with long delay times, but it is possible provided they undergo only one common-envelope phase and/or begin their evolution on the zero-age main sequence with wide orbital separations. In such situations, more luminous supernovae would arise from white dwarf mergers originating from

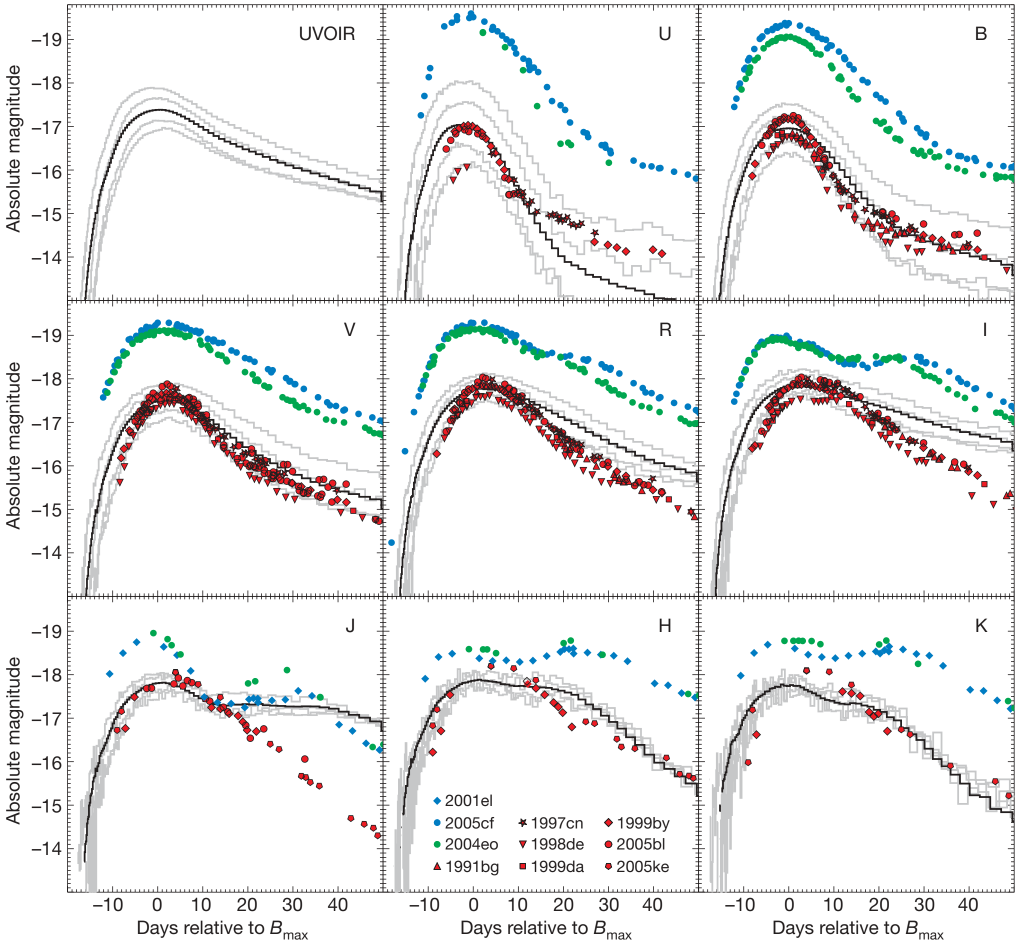


Figure 2 | Synthetic light curves of our model. From top left to bottom right the histograms show ultraviolet–optical–infrared bolometric (UVOIR) and broad-band U,B,V,R,I,J,H and K synthetic light curves of our model, which were obtained using the multi-dimensional Monte Carlo radiative transfer code ARTIS[19] after remapping the explosion ejecta to a 50^3 Cartesian grid. The black histograms show angle-averaged light curves. To indicate the scatter in brightness caused by the model asymmetries, we overplotted four line-of-sight specific light curves (grey histograms). These have been selected from 100 equally sized solid-angle bins such that they represent the full range of the scatter. Time is given relative to B-band maximum, B_{max}. The small-scale fluctuations in the histograms are due to Monte Carlo noise in the simulation, which is largest in the near-infrared bands and at late times in the U band. The region populated by the different lines of sight agrees remarkably well with that populated by the sample of observed 1991bg-like type Ia supernovae shown as red symbols (observations from ref. 25, and references therein). For SN 1999by, polarization measurements[26] revealed a ~20% degree of asphericity, assuming that the object was observed equator-on. Currently, our radiative transfer simulations do not include polarization, but by extracting the shapes of the surfaces of photon last-scattering we find an upper limit on the asphericity of around 40%. This is consistent with the value obtained for SN 1999by. For comparison, the light curves of normal type Ia supernovae (SN 2005cf[27] as blue circles and SN 2001el[28] as blue diamonds) are also shown. They are much brighter, show a slower decline after maximum light, and distinct secondary maxima in R and redder bands. The same is true for SN 2004eo[29] (shown as green circles) which represents the faint end of normal type Ia supernovae.

more massive progenitors (zero-age main sequence masses of about $6M_\odot$–$8M_\odot$), or from the single degenerate channel.

The use of type Ia supernovae to measure the expansion history of the Universe relies on their homogeneity. The possibility of physically different evolutionary paths leading to type Ia supernovae has therefore been a concern in such studies. We have shown that violent mergers of two massive white dwarfs, even when their total mass exceeds the Chandrasekhar limit, predominantly produce faint events. Therefore, they are not expected to pollute samples of normal type Ia supernovae significantly and can be excluded as a source of systematic errors in cosmological distance measurements.

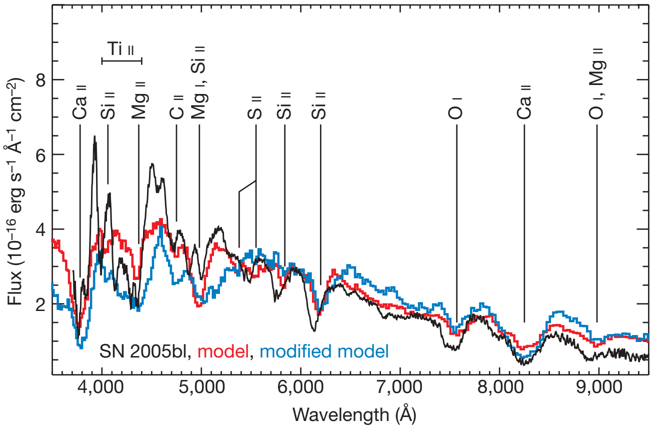


Figure 3 | Synthetic spectrum of our model. Comparison between SN 2005bl (observations from ref. 25) three days before the B-band maximum (black line) and an angle-averaged synthetic spectrum of our model at the corresponding epoch (red histogram). The strongest features in the model spectrum are labelled. The nucleosynthesis yields in the regime of incomplete silicon burning depend strongly on the fuel composition, so the detailed spectral features in our model are highly sensitive to the composition of the white dwarfs. This is illustrated by the blue histogram, which shows the spectrum obtained for a model adopting different initial composition (see Supplementary Information for details). The effect is strongest in the blue-wavelength range, particularly for the Ti II absorption trough between 4,000 Å and 4,400 Å.

Received 22 July; accepted 3 November 2009.

1. Hoyle, F. & Fowler, W. A. Nucleosynthesis in supernovae. *Astrophys. J.* **132,** 565–590 (1960).
2. Kasen, D., Roepke, F. & Woosley, S. E. The diversity of type Ia supernovae from broken symmetries. *Nature* **460,** 869–872 (2009).
3. Leibundgut, B. *et al.* SN 1991bg–A type Ia supernova with a difference. *Astrophys. J.* **105,** 301–313 (1993).
4. Iben, I. Jr & Tutukov, A. V. Supernovae of type I as end products of the evolution of binaries with components of moderate initial mass (M not greater than about 9 solar masses). *Astrophys. J. Suppl. Ser.* **54,** 335–372 (1984).
5. Webbink, R. F. Double white dwarfs as progenitors of R Coronae Borealis stars and type I supernovae. *Astrophys. J.* **277,** 355–360 (1984).
6. Saio, H. & Nomoto, K. Evolution of a merging pair of C+O white dwarfs to form a single neutron star. *Astron. Astrophys.* **150,** L21–L23 (1985).
7. Benz, W., Cameron, A. G. W., Press, W. H. & Bowers, R. L. Dynamic mass exchange in doubly degenerate binaries. I. 0.9 and 1.2 solar mass stars. *Astrophys. J.* **348,** 647–667 (1990).
8. Stritzinger, M., Leibundgut, B., Walch, S. & Contardo, G. Constraints on the progenitor systems of type Ia supernovae. *Astron. Astrophys.* **450,** 241–251 (2006).
9. Mazzali, P. A., Röpke, F. K., Benetti, S. & Hillebrandt, W. A common explosion mechanism for type Ia supernovae. *Science* **315,** 825–828 (2007).
10. Hachinger, S., Mazzali, P. A., Taubenberger, S., Pakmor, R. & Hillebrandt, W. Spectral analysis of the 91bg-like type Ia SN 2005bl: low luminosity, low velocities, incomplete burning. *Mon. Not. R. Astron. Soc.* **399,** 1238–1254 (2009).
11. Thielemann, F.-K., Nomoto, K. & Yokoi, K. Explosive nucleosynthesis in carbon deflagration models of type I supernovae. *Astron. Astrophys.* **158,** 17–33 (1986).
12. Springel, V. The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* **364,** 1105–1134 (2005).
13. Seitenzahl, I. R. *et al.* Spontaneous initiation of detonations in white dwarf environments: determination of critical sizes. *Astrophys. J.* **696,** 515–527 (2009).
14. Röpke, F. K. & Niemeyer, J. C. Delayed detonations in full-star models of type Ia supernova explosions. *Astron. Astrophys.* **464,** 683–686 (2007).
15. Fink, M., Hillebrandt, W. & Röpke, F. K. Double-detonation supernovae of sub-Chandrasekhar mass white dwarfs. *Astron. Astrophys.* **476,** 1133–1143 (2007).
16. Nomoto, K., Thielemann, F.-K. & Yokoi, K. Accreting white dwarf models of type I supernovae. III—carbon deflagration supernovae. *Astrophys. J.* **286,** 644–658 (1984).
17. Howell, D. A. *et al.* The type Ia supernova SNLS-03D3bb from a super-Chandrasekhar-mass white dwarf star. *Nature* **443,** 308–311 (2006).
18. Travaglio, C., Hillebrandt, W., Reinecke, M. & Thielemann, F.-K. Nucleosynthesis in multi-dimensional SN Ia explosions. *Astron. Astrophys.* **425,** 1029–1040 (2004).
19. Kromer, M. & Sim, S. A. Time-dependent three-dimensional spectrum synthesis for type Ia supernovae. *Mon. Not. R. Astron. Soc.* **398,** 1809–1826 (2009).
20. Hicken, M. *et al.* The luminous and carbon-rich supernova 2006gz: a double degenerate merger? *Astrophys. J.* **669,** L17–L20 (2007).
21. Ruiter, A. J., Belczynski, K. & Fryer, C. Rates and delay times of type Ia supernovae. *Astrophys. J.* **699,** 2026–2036 (2009).
22. Belczynski, K., Kalogera, V. & Bulik, T. A comprehensive study of binary compact objects as gravitational wave sources: evolutionary channels, rates, and physical properties. *Astrophys. J.* **572,** 407–431 (2002).
23. Belczynski, K. *et al.* Compact object modeling with the startrack population synthesis code. *Astrophys. J. Suppl. Ser.* **174,** 223–260 (2008).
24. Sullivan, M. *et al.* Rates and properties of type Ia supernovae as a function of mass and star formation in their host galaxies. *Astrophys. J.* **648,** 868–883 (2006).
25. Taubenberger, S, *et al.* The underluminous type Ia supernova 2005bl and the class of objects similar to SN 1991bg. *Mon. Not. R. Astron. Soc.* **385,** 75–96 (2008).
26. Howell, D. A., Höflich, P., Wang, L. & Wheeler, J. C. Evidence for asphericity in a subluminous type Ia supernova: spectropolarimetry of SN 1999by. *Astrophys. J.* **556,** 302–321 (2001).
27. Pastorello, A. *et al.* ESC observations of SN 2005cf—I. Photometric evolution of a normal type Ia supernova. *Mon. Not. R. Astron. Soc.* **376,** 1301–1316 (2007).
28. Krisciunas, K. *et al.* Optical and infrared photometry of the nearby type Ia supernova 2001el. *Astron. J.* **125,** 166–180 (2003).
29. Pastorello, A. *et al.* ESC amd KAIT observations of the transitional type Ia SN 2004eo. *Mon. Not. R. Astron. Soc.* **377,** 1531–1552 (2007).

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

Acknowledgements We thank K. Belczynski for supporting the population synthesis analysis and S. Taubenberger and S. Hachinger for discussions. This work was supported by the Transregional Research Centre ''The Dark Universe'' of the German Research Foundation and by the Excellence Cluster ''Origin and Structure of the Universe''. The work of F.K.R. is supported through the Emmy Noether Program of the German Research Foundation. We are grateful for computer time provided by the Computer Centre of the Max Planck Society in Garching, where all simulations were performed.

Author Contributions R.P. carried out the hydrodynamical simulations and the nucleosynthesis postprocessing. M.K. performed the radiative transfer simulations. F.K.R. and R.P. worked on development of the hydrocodes. S.A.S. and M.K. developed the radiative transfer code. W.H. and F.K.R. started the project. A.J.R. analysed the results of population synthesis calculations. All authors contributed to the interpretation of the simulations and to the writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.P. (rpakmor@mpa-garching.mpg.de).

A lower limit of 50 microgauss for the magnetic field near the Galactic Centre

Roland M. Crocker[1,2], David I. Jones[2,3,4], Fulvio Melia[5], Jürgen Ott[6,7] & Raymond J. Protheroe[3]

The amplitude of the magnetic field near the Galactic Centre has been uncertain by two orders of magnitude for several decades. On a scale of ~100 parsecs (pc), fields of ~1,000 microgauss (μG; refs 1–3) have been reported, implying a magnetic energy density more than 10,000 times stronger than typical for the Galaxy. Alternatively, the assumption of pressure equilibrium between the various phases of the Galactic Centre interstellar medium (including turbulent molecular gas, the contested[4] 'very hot' plasma, and the magnetic field) suggests fields of ~100 μG over ~400 pc size scales[5]. Finally, assuming equipartition, fields of only ~6 μG have been inferred from radio observations[6] for 400 pc scales. Here we report a compilation of previous data that reveals a downward break in the region's non-thermal radio spectrum (attributable to a transition from bremsstrahlung to synchrotron cooling of the *in situ* cosmic-ray electron population). We show that the spectral break requires that the Galactic Centre field be at least ~50 μG on 400 pc scales, lest the synchrotron-emitting electrons produce too much γ-ray emission, given other existing constraints[7]. Other considerations support a field of 100 μG, implying that over 10% of the Galaxy's magnetic energy is contained in only $\lesssim$0.05% of its volume.

Pinning down the large-scale magnetic field of the Galactic Centre—which provides our closest view of a galactic nucleus—within the uncertainty (greater than two orders of magnitude) implied by existing, rival analyses has long been of interest. A 1,000 μG field would have substantial ramifications for the region's dynamics, including limiting the diffusion distances of relativistic particles (thereby excluding scenarios where the diffuse, ~TeV γ-ray glow of the Galactic Centre[8] is ultimately explained as due to a single, astrophysical accelerator[9]). It would also generate a large enough magnetic drag to enhance the spiralling-in rate of giant molecular gas clouds towards the Galactic Centre[3], limiting the lifetime of these to ~100 million years, implying 'starbursts' on the same timescale. On the other hand, the formation of individual stars might be inhibited (or the stellar initial mass function biased) by the tendency of a strong magnetic field to support gas against gravitational collapse.

Radio observations at 74 MHz and 330 MHz (ref. 6) reveal a diffuse (but distinct) region of non-thermal radio emission covering the Galactic Centre (out to about ±3° or about ±420 pc from the Galactic Centre along the Galactic plane). Invoking the 'equipartition' condition (minimizing the total energy in magnetic field and relativistic electrons), a field of only 6 μG is inferred[6], typical for the Galaxy at large, climbing to 11 μG over the inner ~±0.8° (for extremal parameter values, the field might reach 100 μG).

To probe this radio structure at higher frequencies, we have assembled total-intensity, single-dish flux density data at 1.4, 2.4, 2.7 and 10 GHz (refs 10–13). These data are polluted by line-of-sight synchrotron emission in the Galactic plane (both behind and in front of the Galactic Centre) and the flux density contributed by discrete sources; we used a combination of low-pass (spatial wavenumber) filtering to remove the latter and all-sky Galactic synchrotron background observations to remove the former (see Supplementary Information). After this processing, a distinct, non-thermal, radio structure is revealed in all radio maps up to 10 GHz (Fig. 1). However, the structure's spectrum from 74 MHz to 10 GHz is not described by a pure power law: attempting to fit such a power law to the cleaned data we find a minimum χ^2 of 4.9 per degree of freedom (d.f. = 4), excluded at a confidence level of 3.4σ (see below).

In fact, fitting separate power laws to the background-subtracted lower three (1.4–2.7 GHz) and upper three (2.4–10 GHz) radio data, we find that these extrapolations intersect at ~1.7 GHz with a downbreak of ~0.6 in spectral index. This is close to the canonical break of 1/2 produced by a steady-state synchrotron-radiating electron population that transitions (with increasing energy) from bremsstrahlung-cooled to synchrotron-cooled. As the synchrotron cooling rate is a function of magnetic field, B, and bremsstrahlung a function of ambient hydrogen number density, n_H, the break frequency is a function of both parameters. Observation of a spectral break then determines acceptable pairs of B and n_H for the environment of the synchrotron-emitting electrons.

We have modelled the cooled electron distribution and resulting synchrotron emission as a function of B, n_H and spectral index at injection of the electron population. We have accounted for losses due to inverse Compton emission following collisions with ambient light[14], ionization, bremsstrahlung and synchrotron emission.

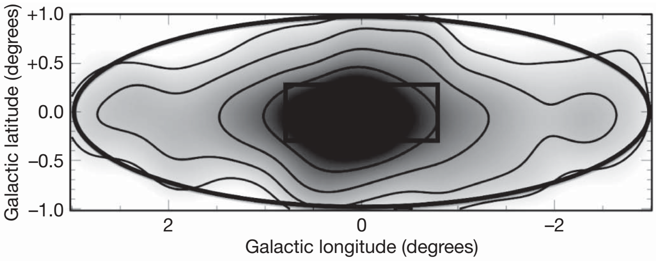


Figure 1 | Total intensity image of the region at 10 GHz. Radio map[13] convolved to a resolution of 1.2° × 1.2° with contours at 10, 20, 40, 80, 160 and 240 Jy per beam. (Native resolution and convolved images at $\nu \geq 1.4$ GHz are available in the Supplementary Information.) There is a striking constancy in the appearance of the radio structure from 74 MHz to at least 10 GHz (the large ellipse traces the diffuse, non-thermal radio emission region first identified at 74 and 330 MHz; ref. 6). The small rectangle delineates the region from which the HESS collaboration determines a diffuse ~TeV γ-ray intensity[8].

[1] School of Physics, Monash University, Victoria 3110, Australia. [2] Max-Planck-Institut für Kernphysik, PO Box 103980, Heidelberg, Germany. [3] Department of Physics, School of Physics and Chemistry, University of Adelaide, North Terrace, South Australia 5005, Australia. [4] Australia Telescope National Facility, Marsfield, New South Wales 2122, Australia. [5] Physics Department, The Applied Math Program, and Steward Observatory, The University of Arizona, Tucson, Arizona 85721, USA. [6] National Radio Astronomical Observatory, Charlottesville, PO Box O, 1003 Lopezville Road, Socorro, New Mexico 87801-0387, USA. [7] California Institute of Technology, 1200 E. California Blvd, Caltech Astronomy, 105-24, Pasadena, California 91125, USA.

The results of our fitting procedure are displayed in Figs 2 and 3. A very important constraint is proffered by γ-ray data covering the non-thermal emission region: bremsstrahlung and inverse Compton emission will inescapably be generated by the electrons responsible for the observed radio emission (the energies of electrons synchrotron-radiating at ~GHz and bremsstrahlung-radiating at $\gtrsim$100 MeV are very similar). This must not surpass the 300 MeV γ-ray flux from the region measured by EGRET[7] (Fig. 2b). This consideration rules out field amplitudes $\lesssim$50 μG.

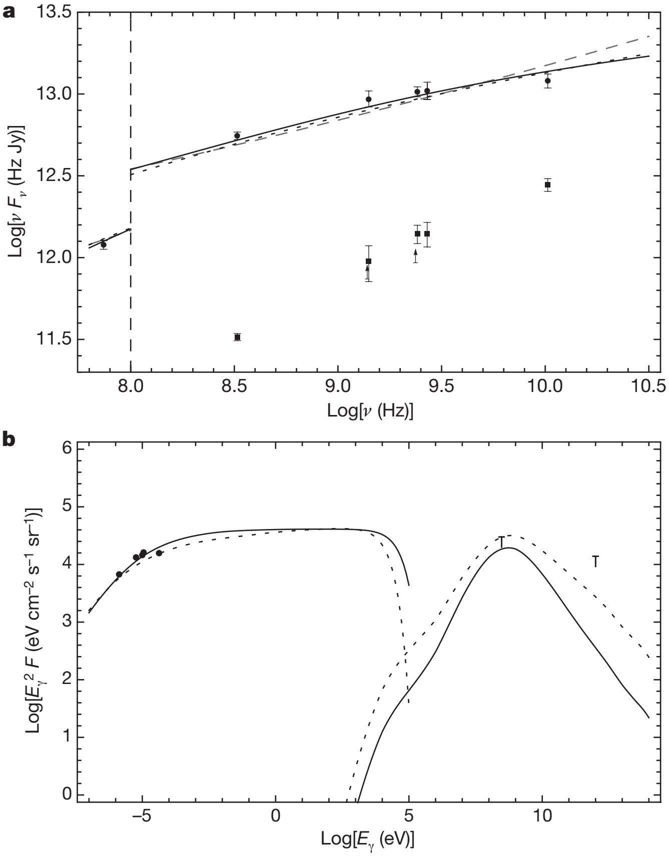


Figure 2 | Spectrum of the region: data and models. a, Radio spectrum; **b**, broadband spectrum. **a**, Circles, flux density of the radio structure after removal of flux from sources with sizes <1.2°. The vertical line divides the 74 MHz datum (that receives no contribution from Galactic plane synchrotron background/foreground) from all the other data (that do)—hence the discontinuity in the model curves at 100 MHz. Squares, discrete source flux densities (measured directly by the VLA at 330 MHz (ref. 6) and otherwise obtained through the Fourier technique described in Supplementary Information). Arrows, discrete source lower limits obtained from ATCA. Error bars, 68% confidence intervals. Curves, the best-fit flux density, F_ν, due to synchrotron emission from a cooled electron distribution plus Galactic plane synchrotron for the case of 100 μG fields (solid line) and 30 μG fields (dotted line; the break in the radio synchrotron curve mirrors a corresponding break in the distribution of radiating electrons; we assume the electrons are injected with a power law in momentum[22]). Dashed line, the best-fit null hypothesis case of a pure power-law signal plus Galactic plane synchrotron. **b**, Models for flux, F, as a function of photon energy, E_γ, for a magnetic field of 100 μG (solid lines) and 30 μG (dashed lines). The injected electron spectrum is assumed to follow a power law up to ~100TeV. (Note that the cut-off energy is only weakly constrained; our conclusions are not sensitive to the exact cut-off energy, however, given that radio observations guarantee the existence of ~GeV electrons whose bremsstrahlung emission we can compare against the EGRET intensity datum.) The lower-energy curves are synchrotron emission. The higher are bremsstrahlung plus inverse Compton emission from the same electrons responsible for the synchrotron (a figure showing the individual bremsstrahlung and inverse Compton contributions is shown in Supplementary Information). The upper limits are due to observations by EGRET[7] at 300 MeV and HESS[8] at 1 TeV.

In the context of diffuse, ~GeV emission around the Galactic Centre, we await results from the Fermi Gamma-ray Space Telescope[15], whose sensitivity (more than an order of magnitude better than EGRET's) promises, at worst, an increased lower limit to the large-scale magnetic field or, more optimistically, a measurement of this field (in concert with radio observations). For now, we emphasise that it is a novel analysis—not new data—that has allowed the new constraint on the Galactic Centre magnetic field.

Figure 4 shows the energy densities of various Galactic Centre interstellar medium phases. For acceptable field amplitudes, the cosmic-ray electron population is considerably sub-equipartition with respect to the other Galactic Centre interstellar medium phases (even after accounting for filling factor effects), in particular, the magnetic field, explaining why the magnetic field estimate arrived at assuming equipartition[6] is too low. We have also calculated the maximum possible energy density in the Galactic Centre cosmic-ray proton population, given the EGRET γ-ray constraints. Note that at ~100 μG the magnetic field reaches equipartition (at ~300 eV cm^{-3}) with the putative 'very hot' (~8 keV) phase of the X-ray-emitting plasma supposedly detected throughout the central few degrees along the Galactic plane[16]. Moreover, the energy density of the gas turbulence kinetic energy for the derived n_H is within a factor of a few of equipartition with these other phases up to magnetic field amplitudes of ~100 μG.

This situation—implying near pressure equilibrium between a number of Galactic Centre interstellar medium phases (including the plasma)—mirrors that of the Galactic disk, albeit at much higher pressure. Such considerations led to the prediction[5] that the real amplitude of the Galactic Centre's magnetic field lies close to ~100 μG. Very recently, however, observations have shown[4] that the X-ray emission from a region around Galactic longitude $l = 0.08°$ and latitude $b = 1.42°$ (taken to be typical of the so-called X-ray Ridge) is due to unresolved point sources, implying that the very hot X-ray plasma is illusory. This casts the pressure equilibrium argument[5] into some doubt. Note, however, that our magnetic field lower limit holds irrespective of whether the 8 keV plasma is real or not.

There are a number of intriguing parallels between the situation described above and the situation apparently pertaining within the inner regions of starburst galaxies. These independently support the idea that the field is in fact close to 100 μG. In starburst environments, it is contended[17], equipartition magnetic field values obtained from radio observations significantly underestimate the real field. Fields are actually sufficiently high to be in hydrostatic equilibrium with the self-gravity

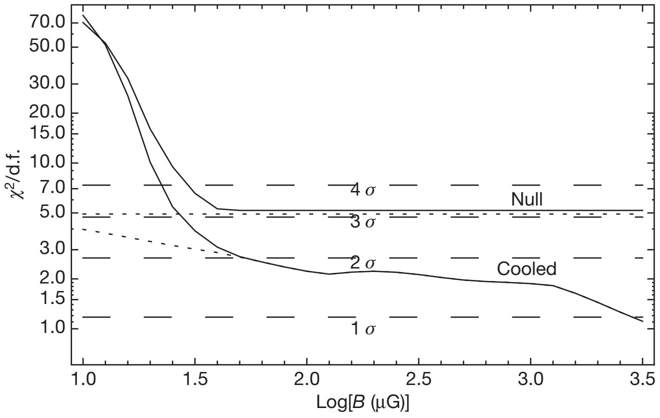


Figure 3 | Plot of the χ^2 per degree of freedom as a function of magnetic field amplitude. Curves, model for a cooled primary electron model (with 3 degrees of freedom, d.f.; 'cooled') and the null hypothesis of a pure power-law electron distribution (with 4 d.f.; 'Null'). The solid curves are constrained by the requirement that the γ-ray emission from the synchrotron-radiating electron population be less than the upper limit obtained from EGRET data[7] (dotted curves are not so constrained). The horizontal dashed lines mark the 1,2,3,4σ confidence limits for a model with 3 d.f. and do not apply to the null hypothesis (which only achieves a best fit acceptable at the ~3.4σ level).

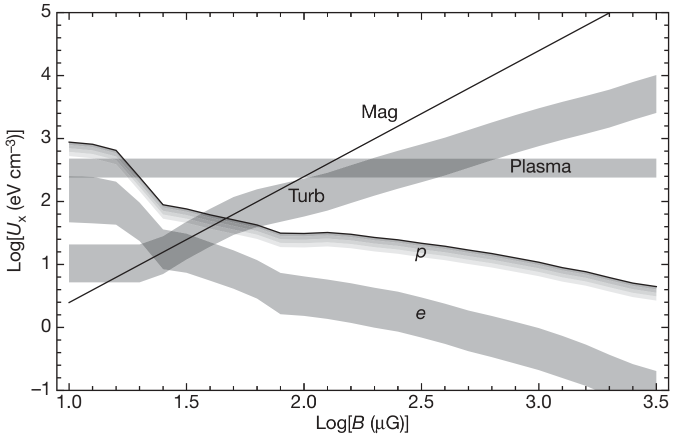


Figure 4 | Energy density, U_{X}, in phase 'X' of the Galactic Centre interstellar medium as a function of magnetic field. Bands show 68% confidence limits (except for p). Mag, magnetic field. Plasma, the disputed[4] X-ray emitting 8 keV plasma (with a density[23] 0.03–0.06 cm^{-3}). Turb, turbulent motions of the local gas (assuming it is at the best-fit n_{H} and has a velocity dispersion in the range 15–30 km s^{-1} typical for Galactic Centre molecular clouds[24]; note that, as the n_{H} required to fit the radio spectrum increases with B, the turbulent energy density is an increasing function of B). e, Cosmic ray electrons. p, a conservative upper bound on the cosmic ray proton energy density inferred from the 300 MeV γ-ray upper limit (a putative proton population, colliding with ambient gas at the best-fit n_{H}, would produce γ-rays mostly via neutral meson decay). The electron and proton energy densities take into account the filling factor of gas at the given value of n_{H} (ref. 25; we take the filling factor of gas within 1σ around the best-fit n_{H} value); this is a correction upwards to the values of U_e and U_p by an amount 100–1,000. Without this correction, one would have $U_e \approx U_{mag}$ at ~10 μG in agreement with the estimate derived assuming equipartion[6]. A factor of 2 uncertainty[26] in the hydrogen mass of the region is also accounted for in determining the electron and proton energy densities.

of the gaseous disk of the starburst. Such strong fields (together with high gas densities) guarantee that the relativistic particle population dumps all its energy before being transported out of the system. This calorimetric limit[18] may explain[17] why even very luminous starbursting galaxies fall on the far-infrared/radio-continuum correlation[19]. Circumstantial evidence also suggests that radio emission from starbursts is dominated by secondary electrons created in collisions between cosmic-ray ions and gas[17] (rather than directly accelerated electrons).

It is noteworthy, then, that the 60 μm and 1.4 GHz emission from the radio emission region place it within scatter of the far-infrared/radio-continuum correlation[19]. Furthermore, in the Galactic Centre a field of ~100 μG is precisely in the range required to establish hydrostatic equilibrium (given the total and gaseous surface densities). A situation wherein a magnetic field provides significant pressure support against gravity may lead to the development of the Parker instability[20], and exactly this is suggested by millimetre-wave observations[21] of molecular filaments of several hundred parsec length within ~1 kpc of the Galactic Centre. These observations independently suggest a ~100 μG field. Finally, the diffuse, ~TeV γ-ray glow from the vicinity of the Galactic Centre[8] is most probably explained by cosmic-ray impacts with gas. Unavoidably, such collisions would also produce copious secondary electrons, which could then contribute significantly to the region's synchrotron radio emission (for fields ≳300 μG, 100% of the radio emission from the rectangular region shown in Fig. 1 could be attributed to secondary electrons). Taken altogether, these facts paint a picture of the Galactic Centre as akin to a weak starburst with a magnetic field of ~100 μG.

Received 25 November 2008; accepted 30 October 2009.

1. Yusef-Zadeh, F. & Morris, M. G0.18–0.04 — Interaction of thermal and nonthermal radio structures in the arc near the galactic center. *Astron. J.* 94, 1178–1184 (1987).
2. Morris, M. & Yusef-Zadeh, F. The thermal, arched filaments of the radio arc near the Galactic center — magnetohydrodynamic-induced ionization? *Astrophys. J.* 343, 703–712 (1989).
3. Morris, M. The Galactic centre magnetosphere. Preprint at ⟨http://arXiv.org/abs/astro-ph/0701050⟩ (2007).
4. Revnivtsev, M. *et al.* Discrete sources as the origin of the Galactic X-ray ridge emission. *Nature* 458, 1142–1144 (2009).
5. Spergel, D. N. & Blitz, L. Extreme gas pressures in the Galactic bulge. *Nature* 357, 665–667 (1992).
6. LaRosa, T. N. *et al.* Evidence of a weak Galactic Centre magnetic field from diffuse low-frequency nonthermal radio emission. *Astrophys. J.* 626, L23–L27 (2005).
7. Hunter, S. D. *et al.* EGRET observations of the diffuse gamma-ray emission from the Galactic plane. *Astrophys. J.* 481, 205–240 (1997).
8. F. A, *et al.* Discovery of very-high-energy γ-rays from the Galactic Centre ridge. *Nature* 439, 695–698 (2006).
9. Wommer, E., Melia, F. & Fatuzzo, M. Diffuse TeV emission at the Galactic Centre. *Mon. Not. R. Astron. Soc.* 387, 987–997 (2008).
10. Reich, W., Reich, P. & Fuerst, E. The Effelsberg 21 cm radio continuum survey of the Galactic plane between L = 357 deg and L = 95.5 deg. *Astron. Astrophys.* 83 (Suppl.), 539–568 (1990).
11. Reich, W., Fuerst, E., Steffen, P., Reif, K. & Haslam, C. G. T. A radio continuum survey of the Galactic Plane at 11 cm wavelength. I — The area L = 357.4 to 76 deg, B = −1.5 to +1.5 deg. *Astron. Astrophys.* 58 (Suppl.), 197–248 (1984).
12. Duncan, A. R. *et al.* A deep radio continuum survey of the southern Galactic plane at 2.4 GHz. *Mon. Not. R. Astron. Soc.* 277, 36–52 (1995).
13. Handa, T. *et al.* A radio continuum survey of the Galactic plane at 10 GHz. *Proc. Astron. Soc. Jpn* 39, 709–753 (1987).
14. Porter, T. A., Moskalenko, I. V. & Strong, A. W. Inverse Compton emission from galactic supernova remnants: effect of the interstellar radiation field. *Astrophys. J.* 648, L29–L32 (2006).
15. Atwood, W. B. *et al.* The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission. *Astrophys. J.* 697, 1071–1102 (2009).
16. Koyama, K., Awaki, H., Kunieda, H., Takano, S. & Tawara, Y. Intense 6.7-keV iron line emission from the Galactic Centre. *Nature* 339, 603–605 (1989).
17. Thompson, T. A., Quataert, E., Waxman, E., Murray, N. & Martin, C. L. Magnetic fields in starburst galaxies and the origin of the FIR-radio correlation. *Astrophys. J.* 645, 186–198 (2006).
18. Voelk, H. J. The correlation between radio and far infrared emission for disk galaxies: a calorimeter theory. *Astron. Astrophys.* 218, 67–70 (1989).
19. Yun, M. S., Reddy, N. A. & Condon, J. J. Radio properties of infrared-selected galaxies in the IRAS 2 Jy sample. *Astrophys. J.* 554, 803–822 (2001).
20. Parker, E. N. The dynamical state of the interstellar gas and field. *Astrophys. J.* 145, 811–833 (1966).
21. Fukui, Y. *et al.* Molecular loops in the Galactic Center: evidence for magnetic flotation. *Science* 314, 106–109 (2006).
22. Crocker, R. M. *et al.* The cosmic ray distribution in Sagittarius B. *Astrophys. J.* 666, 934–948 (2007).
23. Yamauchi, S. *et al.* Optically thin hot plasma near the Galactic center — Mapping observations of the 6.7 keV iron line. *Astrophys. J.* 365, 532–538 (1990).
24. Güsten, R. & Philipp, S. D. in *Proc. 4th Cologne-Bonn-Zermatt Symp.* (eds Pfalzner, S., Kramer, C., Staubmeier, C. & Heithausen, A.) 253–263 (Springer Proceedings in Physics, Vol. 91, Springer, 2004).
25. Paglione, T. A. D., Jackson, J. M., Bolatto, A. D. & Heyer, M. H. Interpreting the HCN/CO intensity ratio in the Galactic Centre. *Astrophys. J.* 493, 680–693 (1998).
26. Ferrière, K., Gillard, W. & Jean, P. Spatial distribution of interstellar gas in the innermost 3 kpc of our galaxy. *Astron. Astrophys.* 467, 611–627 (2007).

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

Acknowledgements R.M.C. thanks T. Porter for conversations about the Galactic Centre interstellar radiation field. D.I.J. thanks Monash University for hospitality. R.M.C. and D.I.J. thank J. Dickey for advice about radio data analysis. J.O. is a Jansky Fellow at NRAO; R.M.C. is a J. L. William Fellow at Monash University and a Marie Curie Fellow at Max-Planck-Institut für Kernphysik.

Author Contributions R.M.C. led the work and performed the main analysis. D.I.J. performed the analysis of radio data, including development of the Fourier-based technique for background and foreground removal, was responsible for original radio observations, and provided critical scientific discussion. F.M. provided input on theoretical and statistical problems, and critical discussion of scientific interpretation. J.O. supervised the analysis of archival radio data and the taking of original radio data, and provided input on statistics. R.J.P. provided input on thermal and relevant non-thermal processes and critical discussion of scientific interpretation. R.J.P. and R.M.C. provided supervision of D.I.J. as doctoral candidate. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.M.C. (Roland.Crocker@mpi-hd.mpg.de).

LETTERS

Quantum simulation of the Dirac equation

R. Gerritsma[1,2], G. Kirchmair[1,2], F. Zähringer[1,2], E. Solano[3,4], R. Blatt[1,2] & C. F. Roos[1,2]

The Dirac equation[1] successfully merges quantum mechanics with special relativity. It provides a natural description of the electron spin, predicts the existence of antimatter[2] and is able to reproduce accurately the spectrum of the hydrogen atom. The realm of the Dirac equation—relativistic quantum mechanics—is considered to be the natural transition to quantum field theory. However, the Dirac equation also predicts some peculiar effects, such as Klein's paradox[3] and 'Zitterbewegung', an unexpected quivering motion of a free relativistic quantum particle[4]. These and other predicted phenomena are key fundamental examples for understanding relativistic quantum effects, but are difficult to observe in real particles. In recent years, there has been increased interest in simulations of relativistic quantum effects using different physical set-ups[5–11], in which parameter tunability allows access to different physical regimes. Here we perform a proof-of-principle quantum simulation of the one-dimensional Dirac equation using a single trapped ion[7] set to behave as a free relativistic quantum particle. We measure the particle position as a function of time and study Zitterbewegung for different initial superpositions of positive- and negative-energy spinor states, as well as the crossover from relativistic to non-relativistic dynamics. The high level of control of trapped-ion experimental parameters makes it possible to simulate textbook examples of relativistic quantum physics.

The Dirac equation for a spin-1/2 particle with rest mass m is given by[1]

$$i\hbar\frac{\partial\psi}{\partial t}=(c\boldsymbol{\alpha}\cdot\hat{\mathbf{p}}+\beta mc^2)\psi$$

Here c is the speed of light, $\hat{\mathbf{p}}$ is the momentum operator, α_j ($j=1$, 2, 3; $(\boldsymbol{\alpha})_j=\alpha_j$) and β are the Dirac matrices (which are usually given in terms of the Pauli matrices, σ_{x}, σ_{y} and σ_{z}), the wavefunctions ψ are four-component spinors and $\hbar$ is Planck's constant divided by 2π. A general Dirac spinor can be decomposed into parts with positive and negative energies $E=\pm\sqrt{p^2c^2+m^2c^4}$. Zitterbewegung is understood to be an interference effect between the positive- and negative-energy parts of the spinor and does not appear for spinors that consist entirely of positive-energy (or negative-energy) parts. Furthermore, it is only present when these parts have significant overlap in position and momentum space and is therefore not a sustained effect under most circumstances[1]. For a free electron, the Dirac equation predicts the Zitterbewegung to have an amplitude of the order of the Compton wavelength, $R_{\mathrm{ZB}}\approx 10^{-12}$ m, and a frequency of $\omega_{\mathrm{ZB}}\approx 10^{21}$ Hz, and the effect has so far been experimentally inaccessible. The existence of Zitterbewegung, in relativistic quantum mechanics and in quantum field theory, has been a recurrent subject of discussion in the past years[12,13].

Quantum simulation aims to simulate a quantum system using a controllable laboratory system that underlies the same mathematical model. In this way, it is possible to simulate quantum systems that can be neither efficiently simulated on a classical computer[14] nor easily accessed experimentally, while allowing parameter tunability over a wide range. The difficulties in observing real quantum relativistic effects have generated significant interest in the quantum simulation of their dynamics. Examples include black holes in Bose–Einstein condensates[5] and Zitterbewegung for massive fermions in solid-state physics[6], neither of which have been experimentally realized so far. Also, graphene is studied widely in connection to the Dirac equation[15–17].

Trapped ions are particularly interesting for the purpose of quantum simulation[18–20], as they allow exceptional control of experimental parameters, and initialization and read-out can be achieved with high fidelity. Recently, for example, a proof-of-principle simulation of a quantum magnet was performed[21] using trapped ions. The full, three-dimensional, Dirac equation Hamiltonian can be simulated using lasers coupling to the three vibrational eigenmodes and the internal states of a single trapped ion[7]. The set-up can be significantly simplified when simulating the Dirac equation in $1+1$ dimensions, yet the most unexpected features of the Dirac equation, such as Zitterbewegung and the Klein paradox, remain. In the Dirac equation in $1+1$ dimensions, that is

$$i\hbar\frac{\partial\psi}{\partial t}=H_{\mathrm{D}}\psi=(c\hat{p}\sigma_x+mc^2\sigma_z)\psi$$

there is only one motional degree of freedom and the spinor is encoded in two internal levels, related to positive- and negative-energy states[7]. We find that the velocity of the free Dirac particle is $\mathrm{d}\hat{x}/\mathrm{d}t=[\hat{x},H_{\mathrm{D}}]/i\hbar=c\sigma_x$ in the Heisenberg picture. For a massless particle, $[\sigma_x, H_{\mathrm{D}}]=0$ and, hence, σ_x is a constant of motion. For a massive particle, $[\sigma_x, H_D]\neq 0$ and the evolution of the particle is described by

$$\hat{x}(t)=\hat{x}(0)+\hat{p}c^2H_{\mathrm{D}}^{-1}t+i\hat{\xi}(\mathrm{e}^{2iH_{\mathrm{D}}t/\hbar}-1)$$

where $\hat{\xi}=(1/2)\hbar c(\sigma_x-\hat{p}cH_{\mathrm{D}}^{-1})H_{\mathrm{D}}^{-1}$. The first two terms represent evolution that is linear in time, as expected for a free particle, whereas the third, oscillating, term may induce Zitterbewegung.

For the simulation, we trapped a single $^{40}\mathrm{Ca}^+$ ion in a linear Paul trap[22] with axial trapping frequency $\omega_{\mathrm{ax}}=2\pi\times 1.36$ MHz and radial trapping frequency $\omega_{\mathrm{rad}}=2\pi\times 3$ MHz. Doppler cooling, optical pumping and resolved sideband cooling on the $S_{1/2}\leftrightarrow D_{5/2}$ transition in a magnetic field of 4 G prepare the ion in the axial motional ground state and in the internal state $|S_{1/2}, m_{\mathrm{J}}=1/2\rangle$ (m_{J}, magnetic quantum number). A narrow-linewidth laser at 729 nm couples the states $\binom{0}{1}\equiv|S_{1/2}, m_{\mathrm{J}}=1/2\rangle$ and $\binom{1}{0}\equiv|D_{5/2}, m_{\mathrm{J}}=3/2\rangle$, which we identify as our spinor states. A bichromatic light field resonant with the upper and lower axial motional sidebands of the $\binom{1}{0}\leftrightarrow\binom{0}{1}$ transition with appropriately set phases and frequency realizes the Hamiltonian[7]

$$H_{\mathrm{D}}=2\eta\Delta\tilde{\Omega}\sigma_x\hat{p}+\hbar\Omega\sigma_z \tag{1}$$

Here $\Delta=\sqrt{\hbar/2\tilde{m}\omega_{\mathrm{ax}}}$ is the size of the ground-state wavefunction, with $\tilde{m}$ the ion's mass (not to be confused with the mass, m, of the

[1]Institut für Quantenoptik und Quanteninformation, Österreichische Akademie der Wissenschaften, Otto-Hittmair-Platz 1, A-6020 Innsbruck, Austria. [2]Institut für Experimentalphysik, Universität Innsbruck, Technikerstrasse 25, A-6020 Innsbruck, Austria. [3]Departamento de Química Física, Universidad del País Vasco - Euskal Herriko Unibertsitatea, Apartado 644, 48080 Bilbao, Spain. [4]IKERBASQUE, Basque Foundation for Science, Alameda Urquijo 36, 48011 Bilbao, Spain.

simulated particle); $\eta = 0.06$ is the Lamb–Dicke parameter; and $\hat{p} = i\hbar(a^\dagger - a)/2\varDelta$ is the momentum operator, with $a^\dagger$ and a the usual raising and lowering operators for the motional state along the axial direction. The first term in equation (1) describes a state-dependent motional excitation with coupling strength $\eta\tilde{\Omega}$, corresponding to a displacement of the ion's wave packet in the harmonic trap. The parameter $\tilde{\Omega}$ is controlled by setting the intensity of the bichromatic light field. The second term is equivalent to an optical Stark shift and occurs when the bichromatic light field is detuned from resonance by 2Ω. Equation (1) reduces to the 1 + 1 dimensional Dirac Hamiltonian if we make the identifications $c \equiv 2\eta\tilde{\Omega}\Delta$ and $mc^2 \equiv \hbar\Omega$. The momentum and position of the Dirac particle are then mapped onto the corresponding quadratures of the trapped-ion harmonic oscillator.

To study relativistic effects such as Zitterbewegung, it is necessary to measure $\langle\hat{x}(t)\rangle$, the expectation value of the position operator of the harmonic oscillator. It has been noted theoretically that such expectation values could be measured using very short probe times, without reconstructing the full quantum state[7,23,24]. To measure $\langle\hat{x}\rangle$ for a motional state ρ_m, we have to (1) prepare the ion's internal state in an eigenstate of σ_y, (2) apply a unitary transformation, $U(\tau)$, that maps information about ρ_m onto the internal states and (3) record the changing excitation as a function of the probe time τ, by measuring fluorescence[22]. In this protocol, the unitary operator $U(\tau) = \exp(-i\eta\Omega_{\rm p}\sigma_x\hat{x}\tau/\varDelta)$, with $\hat{x} = (a^\dagger + a)\varDelta$ and probe Rabi frequency $\Omega_{\rm p}$, effectively transforms the observable σ_z into $\sin k\hat{x}$, with $k = 2\eta\Omega_p\tau/\varDelta$, meaning that $\langle\hat{x}\rangle$ can be determined by monitoring the rate of change of $\langle\sin k\hat{x}\rangle$ for short probe times (Methods). Because the Dirac Hamiltonian generally entangles the motional and internal states of the ion, we first incoherently recombine the internal state population in $\binom{0}{1}$ (Methods) before proceeding to step 1. Then we apply the Hamiltonian generating U with the probe Rabi frequency set to $\Omega_{\rm p} = 2\pi \times 13$ kHz for interaction times τ of up to 14 μs, in 1–2-μs steps. The change of excitation was obtained by linear fits, each based on 10^4 to 3×10^4 measurements.

We simulate the Dirac equation by applying $H_{\rm D}$ for varying amounts of time and for different particle masses. In the experiment, we set $\tilde{\Omega} = 2\pi \times 68$ kHz, corresponding to a simulated speed of light of $c = 0.052\varDelta\ \mu{\rm s}^{-1}$. The measured expectation values, $\langle\hat{x}(t)\rangle$, are shown in Fig. 1 for a particle initially prepared in the spinor state $\psi(x; t = 0) = (\sqrt{2\pi}2\varDelta)^{-1/2}{\rm e}^{-x^2/4\varDelta^2}\binom{1}{1}$ by sideband cooling and application of a π/2 pulse. Zitterbewegung appears for particles with non-zero mass, and is obtained by varying Ω in the range $0 < \Omega \leq 2\pi \times 13$ kHz by changing the detuning of the bichromatic lasers.

We investigate the particle dynamics in the crossover from relativistic to non-relativistic dynamics. The data in Fig. 1 well match numerical simulations based on equation (1), which are shown as solid lines. The error bars are obtained from a linear fit assuming quantum projection noise, which dominates noise caused by fluctuations of control parameters. In addition, the data were fitted with a heuristic model function of the form $\langle\hat{x}(t)\rangle = at + R_{\rm ZB}\sin\omega_{\rm ZB}t$ to extract the effective amplitude, $R_{\rm ZB}$, and frequency, $\omega_{\rm ZB}$, of the Zitterbewegung shown in the inset. As the particle's initial momentum is not dispersion free, the amplitude and frequency are only approximate concepts. From these data, it can be seen that the frequency, $\omega_{\rm ZB} \approx 2\Omega$, grows linearly with increasing mass, whereas the amplitude decreases as the mass is increased. Because the mass of the particle increases but the momentum and the simulated speed of light remain constant, the data in Fig. 1 show the crossover from the far relativistic to non-relativistic limits. Hence, the data confirm that Zitterbewegung decreases in both limits, as theoretically expected. In the far-relativistic case, this is because $\omega_{\rm ZB}$ vanishes; in the non-relativistic case, it is because $R_{\rm ZB}$ vanishes.

The tools with which we simulate the Dirac equation can also be used to set the initial state of the simulated particle precisely. The particle in Fig. 2a was given an average initial momentum $\langle\hat{p}(t = 0)\rangle = \hbar/\varDelta$ by means of a displacement operation using the Hamiltonian $H = \hbar\eta\tilde{\Omega}\sigma_x\hat{x}/\varDelta$. The initial state of this particle consists

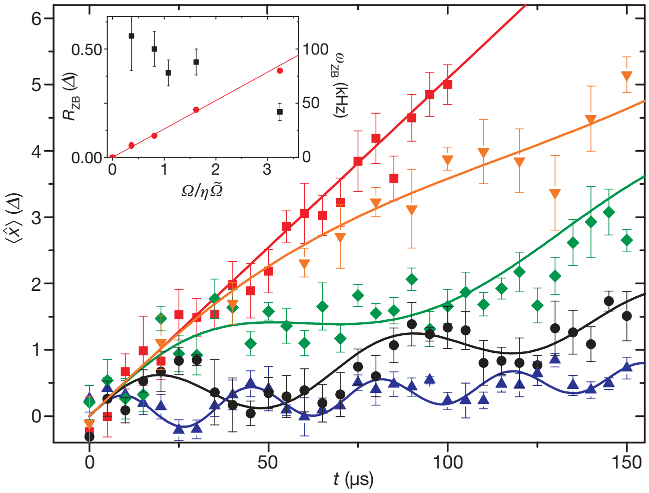


Figure 1 | Expectation values, $\langle\hat{x}(t)\rangle$, for particles with different masses. The linear curve (squares) represents a massless particle ($\Omega = 0$) moving at the speed of light, which is given by $c = 2\eta\tilde{\Omega}\Delta = 0.052\varDelta\ \mu{\rm s}^{-1}$ for all curves. From the top, the other curves represent particles of increasing masses. Their Compton wavelengths are given by $\lambda_{\rm C} \equiv 2\eta\tilde{\Omega}\Delta/\Omega = 5.4\varDelta$ (down triangles), $2.5\varDelta$ (diamonds), $1.2\varDelta$ (circles) and $0.6\varDelta$ (up triangles), respectively. The solid curves represent numerical simulations. The figure shows Zitterbewegung for the crossover from the relativistic limit, $2\eta\tilde{\Omega} \gg \Omega$, to the non-relativistic limit, $2\eta\tilde{\Omega} \ll \Omega$. Inset, fitted Zitterbewegung amplitude, $R_{\rm ZB}$ (squares), and frequency, $\omega_{\rm ZB}$ (circles), versus the parameter $\Omega/\eta\tilde{\Omega}$ (which is proportional to the mass). Error bars, 1σ.

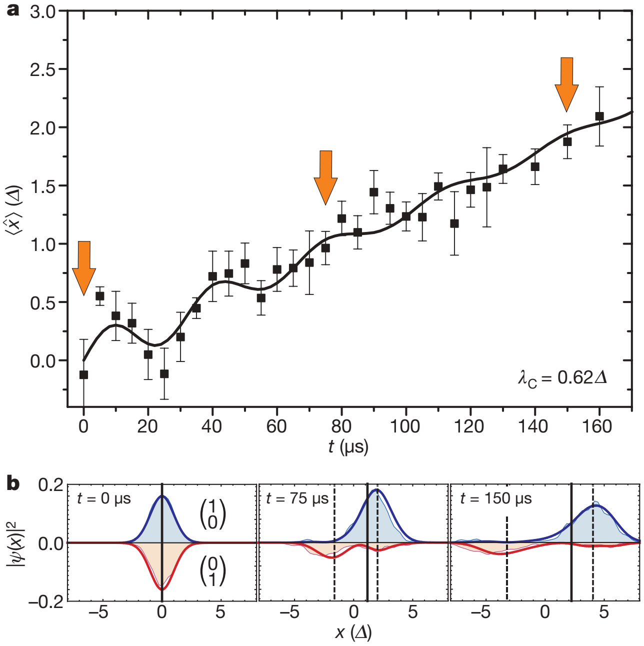


Figure 2 | Zitterbewegung for a state with non-zero average momentum. **a**, Initially, Zitterbewegung appears owing to interference of positive- and negative-energy parts of the state, $\psi(x; t = 0) = {\rm e}^{ix/\varDelta}e^{-x^2/4\varDelta^2}(\sqrt{2\pi}2\varDelta)^{-1/2}\binom{1}{1}$. As these parts separate, the oscillatory motion fades away. The solid curve represents a numerical simulation. Error bars, 1σ. **b**, Measured (filled areas) and numerically calculated (solid lines) probability distributions, $|\psi(x)|^2$, at times $t = 0$, 75 and 150 μs (as indicated by the arrows in **a**). The probability distribution corresponding to the state $\binom{0}{1}$ is inverted for clarity. The vertical solid line represents $\langle\hat{x}\rangle$ as plotted in **a**. The two dashed lines indicate the respective expectation values for the positive- and negative-energy parts of the spinor.

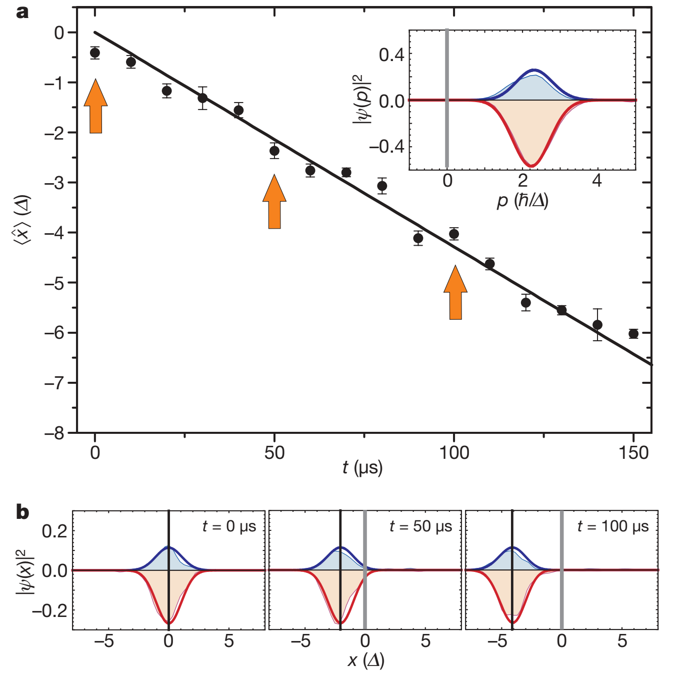


Figure 3 | Time evolution of a negative-energy eigenstate with $\lambda_C = 1.2\Delta$. Laser pulses create the spinor $\psi(p; t=0) = \sqrt{\Delta/\hbar}\begin{pmatrix} -0.48\exp[-(p-2.26)^2\Delta^2/\hbar^2] \\ 0.75\exp[-(p-2.14)^2\Delta^2/\hbar^2] \end{pmatrix}$, which approximates a negative-energy spinor with average momentum $\langle\hat{p}\rangle = 2.2\hbar/\Delta$. The corresponding initial momentum distribution, $|\tilde{\psi}(p)|^2$, is shown in the inset. The filled curves represent data, whereas the solid lines represent a numerical calculation. The data in **a** show no Zitterbewegung. The solid curve represents a numerical simulation. Error bars, 1σ. **b**, Measured probability distributions, $|\psi(x)|^2$, for three different evolution times (indicated by the arrows in **a**). There is no splitting of the wavefunction and the evolution and spreading is as intuitively expected for a free particle.

of a positive-energy component with positive velocity and a negative-energy component with negative velocity[25]. The positive-energy component moves to the right and is contributed to by both spinor states (Methods), whereas the negative-energy component moves to the left. Zitterbewegung is observed as long as these parts overlap, and dies out as they separate. Further information is obtained by a complete reconstruction of the probability distribution[26] $|\psi(x)|^2$, shown in Fig. 2b. It is also possible to initialize the spinor in a pure negative- or positive-energy state (Methods). In Fig. 3a, we show the time evolution, $\langle\hat{x}(t)\rangle$, of a negative-energy spinor with average momentum $\langle\hat{p}\rangle = 2.2\hbar/\Delta$. The corresponding reconstructed probability distributions are displayed in Fig. 3b, and it can be seen that there is neither Zitterbewegung nor splitting of the wavefunction, which occurs only if there are positive- and negative-energy contributions to the wavefunction.

We have implemented a proof-of-principle quantum optical simulation of a tunable relativistic quantum mechanical system. We have demonstrated that the simulated one-dimensional Dirac dynamics for a free particle shows Zitterbewegung and several of its counterintuitive quantum relativistic features. A natural route for the near future will be to move theoretically and experimentally towards the simulation of dynamics that are impossible (or difficult) to calculate in real systems, such as in quantum chemistry[27] or quantized Dirac fields in the context of quantum field theory[1]. Our experiment serves as a first step towards more complex quantum simulations. Furthermore, the mapping between quantum optical systems and relativistic quantum mechanics may be followed by further analogies between the Dirac dynamics and the Jaynes–Cummings model[8,28,29], and in photonic[9] or sonic systems[30].

METHODS SUMMARY

Measurements in position space are carried out by mapping the observable of interest onto the ion's internal state. Applying a state-dependent displacement operation, $U = \exp(-ik\hat{x}\sigma_x/2)$, to the quantum state ρ, followed by a measurement of σ_z, is equivalent to measuring the observable

$$A(k) = U^\dagger\sigma_z U = \cos(k\hat{x})\sigma_z + \sin(k\hat{x})\sigma_y$$

on the original state ρ, where $k = 2\eta\Omega_p t/\Delta$ is proportional to the interaction time, t. If the ion's internal initial state is the eigenstate of σ_z belonging to eigenvalue $+1$, then $\langle A(k)\rangle = \langle\cos k\hat{x}\rangle$. Similarly, for the eigenstate of σ_y belonging to eigenvalue $+1$, $\langle A(k)\rangle = \langle\sin k\hat{x}\rangle$. A Fourier transformation of these measurements yields the probability density $|\psi(x)|^2$ in position space (or equivalently $\langle\delta(\hat{x}-x)\rangle$ if the state is not pure but mixed). Moreover, the coefficients of the Taylor expansion of the observable $A(k)$ are proportional to the moments $\hat{x}^n$, and in particular $\mathrm{d}\langle A(k)\rangle/\mathrm{d}k|_{t=0} \propto \langle\hat{x}\rangle$. The reconstruction of the wave packets associated with the spinor components $\binom{1}{0}$ and $\binom{0}{1}$, shown in Figs 2 and 3, is achieved by projecting either part of the wavefunction onto the $D_{5/2}$ state using a fluorescence measurement followed by the measurement scheme based on post-selected data, described above.

To construct spinors with either purely positive- or negative-energy solutions, it is useful to express a general spinor as $\psi = P^+\psi + P^-\psi$, that is, using the projection operations projecting onto the positive- and negative-energy contributions ($E_\pm = \sqrt{c^2p^2 + m^2c^4}$). In momentum space, the projection operators are given by

$$P^\pm(p) = \frac{1}{2}\left(I_2 \pm \frac{cp\sigma_x + mc^2\sigma_z}{\sqrt{c^2p^2 + m^2c^4}}\right)$$

Here I_2 is the 2×2 identity matrix. The spinor state in Fig. 3 was 'reverse-engineered' by projecting out the negative-energy part of a wave packet with average momentum $\langle\hat{p}\rangle = 2.2\hbar/\Delta$ and renormalizing the spinor. The relative contributions of the two spinor states, and the phase between them, can be set straightforwardly in the experiment. The momentum distributions can be approximated by Gaussians with appropriately set average momenta.

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

Received 31 August; accepted 11 November 2009.

1. Thaller, B. *The Dirac Equation* (Springer, 1992).
2. Anderson, C. D. The positive electron. *Phys. Rev.* **43,** 491–494 (1933).
3. Klein, O. Die Reflexion von Elektronen an einem Potentialsprung nach der relativistischen Dynamik von Dirac. *Z. Phys.* **53,** 157–165 (1929).
4. Schrödinger, E. Über die kräftefreie Bewegung in der relativistischen Quantenmechanik. *Sitz. Preuss. Akad. Wiss. Phys.-Math. Kl.* **24,** 418–428 (1930).
5. Garay, L. J., Anglin, J. R., Cirac, J. I. & Zoller, P. Sonic analog of gravitational black holes in Bose-Einstein condensates. *Phys. Rev. Lett.* **85,** 4643–4647 (2000).
6. Schliemann, J., Loss, D. & Westervelt, R. M. Zitterbewegung of electronic wave packets in III–V zinc-blende semiconductor quantum wells. *Phys. Rev. Lett.* **94,** 206801 (2005).
7. Lamata, L., León, J., Schätz, T. & Solano, E. Dirac equation and quantum relativistic effects in a single trapped ion. *Phys. Rev. Lett.* **98,** 253005 (2007).
8. Bermudez, A., Martin-Delgado, M. A. & Solano, E. Exact mapping of the 2+1 Dirac oscillator onto the Jaynes-Cummings model: ion-trap experimental proposal. *Phys. Rev. A* **76,** 041801(R) (2007).
9. Zhang, X. Observing *Zitterbewegung* for photons near the Dirac point of a two-dimensional photonic crystal. *Phys. Rev. Lett.* **100,** 113903 (2008).
10. Vaishnav, J. Y. & Clark, C. W. Observing *Zitterbewegung* with ultracold atoms. *Phys. Rev. Lett.* **100,** 153002 (2008).
11. Otterbach, J., Unanyan, R. G. & Fleischhauer, M. Confining stationary light: Dirac dynamics and Klein tunneling. *Phys. Rev. Lett.* **102,** 063602 (2009).
12. Krekora, P., Su, Q. & Grobe, R. Relativistic electron localization and the lack of *Zitterbewegung*. *Phys. Rev. Lett.* **93,** 043004 (2004).
13. Wang, Z.-Y. & Xiong, C.-D. *Zitterbewegung* by quantum field-theory considerations. *Phys. Rev. A* **77,** 045402 (2008).
14. Feynman, R. Simulating physics with computers. *Int. J. Theor. Phys.* **21,** 467–488 (1982).
15. Cserti, J. & Dávid, G. Unified description of Zitterbewegung for spintronic, graphene, and superconducting systems. *Phys. Rev. B* **74,** 172305 (2006).
16. Katsnelson, M. I., Novoselov, K. S. & Geim, A. K. Chiral tunnelling and the Klein paradox in graphene. *Nature Phys.* **2,** 620–625 (2006).
17. Neto, A. H. C., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81,** 109–162 (2009).
18. Leibfried, D. *et al.* Trapped-ion quantum simulator: experimental application to nonlinear interferometers. *Phys. Rev. Lett.* **89,** 247901 (2002).
19. Porras, D. & Cirac, J. I. Effective quantum spin systems with trapped ions. *Phys. Rev. Lett.* **92,** 207901 (2004).
20. Johanning, M., Varón, A. F. & Wunderlich, C. Quantum simulations with cold trapped ions. *J. Phys. B* **42,** 154009 (2009).
21. Friedenauer, H., Schmitz, H., Glueckert, J., Porras, D. & Schaetz, T. Simulating a quantum magnet with trapped ions. *Nature Phys.* **4,** 757–761 (2008).
22. Kirchmair, G. *et al.* Deterministic entanglement of ions in thermal states of motion. *New J. Phys.* **11,** 023002 (2009).

23. Lougovski, P., Walther, H. & Solano, E. Instantaneous measurement of field quadrature moments and entanglement. *Eur. Phys. J. D* **38,** 423–426 (2006).
24. Santos, M. F., Giedke, G. & Solano, E. Noise-free measurement of harmonic oscillators with instantaneous interactions. *Phys. Rev. Lett.* **98,** 020401 (2007).
25. Thaller, B. Visualizing the kinematics of relativistic wave packets. Preprint at ⟨http://arxiv.org/abs/quant-ph/0409079⟩ (2004).
26. Wallentowitz, S. & Vogel, W. Reconstruction of the quantum mechanical state of a trapped ion. *Phys. Rev. Lett.* **75,** 2932–2935 (1995).
27. Aspuru-Guzik, A., Dutoi, A. D., Love, P. J. & Head-Gordon, M. Simulated quantum computation of molecular energies. *Science* **309,** 1704–1707 (2005).
28. Rozmej, P. & Arvieu, R. The Dirac oscillator: a relativistic version of the Jaynes-Cummings model. *J. Phys. A* **32,** 5367–5382 (1999).
29. Bermudez, A., Martin-Delgado, M. A. & Solano, E. Dirac cat states in relativistic Landau levels. *Phys. Rev. Lett.* **99,** 123602 (2007).
30. Zhang, X. & Liu, Z. Extremal transmission and beating effect of acoustic waves in two-dimensional sonic crystals. *Phys. Rev. Lett.* **101,** 264303 (2008).

Acknowledgements We gratefully acknowledge support from the Austrian Science Fund, the European Commission (Marie-Curie programme) and the Institut für Quanteninformation GmbH. E.S. acknowledges support of Universidad del País Vasco - Euskal Herriko Unibertsitatea grant GIU07/40 and European Union project EuroSQIP. This material is based upon work supported in part by Intelligence Advanced Research Projects Activity. We thank H. Häffner and M. Baranov for comments on the manuscript.

Author Contributions G.K., F.Z. and R.G. performed the experiment and analysed the data; E.S. provided the theoretical analysis; C.F.R. and R.G. designed the experiment and carried out numerical simulations; R.B., G.K., F.Z., R.G. and C.F.R. contributed to the experimental set-up; and all authors co-wrote the paper.

Author Information 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 C.F.R. (christian.roos@uibk.ac.at).

METHODS

Measurement of $\langle\hat{x}\rangle$ and $|\psi(x)|^2$. In ion-trap experiments, the only observable that can directly be measured by fluorescence detection is σ_z. Additional laser pulses can be used to map other observables onto σ_z. In the experiment, we apply a state-dependent displacement operation, $U = \exp(-ik\hat{x}\sigma_x/2)$, to the quantum state ρ, and then measure σ_z, which is equivalent to measuring the observable

$$A(k) = U^{\dagger}\sigma_z U = \cos(k\hat{x})\sigma_z + \sin(k\hat{x})\sigma_y$$

on the initial state ρ, because $\mathrm{Tr}((U^{\dagger}\rho U)\sigma_z) = \mathrm{Tr}(\rho(U\sigma_z U^{\dagger}))$. Here $k = 2\eta\Omega_{\mathrm{p}}t/\Delta$ is proportional to the interaction time, t. We have $\langle A(k)\rangle = \langle\cos k\hat{x}\rangle$ if the ion's internal initial state is prepared in the eigenstate of σ_z belonging to eigenvalue $+1$; for an ion prepared in the eigenstate of σ_y belonging to eigenvalue $+1$, we obtain $\langle A(k)\rangle = \langle\sin k\hat{x}\rangle$. A Fourier transformation of these measurements yields the probability density $|\psi(x)|^2$ in position space.

For the position operator, we have $\mathrm{d}\langle A(k)\rangle/\mathrm{d}k|_{t=0} \propto \langle\hat{x}\sigma_y\rangle$. Measuring $\langle\hat{x}\rangle$ thus requires the preparation of an eigenstate of σ_y, which cannot be done directly when the motional state is entangled with the internal state. To solve this problem, we first incoherently recombine the internal state in $\binom{0}{1}$. This is done by first shelving the population initially in $\binom{0}{1}$ to $|D_{5/2}, m_{\mathrm{J}} = 5/2\rangle$ using a rapid adiabatic passage transfer. A second such transfer shifts the population in $\binom{1}{0}$ to $\binom{0}{1}$. A 100-μs laser pulse at 854 nm transfers the population in $|D_{5/2}, m_{\mathrm{J}} = 5/2\rangle$ to $|P_{3/2}, m_{\mathrm{J}} = 3/2\rangle$, from which it spontaneously decays to $\binom{0}{1}$. The transfer efficiency is >99%, limited by the small branching ratio to the $D_{3/2}$ state. In the transfer steps, a probability exists that the motional state of the ion is changed. This probability is however very small, owing to the small Lamb–Dicke parameter, but could be eliminated completely by a separate measurement of the motional states of the spinor states $\binom{0}{1}$ and $\binom{1}{0}$, at the expense of a longer data acquisition time.

To distinguish between populations in the states $\binom{1}{0}$ and $\binom{0}{1}$, when reconstructing $|\psi(x)|^2$ (as shown in Figs 2 and 3), we applied a short (200-μs) fluorescence detection to measure the internal state. We used only cases in which $\binom{1}{0}$ was measured (leaving the motional state unchanged as no photons were scattered) for the subsequent analysis. To reconstruct $|\psi(x)|^2$ belonging to $\binom{0}{1}$, a π pulse before the short detection was used to interchange the internal state populations.

Constructing a pure negative-energy spinor. A general spinor is built up out of positive- and negative-energy components (energies $E_{\pm} = \pm\sqrt{c^2p^2 + m^2c^4}$) such that $\psi = P^{+}\psi + P^{-}\psi$. In momentum space, the projection operators are given by

$$P^{\pm}(p) = \frac{1}{2}\left(I_2 \pm \frac{cp\sigma_x + mc^2\sigma_z}{\sqrt{c^2p^2 + m^2c^4}}\right) \tag{2}$$

Here I_2 is the 2 × 2 identity matrix. In general, the projection operators do not project onto the spinor basis states. The exception is when $p = 0$, because in this case the projector in equation (2) becomes diagonal in the spinor basis. The spinor state in Fig. 3 was 'reverse-engineered' by projecting out the negative-energy part of a wave packet with average momentum $\langle\hat{p}\rangle = 2.2\hbar/\Delta$ and renormalizing the spinor.

The complete sequence for approximating the negative-energy state is conveniently described in the basis of the eigenstates $|\pm\rangle_y = (1/\sqrt{2})\binom{1}{\pm i}$ of σ_y. After ground-state cooling, we prepare the state $|+\rangle_y$. Then we displace this state to one with average momentum $\langle\hat{p}\rangle = 2.2\hbar/\Delta$ by using the displacement Hamiltonian $H = \hbar\eta\tilde{\Omega}\sigma_y\hat{x}/\Delta$. Next, a far-detuned laser pulse rotates the internal state to $0.84|+\rangle_y + i0.53|-\rangle_y$. The displacement Hamiltonian $H = -\hbar\eta\tilde{\Omega}\sigma_y\hat{x}/\Delta$ shifts these parts in opposite directions to create the required asymmetry between the average momenta of the components. A final π/2 pulse creates the state shown in Fig. 3. This state has >99% overlap with the desired negative-energy state.

nature Vol 463 | 7 January 2010 | doi:10.1038/nature08681

LETTERS

Preparation and detection of a mechanical resonator near the ground state of motion

T. Rocheleau[1]*, T. Ndukum[1]*, C. Macklin[1], J. B. Hertzberg[2], A. A. Clerk[3] & K. C. Schwab[4]

Cold, macroscopic mechanical systems are expected to behave contrary to our usual classical understanding of reality; the most striking and counterintuitive predictions involve the existence of states in which the mechanical system is located in two places simultaneously. Various schemes have been proposed to generate and detect such states[1,2], and all require starting from mechanical states that are close to the lowest energy eigenstate, the mechanical ground state. Here we report the cooling of the motion of a radio-frequency nanomechanical resonator by parametric coupling to a driven, microwave-frequency superconducting resonator. Starting from a thermal occupation of 480 quanta, we have observed occupation factors as low as 3.8 ± 1.3 and expect the mechanical resonator to be found with probability 0.21 in the quantum ground state of motion. Further cooling is limited by random excitation of the microwave resonator and heating of the dissipative mechanical bath. This level of cooling is expected to make possible a series of fundamental quantum mechanical observations including direct measurement of the Heisenberg uncertainty principle and quantum entanglement with qubits.

Naively treating the motion of a mechanical resonator quantum mechanically produces the elementary result that the energy should be quantized: $E_n = \hbar\omega_m(n + 1/2)$, where n is an integer, ω_m is the resonant frequency (m denoting the mechanical resonator) and $\hbar$ is Planck's constant divided by 2π. In thermal equilibrium, an average occupation factor is expected to follow the Bose–Einstein distribution: $\bar{n}_m^T = (e^{\hbar\omega_m/k_B T} - 1)^{-1}$, where T and k_B are the temperature and Boltzmann's constant, respectively. Cooling a resonator into the quantum regime where $\bar{n}_m^T \ll 1$, and measuring the very small motions, has been challenging for a number of technical reasons; not only are very low temperatures necessary to freeze out the mode, but detection with sensitivity at the quantum zero-point level, that is, on length scales $x_{zp} = \sqrt{\hbar/2m\omega_m}$, is required. Furthermore, this strong position measurement must not heat the mode with measurement back-action[3].

Many strategies have been proposed[4–8] and applied to realize the quantum regime, with increasing success. Experiments with nanoelectromechanical structures have been able to reach a mechanical occupation of $\bar{n}_m = 25$ by passively cooling a nanomechanical resonator[3], detected with a superconducting single-electron transistor. (The mechanical occupation is not necessarily in equilibrium with the thermal occupation, $\bar{n}_m^T$.) Researchers experimenting with optomechanical systems have been able to use extremely sensitive optical detection and radiation pressure to both cool and detect $\bar{n}_m = 65$ in a toroidal resonator[9], $\bar{n}_m = 37$ in microsphere resonator[10] and $\bar{n}_m = 35$ in an optical cavity[11].

The technique we use to both cool and detect the motion of a nanomechanical resonator close to the ground state involves parametrically coupling the motion to a superconducting microwave resonator (SMR)[12,13] (Fig. 1). The nanomechanical resonator has a fundamental in-plane flexural resonance of $\omega_m = 2\pi \times 6.3$ MHz and is capacitively coupled to a symmetric, two-port, half-wave SMR that resonates at $\omega_{\rm SMR} = 2\pi \times 7.5$ GHz. The device is located in a dilution refrigerator and pumped through carefully filtered and cooled leads. The thermal occupation of the SMR, $\bar{n}_{\rm SMR}^T$, is expected to be 0.09 at 146 mK.

The nanomechanical-resonator damping rate, Γ_m^T, has an unusual linear temperature dependence below 600 mK, reaching a resonator quality factor of $Q \approx 10^6$ at 100 mK. The SMR damping rate, $\kappa = 2\pi \times 600$ kHz, is essentially temperature independent below 700 mK and is a factor of 2.4 higher than expected from design owing to internal losses.

The Hamiltonian that describes the coupled resonators is given by[7,8]

$$\hat{H} = \hbar\left(\omega_{\rm SMR} + g\hat{x} - \frac{1}{2}\lambda\hat{x}^2\right)\left(\hat{b}^\dagger\hat{b} + \frac{1}{2}\right) + \hbar\omega_m\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)$$

where $\hat{a}$ and $\hat{a}^\dagger$ are respectively the nanomechanical-resonator annihilation and creation operators, and $\hat{b}$ and $\hat{b}^\dagger$ are those of the SMR. The first term shows the ponderomotive-like coupling of the SMR's frequency to the mechanical motion: $\hat{x} = x_{\rm zp}(\hat{a}^\dagger + \hat{a})$ and $g = \partial\omega_{\rm SMR}/\partial x = (\omega_{\rm SMR}/2C_t)(\partial C_g/\partial x)$, where $C_g(x) = 450 \pm 50$ aF is the coupling capacitance and $C_t = 260$ fF is the SMR's total effective capacitance. The term proportional to $\hat{x}^2$ results from the electrostatic frequency-pulling of the mechanical resonator by the SMR[14], where $\lambda = (\omega_{\rm SMR}/2C_t)(\partial^2 C_g/\partial x^2)$, and is responsible for parametric instabilities under certain pump configurations.

When pumping the SMR at $\omega_p = \omega_{\rm SMR} - \omega_m$, harmonic motion of the nanomechanical resonator preferentially up-converts microwave photons to frequency $\omega_{\rm SMR}$, extracting one radio-frequency nanomechanical-resonator quantum for each up-converted microwave SMR photon, a process that both damps and cools the nanomechanical resonator's motion[7,8,15–17]. This cooling process is analogous to Raman scattering and the process used to cool an atomic ion to the quantum ground state of motion[8,18]. In the sideband-resolved limit, $\kappa < \omega_m$, the rate of this up-conversion process is given by $\Gamma_{\rm opt} \approx 4x_{\rm zp}^2 g^2 \bar{n}_p/\kappa$, where $\bar{n}_p$ is the occupation of the SMR resulting from the pumping.

From detailed balance, the nanomechanical-resonator occupation factor is expected to follow

$$\bar{n}_m = \frac{\Gamma_m^T \bar{n}_m^T + \Gamma_{\rm opt}\bar{n}_{\rm SMR}}{\Gamma_m^T + \Gamma_{\rm opt}} \qquad (1)$$

where $\bar{n}_{\rm SMR} = (\kappa/4\omega_m)^2 + \bar{n}_{\rm SMR}^T[1 + 2(\kappa/4\omega_m)^2]$ is the effective occupancy associated with the SMR's back-action when $\Gamma_{\rm opt} < \kappa$ (ref. 19). The first term in the expression for $\bar{n}_{\rm SMR}$ is due to the

[1]Department of Physics, Cornell University, Ithaca, New York 14853, USA. [2]Department of Physics, University of Maryland, College Park, Maryland 20742, USA. [3]Department of Physics, McGill University, Montreal, Quebec H3A 2T8, Canada. [4]Applied Physics, Caltech, Pasadena, California 91125, USA.
*These authors contributed equally to this work.

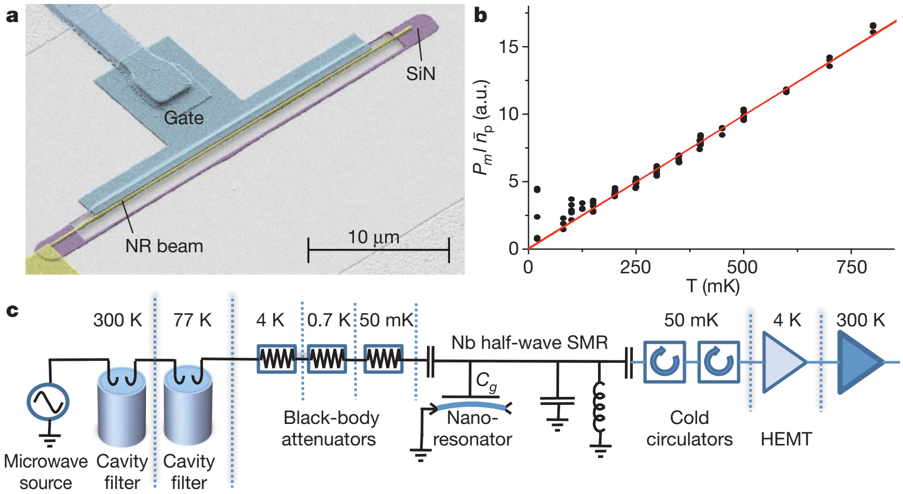


Figure 1 | Nanomechanical device, measurement diagram and thermal calibration. **a,** Nb–Al–SiN sample: the nanomechanical resonator is 30 µm long, 170 nm wide and 140 nm thick, is formed of 60 nm of stoichiometric, high-stress, low-pressure chemical-vapour-deposition SiN[29] and 80 nm of Al, and is located 75 nm from the gate electrode connected to the SMR. The SMR is made from a 345-nm-thick Nb film and has a waveguide characteristic impedance of 126 Ω. **b,** Thermal calibration of the up-converted noise power. **c,** Ultralow-noise, cryogenic measurement circuit with the SMR shown schematically as an equivalent inductance and capacitance. NR, nanomechanical resonator; HEMT, high-electron-mobility transistor; a.u., arbitrary units.

quantum fluctuations of the pump field, and the second term is due to the thermal occupation of the SMR, $\bar{n}^T_{\mathrm{SMR}}$. The expressions above show that the minimum mechanical occupation possible is the effective occupation of the SMR.

The first realization of cooling in a parametrically coupled, electromechanical microwave system was with a kilogram-scale gravitational wave transducer[20], cooling from $\bar{n}_m = 10^8$ to $\bar{n}_m = 10^5$. Cooling of a nanomechanical resonator with an SMR was recently demonstrated[21] and achieved cooling from $\bar{n}_m = 700$ to $\bar{n}_m = 120$ using a scheme similar to that presented here. Our results are made possible by improvements in device engineering (the coupling strength between the nanomechanical resonator and the SMR, g; and the maximum SMR occupation, $\bar{n}_{\mathrm{SMR}}$), which leads to an improvement in Γ_{opt} and resulting cooling rates of two orders of magnitude.

The up-converted noise power is calibrated by applying a weak pump signal ($\Gamma_{\mathrm{opt}} < \Gamma^T_m$) and measuring the resulting integrated sideband power, P_m, normalized by the applied microwave pump power, $\bar{n}_{\mathrm{p}}$, as a function of refrigerator temperature, T (Fig. 1b). For temperatures above ~150 mK, we observe the expected behaviour consistent with equipartition and use this curve to establish the relationship between measured output noise power and $\bar{n}_m$. For temperatures below 150 mK, we observe fluctuations in $\bar{n}_m$ apparently due to a non-thermal, intermittent force noise at the level of $10^{-18}\,\mathrm{N\,Hz}^{-1/2}$, which is observed in other similar samples[22] and is similar to anomalous heating effects in other systems[23]. Furthermore, the linear temperature dependence of Γ^T_m causes the nanomechanical resonator to decouple from the thermal environment at the lowest measured temperatures. Although the behaviour of the nanomechanics seems consistent with a non-thermal force, we do not currently understand its source.

The measured signal powers are consistent with our knowledge of the attenuation and gain of our measurement circuit and estimates of the device parameters. We find that $g/2\pi = 84 \pm 5\,\mathrm{kHz\,nm}^{-1}$, which is the largest coupling strength so far demonstrated in a system of this type. From measurements of ω_m versus $\bar{n}_{\mathrm{p}}$ and pump frequency, we determine that $\lambda/2\pi = 2.1 \pm 0.7\,\mathrm{kHz\,nm}^{-2}$.

With the refrigerator stabilized at $T = 146$ mK ($\bar{n}^T_m = 480$), we measure output noise spectra, $S_x(\omega)$, versus the SMR pump occupation, $\bar{n}_{\mathrm{p}}$. $S_x(\omega)$ is referred to the oscillator position using the nanomechanical-resonator thermal noise calibration, and is composed of up-converted microwave photons due to $\bar{n}_m$, SMR noise due to $\bar{n}_{\mathrm{SMR}}$, and HEMT amplifier noise. We measure $\bar{n}_{\mathrm{SMR}}$ directly by observing the noise spanning the SMR resonance. Back-action correlations between the nanomechanical-resonator motion and the SMR field are important in our measured noise spectra at the lowest mechanical occupation factors. Fluctuations in the SMR voltage, due to $\bar{n}_{\mathrm{SMR}}$, together with the pump, produce forces at frequency ω_m. The resulting motion, together with the pump, produces noise at frequency ω_{SMR}, but 180° out of phase with the original SMR fluctuations. This correlation results in an inverted noise peak, similar to noise squashing[24], which adds incoherently to the noise power driven by the thermal bath. The resulting observed noise peak or dip, $\bar{n}_{\mathrm{eff}}$, is calibrated using thermal noise. Our analysis shows that the nanomechanical resonator occupation factor is given by $\bar{n}_m = \bar{n}_{\mathrm{eff}} + 2\bar{n}_{\mathrm{SMR}}$ (Supplementary Information). Figure 2 shows measurements of $S_x(\omega)$ in three cases at low occupation factors: $\bar{n}_{\mathrm{eff}} > 0$, $\bar{n}_{\mathrm{eff}} \approx 0$ and, showing the squashed output noise, $\bar{n}_{\mathrm{eff}} < 0$.

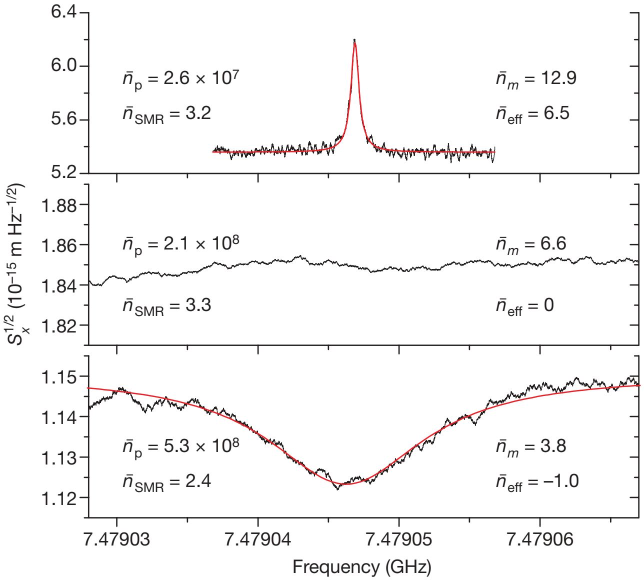


Figure 2 | Measured noise spectra. The noise squashing effect on $S_x(\omega)$ due to the finite occupation of the SMR can be seen in three situations: $\bar{n}_{\mathrm{eff}} > 0$ (top), $\bar{n}_{\mathrm{eff}} \approx 0$ (middle) and $\bar{n}_{\mathrm{eff}} < 0$ (bottom). The red curves show Lorentzian fits through the mechanically up-converted sideband.

Taking the effects of $\bar{n}_{\rm SMR}$ into account in this way, the lowest mechanical occupation we have observed is $\bar{n}_m = 3.8 \pm 1.3$, shown in Fig. 2, with the uncertainty dominated by the uncertainty in $\bar{n}_{\rm SMR}$. At this low occupation factor, the resonator is expected to be found in the ground state with probability $P_0 = 1/(\bar{n}_m + 1) = 0.21$. The cooling power of this refrigeration technique is $\dot{Q} = \hbar\omega_m\Gamma_{\rm opt} = 10^{-22}$ W.

We lowered the refrigerator temperature to 20 mK and did not observe a decrease in the minimum $\bar{n}_m$ value. Using the detailed balance relationship and the measured values of $\bar{n}_m$, $\bar{n}_{\rm SMR}$ and $\Gamma_{\rm opt}$ values, we can compute the bath heating rate, $\dot{n}_T = \Gamma_m^T \bar{n}_m^T$, as a function of $\bar{n}_{\rm p}$ (Fig. 3). It is clear that as $\bar{n}_{\rm p}$ increases above 3×10^7, $\dot{n}_T$ begins to increase, nullifying the benefit of starting at low temperatures. This level of heating is consistent with ohmic losses in the metal film on top of the nanomechanical resonator, and the thermal conductance of a normal-state electron gas.

To check the behaviour of our system (nanomechanical resonator and SMR) in a range where $\bar{n}_{\rm SMR}^T$ is not a complicating factor, we applied radio-frequency electrostatic force noise at the nanomechanical-resonator frequency. Starting from $\bar{n}_m = 2.5 \times 10^5$, we observe cooling, by a factor of 3,000, to $\bar{n}_m = 80$, which closely follows the expected cooling curve over the full range of $\bar{n}_{\rm p}$ (Fig. 4). This suggests that the increase in bath rate at high pump power is due primarily to an increase in bath temperature and not a significant increase in Γ_m^T. To check the stability of our system, we cycled the device from millikelvin temperatures to 77 K and back. To within a few per cent, we observed no change in the microwave signal levels generated by thermal noise, and no change in $\Gamma_{\rm opt}$ as a function of $\bar{n}_{\rm p}$. We also found the out-of-plane mechanical resonance to be 150 kHz lower than the in-plane resonance used in this work. This additional resonance is sufficiently different in frequency that we do not expect any significant interaction.

These measurements identify three effects that work against the cooling process: excess fluctuations of the SMR ($\bar{n}_{\rm SMR}$), heating of the nanomechanical resonator thermal bath at high pump powers, and the non-thermal force noise at low temperatures.

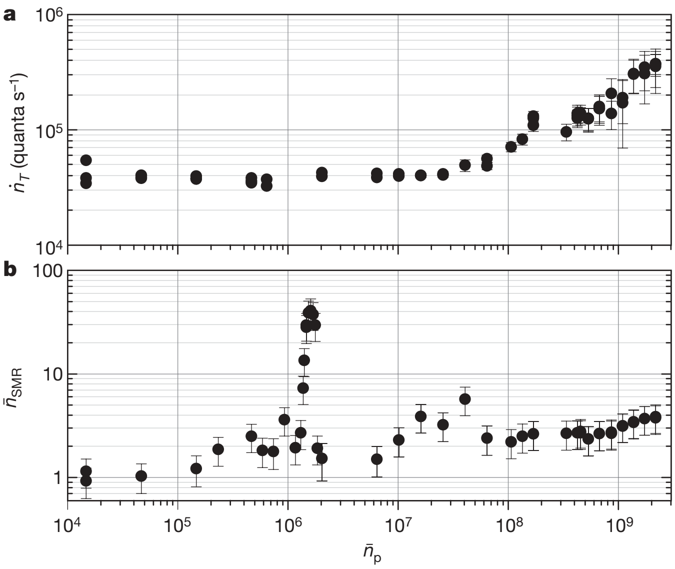


Figure 3 | Nanomechanical heating rate and superconducting-resonator occupation versus pump strength. **a**, The upper figure shows the bath heating rate, $\dot{n}_T$, and the onset of excess heating above $\bar{n}_{\rm p} = 3 \times 10^7$. **b**, The lower figure shows the measured value of thermal occupation of the SMR; the structure is suspected to be related to temporal dynamics of the transition between superconducting and normal states of the metal films and resulting microwave sideband generation[30]. The error bars on $\bar{n}_{\rm SMR}$ (s.e.m.) are dominated by the uncertainty in the transmission of our microwave circuit: gain of our HEMT amplifier and attenuation of our microwave cables and isolators. The error bars on $\dot{n}_T$ (s.e.m.) result from standard error propagation in equation (1) and are dominated by the uncertainty in $\bar{n}_{\rm SMR}$.

We believe that the excess SMR occupation, $\bar{n}_{\rm SMR}$, is not a result of phase or amplitude noise of our microwave source: the pump signal is filtered using tunable, copper microwave cavities (one at 300 K ($Q = 9.5 \times 10^3$) and a second at 77 K ($Q = 2.6 \times 10^4$)) achieving a noise power measured 6.3 MHz from, and relative to, $\omega_{\rm p}$ of less than $-195\,{\rm dB}_c\,{\rm Hz}^{-1}$ (where ${\rm dB}_c$ denotes units of noise power measured relative to the pump power) and contributing less than 0.04 photons into the SMR at our highest value of $\bar{n}_{\rm p}$. Without these cavities, the SMR would be excited to $\bar{n}_{\rm SMR} = 35$. We also believe that this excess SMR occupation is not due to ohmic heating of, and resulting thermal radiation from, the cryogenic attenuator network because $\bar{n}_{\rm SMR}$ increases only weakly over a wide range of $\bar{n}_{\rm p}$ values. Tests of Nb SMR devices at 1.2 K before the surface micromachining of the nanomechanical resonator do not show excess dissipation and suggest that the excess losses are related to our fabrication process.

Increasing $\Gamma_{\rm opt}$ by engineering a larger coupling strength, g, and/or a smaller κ value should be very beneficial as it will lead to higher cooling rates at lower pump powers, minimizing the effect of excess bath heating, $\dot{n}_T$. By increasing $\Gamma_{\rm opt}$ by a factor of ten, maintaining the same value of $\dot{n}_T$, we expect to obtain $\bar{n}_m \approx 0.5$ with $P_0 = 0.67$. This approach will be limited when $\Gamma_{\rm opt}$ becomes comparable to κ, which constrains the rate of cooling[19,25].

The deep quantum limit, $\bar{n}_m \ll 1$, will be accessible when it is possible to reach lower refrigerator temperatures and lower mechanical

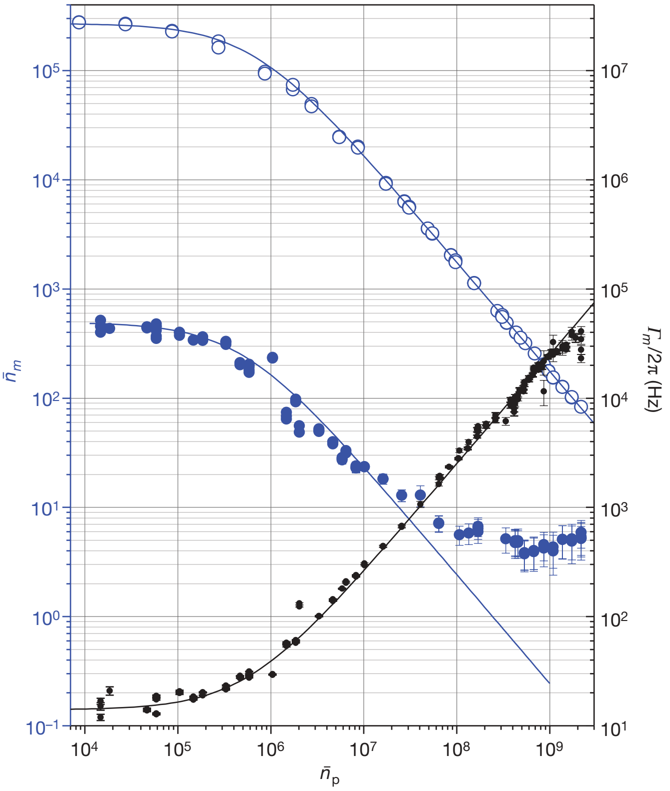


Figure 4 | Mechanical linewidth broadening and cooling versus pump strength. Mechanical occupation factor, $\bar{n}_m$ (blue), and total mechanical linewidth, $\Gamma_m = \Gamma_m^T + \Gamma_{\rm opt}$ (black), versus drive photon occupation, $\bar{n}_{\rm p}$. The lower cooling curve starts from a refrigerator temperature of 146 mK, and the upper curve starts from an effective temperature of 80 K, which is generated by applying an electrostatic force noise to the nanomechanical resonator. The solid black curve is a fit to the measured Γ_m values. The solid blue curves are the expected values of $\bar{n}_m$ assuming ideal values of $\bar{n}_{\rm SMR}$ and $\dot{n}_T = 4 \times 10^4$ quanta per second (filled points) and $\dot{n}_T = 2 \times 10^6$ quanta per second (open points). The error bars on Γ_m arise from the statistical fluctuations in our measured noise power and the resulting standard error in the Lorentzian fit parameters. The error bars on $\bar{n}_m$ (s.e.m.) include the same source of statistical error, as well as the uncertainty in SMR occupation, which is calculated from best estimates of uncertainties in line loss and amplifier gain (Supplementary Information).

damping rates at these temperatures. Understanding and eliminating the excess bath heating and the non-thermal force noise will be required, although it should be pointed out that, even without these improvements, the device described here should achieve $\bar{n}_m < 0.5$ if excess SMR occupation is reduced. Furthermore, superconducting metals on the nanomechanical resonator also appear to be required owing to the expected mechanical force noise from transport and electron momentum scattering in diffusive conductors[26]. For a normal-state conductor, each electron scattering event imparts a momentum change on the order of the Fermi momentum. The diffusive scattering of current through the beam directly produces force noise on the nanomechanical resonator. We estimate that this heating mechanism will result in a limit of $\bar{n}_m > 3$ at $\bar{n}_\mathrm{p} = 3 \times 10^8$, assuming our current device parameters and a resistance of $100\,\Omega$ through the nanomechanical resonator.

These measurements show that detection with sensitivity to resolve motions approaching the ground state is possible with existing HEMT-based amplifiers. Eliminating internal SMR losses, unbalancing the SMR couplings and using improved microwave amplifiers[27] would significantly reduce the measurement time.

Nonetheless, the production and detection of a nanomechanical resonator with $\bar{n}_m = 3.8$ is sufficient to allow future experiments. Owing to uncertainty-principle fluctuations of the mechanical motion, and resulting spontaneous emission, the rate of microwave-photon up-conversion is expected to differ from the rate of down-conversion. This difference can be used as a fundamental thermometry technique[7,8,18], and would allow the quantitative measurement of the zero-point motion of a mechanical structure.

This level of cooling is essential for the formation of entangled states between superconducting quantum bits and the motion of a nanomechanical device. Proposed schemes[1,28] allow the generation and detection of nanomechanical resonator/qubit entanglement using a Jaynes–Cummings-type interaction with the mechanical resonator at thermal occupations of the level shown here. Similar to procedures in atomic physics, such an experiment would involve preparing the cold state of the mechanical device and, after the refrigeration is complete, turning the cooling off. The state of the cold beam could then be manipulated before thermalization of the motion. In our implementation of this process, we expect cooling from $\bar{n}_m = 500$ to $\bar{n}_m = 4$ in $\sim 200\,\mu\mathrm{s}$, and we expect one thermal quantum to enter the resonator in $\tau = 1/\dot{n}_T = 2\,\mu\mathrm{s}$, which exceeds superconducting qubit manipulation times and is comparable to qubit measurement and relaxation times.

Received 13 August; accepted 18 November 2009.
Published online 9 December 2009.

1. Armour, A., Blencowe, M. & Schwab, K. Entanglement and decoherence of a micromechanical resonator via coupling to a Cooper-pair box. *Phys. Rev. Lett.* **88,** 148301 (2002).
2. Marshall, W., Simon, C., Penrose, R. & Bouwmeester, D. Towards quantum superposition of a mirror. *Phys. Rev. Lett.* **91,** 130401 (2003).
3. Naik, A. *et al.* Cooling a nanomechanical resonator with quantum back-action. *Nature* **443,** 193–196 (2006).
4. Courty, J. M., Heidmann, A. & Pinard, M. Quantum limits of cold damping with optomechanical coupling. *Eur. Phys. J. D* **17,** 399–408 (2001).
5. Martin, I., Shnirman, A., Tian, L. & Zoller, P. Ground-state cooling of mechanical resonators. *Phys. Rev. B* **69,** 125339 (2004).
6. Blencowe, M. P. & Buks, E. Quantum analysis of a linear dc SQUID mechanical displacement detector. *Phys. Rev. B* **76,** 014511 (2007).
7. Marquardt, F., Chen, J. P., Clerk, A. A. & Girvin, S. M. Quantum theory of cavity-assisted sideband cooling of mechanical motion. *Phys. Rev. Lett.* **99,** 093902 (2007).
8. Wilson-Rae, I., Nooshi, N., Zwerger, W. & Kippenberg, T. J. Theory of ground state cooling of a mechanical oscillator using dynamical backaction. *Phys. Rev. Lett.* **99,** 093901 (2007).
9. Schliesser, A., Arcizet, O., Riviere, R., Anetsberger, G. & Kippenberg, T. J. Resolved-sideband cooling and position measurement of a micromechanical oscillator close to the Heisenberg uncertainty limit. *Nature Phys.* **5,** 509–514 (2009).
10. Park, Y.-S. & Wang, H. Resolved-sideband and cryogenic cooling of an optomechanical resonator. *Nature Phys.* **5,** 489–493 (2009).
11. Groblacher, S. *et al.* Demonstration of an ultracold micro-optomechanical oscillator in a cryogenic cavity. *Nature Phys.* **5,** 485–488 (2009).
12. Day, P. K., LeDuc, H. G., Mazin, B. A., Vayonakis, A. & Zmuidzinas, J. A broadband superconducting detector suitable for use in large arrays. *Nature* **425,** 817–821 (2003).
13. Regal, C. A., Teufel, J. D. & Lehnert, K. W. Measuring nanomechanical motion with a microwave cavity interferometer. *Nature Phys.* **4,** 555–560 (2008).
14. Cleland, A. N. & Roukes, M. L. A nanometre-scale mechanical electrometer. *Nature* **392,** 160–162 (1998).
15. Dykman, M. I. Heating and cooling of local and quasilocal vibrations by nonresonant eld. *Sov. Phys. Solid State* **20,** 1306 (1978).
16. Linthorne, N. P., Veitch, P. J. & Blair, D. G. Interaction of a parametric transducer with a resonant bar gravitational radiation detector. *J. Phys. D* **23,** 1–6 (1990).
17. Xue, F., Wang, Y. D., Liu, Y.-X. & Nori, F. Cooling a micromechanical beam by coupling it to a transmission line. *Phys. Rev. B* **76,** 205302 (2007).
18. Diedrich, F., Bergquist, J. C., Itano, W. & Wineland, D. J. Laser cooling to the zero-point energy of motion. *Phys. Rev. Lett.* **62,** 403–406 (1989).
19. Dobrindt, J. M., Wilson-Rae, I. & Kippenberg, T. J. Parametric normal-mode splitting in cavity optomechanics. *Phys. Rev. Lett.* **101,** 263602 (2008).
20. Blair, D. G. *et al.* High sensitivity gravitational wave antenna with parametric transducer readout. *Phys. Rev. Lett.* **74,** 1908–1911 (1995).
21. Teufel, J. D., Harlow, J. W., Regal, C. A. & Lehnert, K. W. Dynamical backaction of microwave fields on a nanomechanical oscillator. *Phys. Rev. Lett.* **101,** 197203 (2008).
22. Teufel, J. D., Regal, C. A. & Lehnert, K. W. Prospects for cooling nanomechanical motion by coupling to a superconducting microwave resonator. *New J. Phys.* **10,** 095002 (2008).
23. Stipe, B. C., Mamin, H. J., Stowe, T. D., Kenny, T. W. & Rugar, D. Noncontact friction and force fluctuations between closely spaced bodies. *Phys. Rev. Lett.* **87,** 096801 (2001).
24. Poggio, M., Degen, C. L., Mamin, H. J. & Rugar, D. Feedback cooling of a cantilever's fundamental mode below 5mk. *Phys. Rev. Lett.* **99,** 017201 (2007).
25. Grajcar, M., Ashhab, S., Johansson, J. R. & Nori, F. Lower limit on the achievable temperature in resonator-based sideband cooling. *Phys. Rev. B* **78,** 035406 (2008).
26. Shytov, A. V., Levitov, L. S. & Beenakker, C. W. J. Electromechanical noise in a diffusive conductor. *Phys. Rev. Lett.* **88,** 228303 (2002).
27. Castellanos-Beltran, M. A., Irwin, K. D., Hilton, G. C., Vale, L. R. & Lehnert, K. W. Amplification and squeezing of quantum noise with a tunable Josephson metamaterial. *Nature Phys.* **4,** 929–931 (2008).
28. Utami, D. W. & Clerk, A. A. Entanglement dynamics in a dispersively coupled qubit-oscillator system. *Phys. Rev. A* **78,** 042323 (2008).
29. Verbridge, S. S., Craighead, H. G. & Parpia, J. M. A megahertz nanomechanical resonator with room temperature quality factor over a million. *Appl. Phys. Lett.* **92,** 013112 (2008).
30. Segev, E., Abdo, B., Shtempluck, O. & Buks, E. Thermal instability and self-sustained modulation in superconducting NbN stripline resonators. *J. Phys. Condens. Matter* **19,** 096206 (2007).

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

Acknowledgements We acknowledge conversations with M. Blencowe, M. Aspelmeyer, R. Ilic, M. Skvarla, M. Metzler and M. Shaw and assistance from M. Savva, S. Rosenthal and M. Corbett. This work has been supported by the Fundamental Questions Institute (http://fqxi.org) (RFP2-08-27) and the US National Science Foundation (NSF) (DMR-0804567). Device fabrication was performed at the Cornell Nanoscale Facility, a member of the US National Nanotechnology Infrastructure Network (NSF grant ECS-0335765).

Author Contributions T.R. and T.N. contributed equally to device fabrication and measurements. C.M. built key apparatus and assisted in experimental set-up. J.B.H. assisted in planning and analysis. A.A.C. provided theoretical analysis. K.C.S. initiated and oversaw the work.

Author Information 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 K.C.S (schwab@caltech.edu).

Vol 463 | 7 January 2010 | doi:10.1038/nature08676

LETTERS

Slip-stick and the evolution of frictional strength

Oded Ben-David[1], Shmuel M. Rubinstein[1]† & Jay Fineberg[1]

The evolution of frictional strength has great fundamental and practical importance. Applications range from earthquake dynamics[1–4] to hard-drive read/write cycles[5]. Frictional strength is governed by the resistance to shear of the large ensemble of discrete contacts that forms the interface that separates two sliding bodies. An interface's overall strength is determined by both the real contact area and the contacts' shear strength[6,7]. Whereas the average motion of large, slowly sliding bodies is well-described by empirical friction laws[3,8–10], interface strength is a dynamic entity that is inherently related to both fast processes such as detachment/re-attachment[11–14] and the slow process of contact area rejuvenation[6,7,13,15,16]. Here we show how frictional strength evolves from extremely short to long timescales, by continuous measurements of the concurrent local evolution of the real contact area and the corresponding interface motion (slip) from the first microseconds when contact detachment occurs to large (100-second) timescales. We identify four distinct and inter-related phases of evolution. First, all of the local contact area reduction occurs within a few microseconds, on the passage of a crack-like front. This is followed by the onset of rapid slip over a characteristic time, the value of which suggests a fracture-induced reduction of contact strength before any slip occurs. This rapid slip phase culminates with a sharp transition to slip at velocities an order of magnitude slower. At slip arrest, 'ageing' immediately commences as contact area increases on a characteristic timescale determined by the system's local memory of its effective contact time before slip arrest. We show how the singular logarithmic behaviour generally associated with ageing is cut off at short times[16]. These results provide a comprehensive picture of how frictional strength evolves from the short times and rapid slip velocities at the onset of motion to ageing at the long times following slip arrest.

Although frictional motion is often conceptually viewed as the motion of two rigid bodies in perfect contact at a planar interface, in fact, the applied normal load F_N is supported by an ensemble of micro-contacts that comprise only a small fraction of the apparent contact area. The real contact area A of these micro-contacts determines the interface strength[6,7]. The interface's properties differ entirely from those of the surrounding elastic material[17], because it bears local pressures that approach the material yield strength. How these contacts detach and re-attach is central to understanding frictional motion.

Frictional slip[18,19] initiates rapidly, through the fracture of contacts. A number of different fracture mechanisms[11,12,14,20] have recently been reported involving contact area reduction via crack-like "detachment fronts"[12] preceding frictional motion. When the applied shear force F_S is not uniformly distributed along the interface, arrested (precursory) detachment fronts[20] are excited for F_S well below the threshold for macroscopic frictional motion at $F_S = \mu_S F_N$ (where μ_S is the static friction coefficient).

Upon cessation of motion, contacts re-form and strengthen. In many materials, interface strengthening is attributed to growth of A through contact rejuvenation[6,7,13] as frictional strength (embodied in μ_S) increases logarithmically with the time of static contact[8,9,21,22]. These slow changes, referred to as 'ageing', are common to a class of 'glassy' physical systems that are characterized by the existence of a large number of meta-stable configurations in which a system can dwell, while slowly relaxing to states of lower energy[15]. The way in which a frictional system evolves to this asymptotic logarithmic dependence, while regularizing the apparent singularity of μ_S at short times has been, until now, experimentally inaccessible. Here we examine how the local values of the real contact area $A(x, t)$ and concurrent slip $\delta(x, t)$ are related throughout detachment, slip and rehealing. These measurements provide an inclusive picture of processes that occur within frictional slip.

Our experimental system is schematically shown in Fig. 1a. Two poly(methyl methacrylate) (PMMA) blocks form a long (200 mm) and thin (6 mm) interface of approximately 1 μm root-mean-square roughness. The blocks are initially pressed together by a fixed, spatially uniform normal force F_N. F_S is applied to the upper block's 'trailing' edge at $x = 0$. $A(x, t)$ is measured[12,13] (see Supplementary Information) by illuminating the interface area with a laser sheet whose incident angle is well beyond the angle for total internal reflection from non-contacting parts of the interface. Under these conditions, the light intensity transmitted at each point x along the interface is proportional to $A(x, t)$. This light is imaged by a fast camera, capturing changes in $A(x, t)$ that occur over timescales ranging from 10^{-6} s to 10^2 s (each of the camera's 1,280 pixels is mapped to a 0.2 mm × 0.8 mm region of the interface). In parallel to contact measurements, concurrent high-resolution displacement measurements are performed at specific points X adjacent to the interface.

When F_S is slowly increased, the spatially non-uniform shear stress distribution along the interface gives rise to a sequence (Fig. 1b) of arrested (precursory) detachment fronts[20]. Each successive front in the sequence initiates at the sample's trailing edge ($x = 0$), and propagates an increased distance before arresting within the interface. At every point along its path, each detachment front both generates a sharp reduction in $A(x, t)$ and initiates slip $\delta(x, t)$. No overall motion of the blocks occurs[20] until the fronts reach the leading edge ($x = L$). A close look (for example, Fig. 1c) reveals that detachment fronts are extremely rapid, with velocities that approach the Rayleigh-wave speed c_R (1,280 m s^{-1}). These high speeds suggest that the processes of detachment, slip and subsequent rehealing of the interface must initiate at extremely short (microsecond) timescales.

Arrested detachment fronts provide a sharp, well-defined initiation point for the study of the concurrent evolution of the surface area and slip. Acoustic signals initiated by each front were used both to trigger data storage and to suspend F_S for fixed periods (5–1,000 s) to study the resulting contact rejuvenation. In Fig. 2a we present simultaneous measurements of $A(t) \equiv A(X, t)$ and $\delta(t) \equiv \delta(X, t)$ at short timescales within the 500 μs that bracket the passage of a detachment front at point X. These measurements reveal three distinct dynamic phases:

[1]The Racah Institute of Physics, The Hebrew University of Jerusalem, Givat Ram, Jerusalem 91904, Israel. †Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

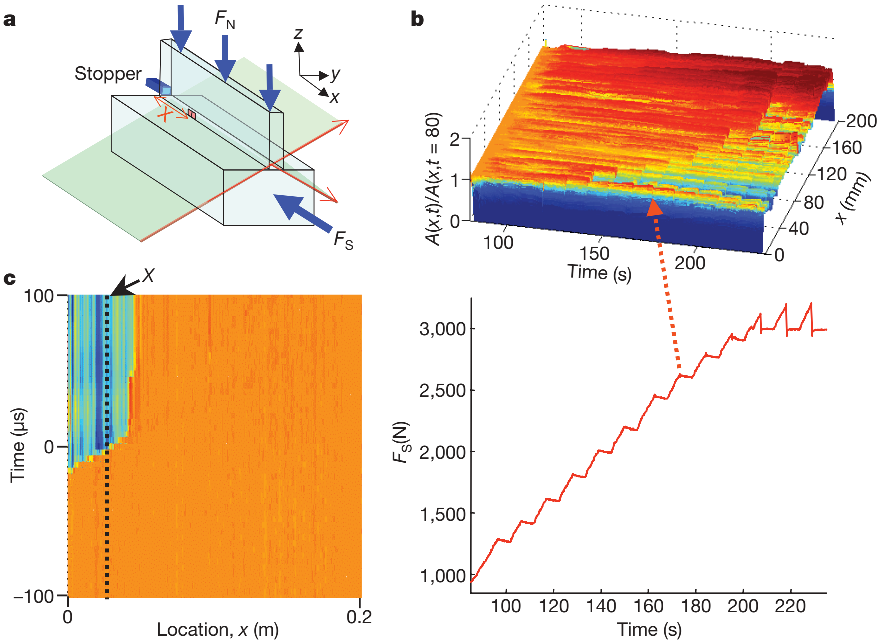


Figure 1 | Rapid precursors to frictional motion enable the precise study of local dynamics. a, Schematic diagram of the experimental system (see Supplementary Information). A uniform normal load F_N is applied to two PMMA blocks (base and top blocks). Shear force F_S is applied in the x direction to the base block, which is mounted on a low-friction stage. F_S is countered by a rigid stopper at the slider's $x = 0$ edge. The real area of contact $A(x, t)$ throughout the contact plane is imaged at 4-μs intervals. Simultaneous measurements of slip both at a distance X from the stopper and at the top block's leading edge were made. **b,** The non-spatially uniform shear loading generates a sequence of rapid precursors before the onset of overall motion, which initiate at $x = 0$ and arrest at successively larger distances within the interface. The top panel shows $A(x, t)$ normalized by $A(x, t = 80\text{ s})$ taken before the first precursor. Hotter colours (reds) denote increased $A(x, t)$. Colder colours (blues) denote reduced $A(x, t)$. Each precursor significantly reduces $A(x, t)$ along its length. The bottom panel shows the corresponding loading curve $F_S(t)$. On acoustic detection of each precursor, the continuous increase of F_S in time is paused for 5–1,000-s intervals. Slip, generated by each precursor at $x = 0$, gives rise to small sharp drops in F_S that are detectable due to the compliance of the loading system. Here, $F_N = 6{,}000$ N and $\mu_S = 0.51$, and the origin of t was the time at which F_S was applied. **c,** Short-time measurements of $A(x, t)$ (at 4-μs intervals) reveal that each precursor is a detachment front propagating along the interface at velocities v approaching the Rayleigh-wave speed $c_R = 1{,}280\text{ m s}^{-1}$. Here $v = 1{,}200 \pm 100\text{ m s}^{-1}$. The dotted line denotes the location X where slip $\delta(t)$ was measured. Here, $A(x, t)$ was normalized by its value at 1 ms before the front's passage.

detachment (phase I), rapid slip (phase II) and slow slip (phase III). The entire net reduction of $A(t)$ occurs in phase I, immediately upon the passage of the detachment front (top panel of Fig. 2b). Simultaneous measurements of $\delta(t)$ reveal that during this 20% drop of $A(t)$, no net slip takes place at all. As shown in the lower panel of Fig. 2b, the drop in $A(t)$ is always preceded by strong fluctuations of $\delta(t)$ of duration approximately 10 μs.

Only upon conclusion of the detachment phase does net slip in the system initiate. In Fig. 2c we present a superposition of 16 different experiments in which $\delta(t)$, normalized by the total slip of each event δ_{tot} is plotted. The rough collapse of the data (where δ_{tot} varies between 4 μm and 20 μm) reveals generic dynamics. Two distinct slip phases exist; a rapid slip phase (phase II) that commences immediately after detachment, followed by a slow slip phase (phase III). A number of characteristic features are apparent in Fig. 2c. First, in both phases II and III, slip occurs at roughly constant slip velocities $\dot{\delta}$, the values of which differ by over an order of magnitude. Second, there is a characteristic duration time for each phase and the transition between them is sharp. The data collapse in Fig. 2c is due both to the existence of this characteristic time and to a roughly constant proportion of the total amount of slip within each phase. These features are independent of the magnitude of δ_{tot}, the measurement location X, the relative location between the front arrest point and X, details of loading and the geometry of the blocks (see Supplementary Information).

Phase II is characterized by its high slip velocities, ranging from 5 cm s^{-1} to over 30 cm s^{-1}, with variations of $A(t)$ that are an order of magnitude smaller than those in phase I. The duration of this phase is a surprisingly constant 60 μs. This timescale is evident from both the transition point between phases II and III in Fig. 2c and the constant value (60 ± 6 μs) of the slope of the slip δ_{rapid} versus the slip velocity V_{rapid} in this phase, as presented in the top panel of Fig. 2d. We observe these features at all locations traversed by rapid $(0.8 < v/c_R < 1)$ detachment fronts.

In phase III the slip is about 30% of δ_{tot}. $\dot{\delta}$ is much slower (in the range $0.1\text{–}2\text{ cm s}^{-1}$) than in phase II, as shown in the bottom panel of Fig. 2d. The slip duration, however, is much longer (typically 350 μs). In this phase, $A(t)$ is constant in time with fluctuations of less than 1%.

Phase III concludes with the cessation of slip. This point marks the initiation of the contact rejuvenation and strengthening that are generally associated with 'ageing' of μ_S. Whereas the short-time (that is, less than a millisecond) evolution of μ_S is nearly impossible to measure directly, we can access the short-time evolution of frictional strength by measuring the growth of $A(t)$ from the time $t = t_0$, when slip ceases. At large times ageing is logarithmic. This growth must, however, be regularized at short times, because frictional strength is finite at the onset of growth. How this regularization occurs is unknown, because the short-time growth of neither μ_S nor $A(t)$ has ever been measured. Both the short- and long-time asymptotic behaviours can, in general, be characterized by the functional form: $A(t) = A(t_0) + b\log[f(t/\tau)]$, where $f(t = t_0) = 1$ and $f(t) \approx t^{\alpha}$ for $t \gg \tau$ (where α is any number), and τ is a characteristic timescale that renders t dimensionless. The simplest such function[21,23] is:

$$A(t) = A(t_0)[1 + \beta\log(1 + (t - t_0)/\tau)] \qquad (1)$$

In Fig. 3a we describe the evolution of $A(t)$ over six orders of magnitude of temporal scales. In this region of growth, denoted phase IV, we observe logarithmic growth at long timescales. Figure 3a demonstrates that equation (1) provides an excellent characterization of $A(t)$ over the entire measurement range and shows

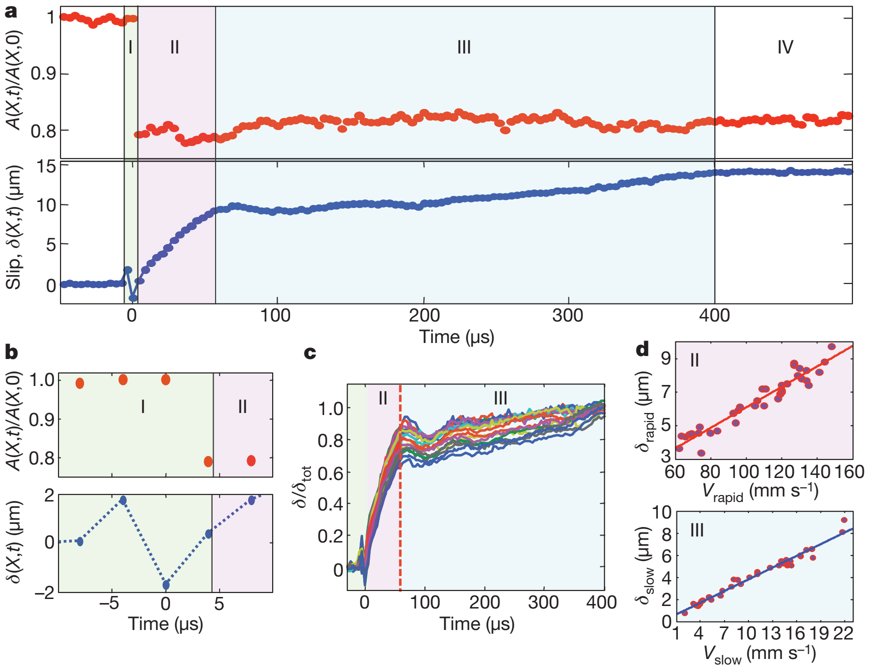


Figure 2 | The detachment process and the evolution of frictional slip. **a**, Simultaneous measurements of the local dynamics of contact area $A(t) \equiv A(X, t)$ (top panel) and slip $\delta(t) \equiv \delta(X, t)$ (bottom panel) before, during and after the passage of a precursor event. The measurements reveal three distinct initial phases of the dynamics: detachment (phase I) at $t = 0$, rapid slip (phase II) and slow slip (phase III). Notice that, whereas $A(t)$ drops by 20% in phase I, it remains nearly constant within phases II and III. **b**, The detachment phase (top panel), is labelled as phase I. No net slip (bottom panel) is observed before contact reduction. Nearly all reduction of $A(t)$ occurs within our 4-μs temporal resolution. The drop in $A(t)$ is concurrent with the conclusion of a short slip fluctuation. **c**, Normalizing $\delta(t)$ by the total slip in an event δ_{tot} yields an approximate collapse of the slip history. 16 events are plotted. Rapid slip (phase II), at roughly constant velocity V_{rapid}, initiates immediately upon the conclusion of the detachment phase. A sharp transition (denoted by the red dashed line) to velocities an order of magnitude lower V_{slow} is apparent (phase III). The approximate collapse indicates that both the relative part of the slip and the duration of each phase are approximately constant. **d**, The top panel shows the slope of the slip during rapid phase δ_{rapid} as a function of the slip velocity V_{rapid}, revealing the remarkably constant 60-μs duration of phase II. The bottom panel shows a similar plot of the slip δ_{slow} as a function of its corresponding velocity V_{slow}, yielding a typical duration of 350 μs for phase III. Solid lines depict linear fits to the data.

how $A(t)$ is regularized for short timescales. We note that the growth depicted in Fig. 3 does not simultaneously occur at every point along the interface, but initiates locally at each point upon cessation of slip.

Any deviation from equation (1) would be expected at times when $0 < (t - t_0)/\tau < 10$. In Fig. 3b, we plot $A(t)$ for a number of typical experiments in this range. Remarkably, the data all collapse and no systematic deviations from equation (1) are evident. The ageing rate β_S of the static friction coefficient, $\mu_S = \mu_S{}^0 + \beta_S \log(t)$, is given by[7] $\beta_S = \beta\mu_S$. The average value derived from our experiments of $\beta_S = 0.009 \pm 0.001$ is compatible with measurements in bulk frictional motion[21].

The timescale τ in equation (1) characterizes the growth rate of A, until $(t - t_0)/\tau \gg 1$. What determines this scale? In Fig. 3c we plot the value of τ as a function of the slip velocity, $\dot{\delta} = V_{\text{slow}}$, in phase III, which immediately preceded the onset of growth of $A(t)$ at t_0. We find that τ is inversely proportional to V_{slow}. A linear fit $\tau = D/V_{\text{slow}}$ (see Fig. 3c) yields $D = 0.9$ μm, a value corresponding to the characteristic size of the asperities in our experiments. We therefore interpret τ to be the effective contact time of the micro-contact population at the onset of ageing.

We now consider the overall evolution of the system, as described by phases I to IV. Two rough surfaces are 'bound' together by interlocking protrusions (asperities) from each surface. For any motion to occur, these asperities must circumvent one another, either by fracture, internal damage or deformation. Each process involves a significant energy cost that should scale with $A(x, t)$. Thus, the onset of slip involves a 'fracture energy' Γ, defined as the energy per unit area needed to detach two contact surfaces. In tensile fracture of PMMA, Γ is dominated by (irreversible) plastic deformation before bond rupture. We assume that similar plastic deformation of interlocking asperities is needed to enable motion.

In phase I the 'instantaneous' drop of $A(t)$ indicates that the contact fracture takes place within the short (about a microsecond) passage time of a detachment front. This drop is immediately preceded by strong fluctuations of $\delta(t)$ (see Fig. 2b), the magnitude of which approaches asperity sizes. We therefore surmise that this is when fracture occurs.

The heat ΓA released by this rapid irreversible deformation is deposited along the interface layer before the net slip. Such a rapid heat injection will cause an 'immediate' large temperature increase ΔT within the thin interface layer, akin to the large ($\Delta T \approx 1{,}000$ °C) temperature rises observed in the tensile fracture of PMMA[24]. Any increase beyond the glass temperature T_g (~110 °C in PMMA[10]) will cause significant shear strength weakening until the heat generated by fracture diffuses away from the interface region. An estimate of Γ (see Supplementary Information) yields cooling times compatible with the observed 60 μs duration of phase II.

An interface's overall strength is determined by both the value of A and the contacts' shear strength. A does not vary significantly between phases II and III, so the rapid slip in phase II clearly suggests contact strength weakening. Although conceptually similar to friction reduction attributed[25–27] to 'flash heating', fracture-generated weakening requires no sustained net slip to reduce shear strength. The thermally induced weakening here is sustained only for a finite time before cooling occurs. Although our interpretations of phases II and III are strictly relevant for glassy materials, it is interesting that the form of the slip profiles and the velocities measured in phases II and III resemble slip within granite blocks[18,19] during propagation of rapid crack-like fronts. This may suggest a similar fracture-induced weakening mechanism in brittle materials such as rock, where weakening via the crushing of interlocking asperities could supplant the thermal softening of PMMA.

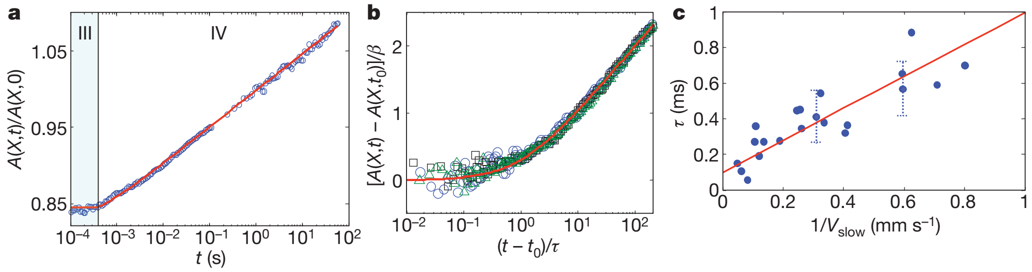


Figure 3 | Cessation of slip and interface ageing. **a**, $A(t)$ is plotted over six orders of magnitude of time, where $t = 0$ marks the drop in $A(t)$ within phase I. The transition from the constant value of $A(t)$ within phase III to purely logarithmic ageing is described by the red line, $A(t) = A(t_0)[1 + \beta\log(1 + (t - t_0)/\tau)]$, where t_0 is the time of slip arrest. **b**, A high-resolution plot of $[A(t) - A(t_0)]/\beta$ as a function of the scaled time $(t - t_0)/\tau$ for three typical experiments with $\tau = 310\ \mu\text{s}$ (squares), $450\ \mu\text{s}$ (circles) and $470\ \mu\text{s}$ (triangles). The excellent superposition indicates that all are described by the same functional form. The function $\log[1 + (t - t_0)/\tau]$ (red line) describes the data with no visible systematic deviations. **c**, The characteristic timescale τ as a function of $1/V_{\text{slow}}$. The data suggest a linear dependence of $\tau = D/V_{\text{slow}}$, where a fit yields a length scale $D \approx 0.9\ \mu\text{m}$, consistent with the characteristic scale of the asperities in our experiments. Thus, τ represents the characteristic timescale of the system's memory and may be interpreted as the effective asperity contact time during phase III. Error bars represent the cumulative error in each measurement (for details, see Supplementary Information).

In glasses, when the softened layer cools to below T_g, contact strengthening will occur and the system will undergo a sharp transition to slow slip, precisely as observed in phase III. The local motion in this phase may be analogous to large-scale frictional motion[8,9] governed by the contact dynamics of 'rigid' (unsoftened) microcontacts. Here, 'collective' heating due to fracture is absent, because slip is accomplished by sporadic rupture of discrete, loosely coupled contacts. The interface resembles a glassy system in which the state of each contact is coupled only to that of its neighbours through the residual shear stress that remained at the transition from phase II.

In phase IV equation (1) both shows how logarithmic ageing is regularized and relates the early time growth rate τ to the slip history immediately before slip cessation. Precisely this behaviour was suggested[16] in recent interpretations of experiments describing granite blocks sliding over a ground quartz (gouge) layer[22]. V_{slow} is an ensemble average over numerous micro-contacts, so τ reflects the collective behaviour or 'state' of the entire ensemble. In our measurements, V_{slow} and thus τ are spatially local, and suggest that the system's state is a local 'coarse-grained' quantity that represents an ensemble average of the contact population at each spatial location. It is significant that τ is determined by the slip rate in phase III. Thus, the overall dynamics of ageing retains a memory of the system's state at the arrest of slip.

These results provide fundamental new insights into how frictional strength evolves throughout the slip–stick cycle. They suggest that a mesoscopic description of friction that incorporates the spatial dependence of the system's state may be necessary when describing dynamics at either the intermediate timescales within phases III and IV, or in large spatially extended systems.

Received 27 August; accepted 10 November 2009.

1. Scholz, C. H. Earthquakes and friction laws. *Nature* **391,** 37–42 (1998).
2. Dieterich, J. H. Earthquake nucleation on faults with rate-dependent and state-dependent strength. *Tectonophysics* **211,** 115–134 (1992).
3. Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* **26,** 643–696 (1998).
4. Ben-Zion, Y. Collective behavior of earthquakes and faults: continuum-discrete transitions, progressive evolutionary changes, and different dynamic regimes. *Rev. Geophys.* **46,** 1–70 (2008).
5. Urbakh, M., Klafter, J., Gourdon, D. & Israelachvili, J. The nonlinear nature of friction. *Nature* **430,** 525–528 (2004).
6. Dieterich, J. H. & Kilgore, B. D. Direct observation of frictional contacts—new insights for state-dependent properties. *Pure Appl. Geophys.* **143,** 283–302 (1994).
7. Bowden, F. P. & Tabor, D. *The Friction and Lubrication of Solids* Ch. 1 (Oxford Univ. Press, 2001).
8. Dieterich, J. H. Modeling of rock friction. 1. Experimental results and constitutive equations. *J. Geophys. Res.* **84,** 2161–2168 (1979).
9. Rice, J. R. & Ruina, A. L. Stability of steady frictional slipping. *J. Appl. Mech.* **50,** 343–349 (1983).
10. Baumberger, T., Berthoud, P. & Caroli, C. Physical analysis of the state- and rate-dependent friction law. II. Dynamic friction. *Phys. Rev. B* **60,** 3928–3939 (1999).
11. Ben-Zion, Y. Dynamic ruptures in recent models of earthquake faults. *J. Mech. Phys. Solids* **49,** 2209–2244 (2001).
12. Rubinstein, S. M., Cohen, G. & Fineberg, J. Detachment fronts and the onset of dynamic friction. *Nature* **430,** 1005–1009 (2004).
13. Rubinstein, S., Cohen, G. & Fineberg, J. Contact area measurements reveal loading-history dependence of static friction. *Phys. Rev. Lett.* **96,** 256103 (2006).
14. Xia, K., Rosakis, A. J. & Kanamori, H. Laboratory earthquakes: the sub-Raleigh-to-supershear rupture transition. *Science* **303,** 1859–1861 (2004).
15. Rottler, J. & Robbins, M. O. Unified description of aging and rate effects in yield of glassy solids. *Phys. Rev. Lett.* **95,** 225504 (2005).
16. Nakatani, M. & Scholz, C. H. Intrinsic and apparent short-time limits for fault healing: theory, observations, and implications for velocity-dependent frictionl. *J. Geophys. Res. Solid Earth* **111,** B12208 (2006).
17. Kim, H. J., Kim, W. K., Falk, M. L. & Rigney, D. A. MD simulations of microstructure evolution during high-velocity sliding between crystalline materials. *Tribol. Lett.* **28,** 299–306 (2007).
18. Ohnaka, M. & Kuwahara, Y. Characteristic features of local breakdown near a crack-tip in the transition zone from nucleation to unstable rupture during stick-slip shear failure. *Tectonophysics* **175,** 197–220 (1990).
19. Okubo, P. G. & Dieterich, J. H. Effects of physical fault properties on frictional instabilities produced on simulated faults. *J. Geophys. Res.* **89,** 5817–5827 (1984).
20. Rubinstein, S. M., Cohen, G. & Fineberg, J. Dynamics of precursors to frictional sliding. *Phys. Rev. Lett.* **98,** 226103 (2007).
21. Berthoud, P. & Baumberger, T. G'Sell, C. & Hiver, J. M. Physical analysis of the state- and rate-dependent friction law: static friction. *Phys. Rev. B* **59,** 14313–14327 (1999).
22. Marone, C. The effect of loading rate on static friction and the rate of fault healing during the earthquake cycle. *Nature* **391,** 69–72 (1998).
23. Dieterich, J. H. Time-dependent friction in rocks. *J. Geophys. Res.* **77,** 3690–3697 (1972).
24. Fuller, K. N. G., Fox, P. G. & Field, J. E. Temperature rise at tip of fast-moving cracks in glassy polymers. *Proc. R. Soc. Lond. A* **341,** 537 (1975).
25. Rice, J. R. Heating and weakening of faults during earthquake slip. *J. Geophys. Res. Solid Earth* **111,** B05311 (2006).
26. Nielsen, S., Di Toro, G., Hirose, T. & Shimamoto, T. Frictional melt and seismic slip. *J. Geophys. Res. Solid Earth* **113,** B01308 (2008).
27. Beeler, N. M., Tullis, T. E. & Goldsby, D. L. Constitutive relationships and physical basis of fault strength due to flash heating. *J. Geophys. Res. Solid Earth* **113,** B01401 (2008).

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

Acknowledgements We acknowledge comments by J. R. Rice, M. O. Robbins, T. Baumberger and C. Caroli. This work, as part of the ESF EUROCORES programme FANAS, was supported by the Israel Science Foundation (grant 57/07). We also acknowledge support of the US–Israel Binational Science Foundation (grant 2006288).

Author Contributions All authors contributed to the design of the experiment. O.B.-D. performed the measurements. All authors contributed to the analysis and writing of the manuscript.

Author Information 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 J.F. (jay@vms.huji.ac.il).

LETTERS

Anthropogenic carbon dioxide transport in the Southern Ocean driven by Ekman flow

T. Ito[1], M. Woloszyn[1] & M. Mazloff[2]

The Southern Ocean, with its large surface area and vigorous overturning circulation, is potentially a substantial sink of anthropogenic CO_2 (refs 1–4). Despite its importance, the mechanism and pathways of anthropogenic CO_2 uptake and transport are poorly understood. Regulation of the Southern Ocean carbon sink by the wind-driven Ekman flow, mesoscale eddies and their interaction is under debate[5–8]. Here we use a high-resolution ocean circulation and carbon cycle model to address the mechanisms controlling the Southern Ocean sink of anthropogenic CO_2. The focus of our study is on the intra-annual variability in anthropogenic CO_2 over a two-year time period. We show that the pattern of carbon uptake is correlated with the oceanic vertical exchange. Zonally integrated carbon uptake peaks at the Antarctic polar front. The carbon is then advected away from the uptake regions by the circulation of the Southern Ocean, which is controlled by the interplay among Ekman flow, ocean eddies and subduction of water masses. Although lateral carbon fluxes are locally dominated by the imprint of mesoscale eddies, the Ekman transport is the primary mechanism for the zonally integrated, cross-frontal transport of anthropogenic CO_2. Intra-annual variability of the cross-frontal transport is dominated by the Ekman flow with little compensation from eddies. A budget analysis in the density coordinate highlights the importance of wind-driven transport across the polar front and subduction at the subtropical front. Our results suggest intimate connections between oceanic carbon uptake and climate variability through the temporal variability of Ekman transport.

Previous modelling studies suggest that more than 40% of the uptake of global oceanic anthropogenic CO_2 (ACO_2) occurs in the Southern Ocean[9–11]. In contrast, observational estimates of the ACO_2 inventory are very low; the area south of 50° S holds only 9% of the global oceanic inventory[12]. This apparent difference between the region's significant carbon uptake and minimal storage implies that a large fraction of ACO_2 absorbed into the Southern Ocean is subsequently exported to the northern basins[13]. It is unclear, however, how this transport is achieved. Ocean eddies and jets are dominant features of the Antarctic Circumpolar Current (ACC). Lateral transport by eddy stirring along constant density (isopycnal) surfaces has been hypothesized to explain the northward transport[14]. Although the distribution of some tracers, such as salinity and nutrients, are primarily oriented along isopycnal surfaces, others, such as potential vorticity, have significant gradients across the ACC[15]. The dynamically relevant potential-vorticity gradients across the ACC may inhibit the eddy stirring of tracers at these latitudes, which challenges the hypothesized role of along-isopycnal eddy stirring. Furthermore, isopycnal eddy stirring leads to tracer homogenization, and a northward ACO_2 flux is clearly up-gradient.

The atmosphere overlying the Southern Ocean is undergoing a significant change characterized by a positive trend in the index of the southern annular mode[16,17] and, thus, increasing westerly winds. This climate trend is likely to continue in the coming decades[18]. The Southern Ocean biogeochemistry and associated carbon fluxes will respond to the intensification of meridional overturning circulation driven by the stronger winds[6,7]. Whether the transport driven by the increasing winds will be compensated by an increase in mesoscale-eddy transport is uncertain[8].

Despite its climatic importance, the mechanisms and pathways of ACO_2 transport are poorly understood. Several complications have slowed progress in studying this topic. Horizontal scales of ocean eddies are in the range of 10–100 km and are not explicitly resolved in typical ocean circulation and carbon cycle models. Physical and chemical properties of the Southern Ocean have been historically undersampled owing to its vast area, remoteness and severe weather conditions. Recently, however, supercomputing resources have become available that allow eddy resolutions to be reached. Furthermore, Southern Ocean observations have been greatly increased owing to augmentation of shipboard measurements by satellite sensors and autonomous floats. A high-resolution ($1/6° \times 1/6°$ longitude–latitude grid) estimate of the Southern Ocean circulation has been developed by using these resources to make an eddy-permitting ocean general circulation model[19,20] consistent with the modern observations (Methods).

In this study, we couple a marine carbon cycle model to the high-resolution circulation estimate, allowing diagnosis of the physical processes controlling the uptake and lateral transport of ACO_2 in the Southern Ocean. Fluxes and distributions of ACO_2 are determined by comparing two simulations in which the model is forced with the observed atmospheric concentration of CO_2 (contemporary run) and with the constant, pre-industrial CO_2 concentration of 278 parts per million by volume (natural run). The two runs are initialized separately using observational estimates[21] of contemporary and natural carbon in 1995. Then the model is integrated for 12 yr, up to the end of 2006. We focus on the intra-annual variability over the 2-yr period from January 2005 to December 2006. Because this is the first time that such a high-resolution ocean state estimate has been used to study the uptake and transport of ACO_2, we perform extensive model validation to evaluate uncertainties in the simulated ACO_2 distribution (Methods and Supplementary Information). Simulated model fields reproduce the observed pattern of water mass distribution, tracer properties and the seasonal variability of areas covered in sea ice. However, our estimate of anthropogenic carbon is not perfect, and the sources of uncertainty include errors in the initial conditions, simplified parameterization of biological processes and the circulation errors. On the basis of direct comparison with *in situ* observations, the largest discrepancy occurs near the base of the mixed layer and at the fronts, and the standard error in local ACO_2 concentration is less than 16%.

Simulated uptake and column inventory of ACO_2 reveals significant spatial variability and distinct patterns associated with the

[1]Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, Colorado 80523-1371, USA. [2]Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0230, USA.

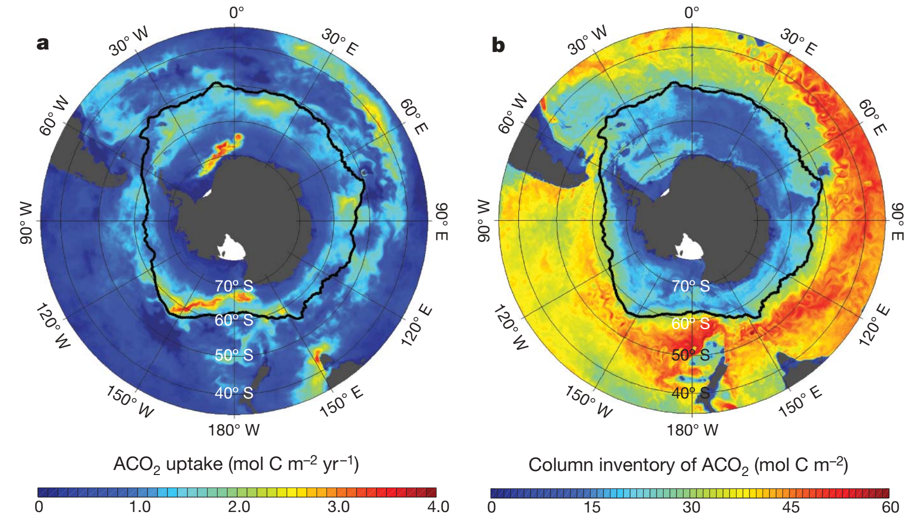


Figure 1 | Uptake and inventory of anthropogenic carbon. a, The 2-yr mean of ACO$_2$ uptake rates evaluated between January 2005 and December 2006; **b,** the column inventory of ACO$_2$ (storage) determined from a 5-d mean in December 2006. The black solid line represents the position of the APF calculated from satellite observations of sea surface temperature[29].

frontal structures of the ACC. Figure 1a highlights the enhanced ACO$_2$ uptake in the three major regions: poleward of the Antarctic polar front (APF), which encircles the globe at ~55° S (marked by the black solid line); the western boundary regions of the subtropical gyres (at the subtropical front at ~40° S); and the high latitudes near the Weddell and Ross seas. These regions are characterized by enhanced oceanic vertical exchange due to convective mixing, wind-driven upwelling and (often topographically induced) mesoscale-eddy variability[22] (Supplementary Information). Spatial correlation (correlation coefficient, $r = 0.51$; significant at the 95% confidence interval) is found between the 2-yr-averaged air–sea ACO$_2$ flux and the logarithm of the standard deviation of vertical velocity at a depth of 225 m. The relatively slow air–sea CO$_2$ exchange (on the order of 1 yr) and the strong lateral advection are probably responsible for the moderate spatial correlation. Other processes (for example air–sea gas exchange rates that depend on the surface wind speed and sea-ice cover) are found to have only minor influences on the uptake. The important role of vertical exchange is consistent with previous studies[5,13] that suggest entrainment of unperturbed deep waters to be the rate-controlling process for the ACO$_2$ uptake.

Figure 1b shows the simulated column inventory of ACO$_2$ in December 2006. The low ACO$_2$ inventory poleward of the APF is in contrast to the enhanced ACO$_2$ uptake there. The location of the APF separates the polar waters, which are relatively depleted of ACO$_2$, from the relatively enriched sub-Antarctic waters; this is consistent with observational studies indicating low ACO$_2$ storage in the polar regions[12]. The mismatch between the regions of uptake and storage indicates significant up-gradient transport of ACO$_2$ (ref. 13). Mesoscale eddies and meandering jets dominate the local northward transport of ACO$_2$, as is demonstrated by the small scale (10–100-km) alternating transport directions in Fig. 2a. The net northward transport is determined by calculating the zonally averaged ACO$_2$ transports. The northward ACO$_2$ transport is decomposed into the zonal mean and the eddy components, where an eddy is defined as a deviation from a zonal mean. Eddy transport can be further decomposed into transient and stationary components. The zonal mean can be defined in two different ways: along constant-latitude circles and along the mean streamline of the ACC. Although the first definition is simple, using the second removes the effects of the meandering of the mean pathway of the ACC, and the eddy fluxes primarily reflect the mesoscale variability.

Compensations between the mean flow and eddies are crucial in quantifying the lateral carbon transport. Figure 2b shows the decomposition using the zonal mean at constant latitude, where the

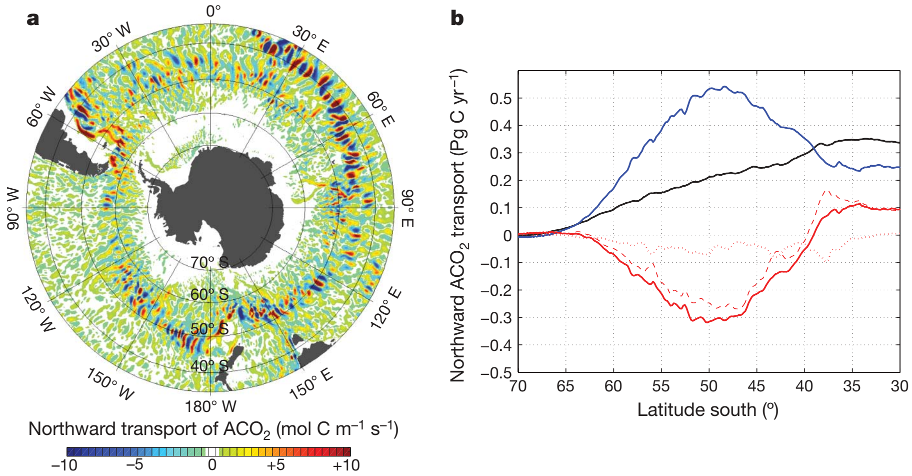


Figure 2 | Northward transport of anthropogenic carbon. a, Vertically integrated northward ACO$_2$ transport determined from a 5-d mean in December 2006; **b,** the zonally and temporally integrated magnitudes of northward ACO$_2$ flux: zonal-mean (blue), eddy (red) and net residual (black) transport. Eddy transport is further decomposed into transient (dotted) and stationary (dashed) components. The zonal and temporal mean was calculated along constant-latitude circles for the 2-yr simulation period.

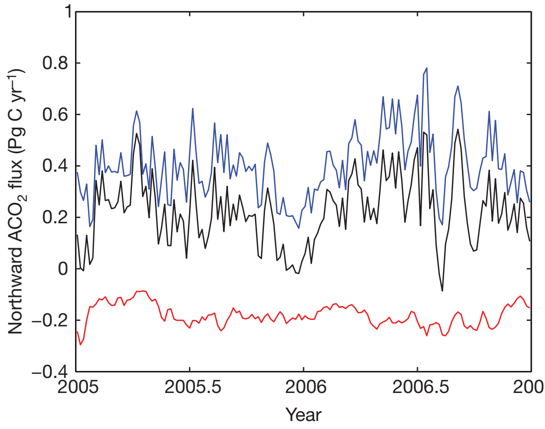


Figure 3 | Variability of anthropogenic carbon transport. Time series of northward cross-frontal ACO_2 flux: zonal-mean (blue), eddy (red) and net (black) transport. The zonal and temporal mean was calculated along the position of the APF for the 2-yr simulation period.

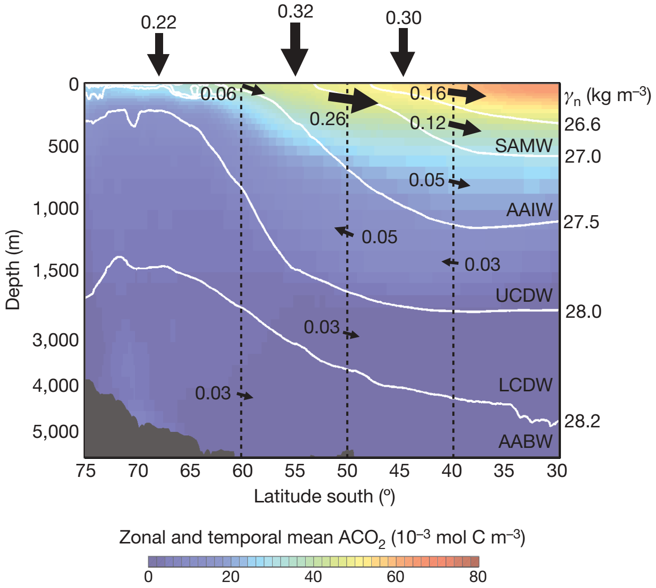


Figure 4 | Vertical structure of anthropogenic carbon transport. Major density layers (differentiated by white lines) are defined using neutral density (γ_n): sub-Antarctic mode water (SAMW), Antarctic intermediate water (AAIW), upper circumpolar deep water (UCDW), lower circumpolar deep water (LCDW) and Antarctic bottom water (AABW). The size of each black arrow is proportional to the magnitude of integrated ACO_2 flux (in units of petagrams of carbon per year); the magnitude is displayed only for fluxes greater than $0.02\,Pg\,C\,yr^{-1}$. Background colour shading indicates zonally and temporally averaged ACO_2. The temporal mean was calculated over the 2-yr simulation period.

net northward ACO_2 transport is achieved by a relatively small residual between a northward mean transport and a southward eddy flux. Eddy transport is primarily dominated by the stationary component. Poleward of 45° S, the mean component results primarily from the strong wind-driven Ekman transport; more than 90% of the mean northward flux occurs in the top 40 m of the ocean at 50° S. Though small-scale eddy fluxes dominate locally, as shown in Fig. 2a, zonally averaged ACO_2 transport is accomplished by means of the zonal-mean circulation.

The main axis of the ACC, as measured by the vertically integrated (barotropic) stream function, is closely aligned with the position of the APF, which is controlled by the topography of the ocean floor. The net cross-frontal ACO_2 transport can then be determined by measuring the zonal mean along the position of the APF, which is essentially identical to the along-streamline mean of the ACO_2 transport. The net northward ACO_2 transport averaged along the APF for the 2-yr simulation period is 0.22 petagrams of carbon (Pg C) per year. Again, this net transport is the residual of a northward zonal-mean transport of $0.41\,Pg\,C\,yr^{-1}$ and a southward eddy transport of $0.19\,Pg\,C\,yr^{-1}$. Similar to the zonal mean along constant-latitude circles, the net northward ACO_2 transport involves significant compensation between a northward zonal-mean transport and a southward eddy flux.

The meridional ACO_2 transport shows significant temporal variability (Fig. 3). The standard deviation in the cross-frontal transport is $0.13\,Pg\,C\,yr^{-1}$, and this variability is dominated by the mean component, which explains 90% of the variance of the net transport. The temporal variability of the mean component is almost entirely controlled by the cross-frontal mass flux by Ekman transport (with $r = 0.98$, significant at the 99% confidence interval). Thus, variability in the net cross-frontal ACO_2 transport is essentially driven by the atmospheric wind variability with a decorrelation timescale of about 10 d.

Considering the overall Southern Ocean budget, the spatially integrated oceanic ACO_2 uptake south of 40° S is $0.83\,Pg\,C\,yr^{-1}$ over the 2-yr period, which is roughly 40% of the global oceanic ACO_2 uptake[13,23]. Across the latitude circle at 40° S, $0.30\,Pg\,C\,yr^{-1}$ (36% of the uptake south of 40° S) is exported to the northern basins. Analysing the vertical structure of the northward ACO_2 transport further illuminates the transport mechanisms (Fig. 4). Although the ACO_2 transport is dominated by the Ekman flow at the latitudes of the APF, the subduction and circulation of thermocline waters become increasingly important northward of the ACC. At 40° S, approximately one-half of the northward ACO_2 transport occurs in the surface layer, probably driven by Ekman dynamics. The other half of the transport occurs below the mixed layer in the density classes of sub-Antarctic mode water and Antarctic intermediate water, indicating the importance of the dynamics associated with the formation and subduction of these water masses. At 35° S, the northward ACO_2 transport is fairly evenly distributed between the surface and a depth of 600 m, confirming the important role of these water masses in the subtropics.

These results offer new perspectives on our understanding and the quantification of the anthropogenic carbon uptake in the Southern Ocean. The zonal-mean transport, which dominates the net northward transport of ACO_2 across the ACC, is synchronous to the Southern Ocean's zonal wind-stress pattern, indicating a link to atmospheric climate variability. Stronger westerly winds over the Southern Ocean may enhance regional uptake of ACO_2 and increase its transport to the northern basins. It is not yet clear how eddy fluxes may respond to the changes in the westerly wind on interannual and longer timescales. Our analysis demonstrates that there is no compensation on short, intra-annual timescales, over which the variability of northward ACO_2 transport is primarily driven by the variability of wind-driven Ekman transport and is not significantly correlated with that of eddy transport ($r = -0.03$). This raises the question of on what timescales the mean flow–eddy compensation can regulate the cross-frontal fluxes, which is beyond the scope of this paper and is left for future study. Climate variability also affects the natural carbon cycle. It has been suggested that the intensification of surface winds over the Southern Ocean may increase the upwelling of carbon-rich deep waters and the outgassing of natural CO_2 (refs 6, 7). Because the large-scale gradients of anthropogenic and natural CO_2 have opposite signs in the Southern Ocean, opposite sensitivities of anthropogenic and natural carbon fluxes are anticipated. The anthropogenic fraction of atmospheric CO_2 concentration is predicted to increase in the future, and the role of ACO_2 fluxes may become even more important in controlling the total CO_2 uptake. Understanding the mechanisms underlying the uptake and transport of both anthropogenic and natural CO_2, and their interplay, can help in predicting the future oceanic sink of CO_2.

METHODS SUMMARY

We developed a high-resolution ocean circulation and carbon cycle model using an eddy-permitting ocean circulation model with a lateral resolution of 1/6°. The circulation of the model was optimized to physical observations in a weighted least-squares sense[24,25]. We used a cost function to compare the model to *in situ* observations, altimetric observations and other data sets. Reduction of the model–observation discrepancy was achieved by systematically adjusting the control variables (prescribed atmospheric state and initial conditions) using the adjoint method[26]. We simulated the ocean carbon cycle and air–sea fluxes of CO_2 in offline mode using 5-d-averaged state-estimate fields and a simple biogeochemistry model based on the Ocean Carbon-Cycle Model Intercomparison Project scheme[27], in which biological carbon uptake is parameterized using the linear relaxation of surface phosphate to the monthly mean climatology of the World Ocean Atlas 2005[28]. Anthropogenic components of carbon concentrations and fluxes were determined by subtracting the natural run from the contemporary run. The two runs were initialized in 1995 using two separate initial conditions based on the Global Ocean Data Analysis Project data set[21], which consists of the contemporary and natural components of dissolved inorganic carbon.

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

Received 6 December 2008; accepted 9 November 2009.

1. Toggweiler, J. R. & Samuels, B. Effect of Drake Passage on the global thermohaline circulation. *Deep-Sea Res. I* **42,** 477–500 (1995).
2. Sarmiento, J. L. *et al.* Response of ocean ecosystems to climate warming. *Glob. Biogeochem. Cycles* **18,** GB3003 (2004).
3. Russell, J. L., Dixon, K. W., Gnanadesikan, A., Stouffer, R. J. & Toggweiler, J. R. The Southern Hemisphere westerlies in a warming world: propping open the door to the deep ocean. *J. Clim.* **19,** 6382–6390 (2006).
4. Marinov, I., Gnanadesikan, A., Toggweiler, R. & Sarmiento, J. L. The Southern Ocean biogeochemical divide. *Nature* **441,** 964–967 (2006).
5. Ito, T., Marshall, J. & Follows, M. What controls the uptake of transient tracers in the Southern Ocean? *Glob. Biogeochem. Cycles* **18,** GB2021 (2004).
6. Le Quéré, C. *et al.* Saturation of the Southern Ocean CO_2 sink due to recent climate change. *Science* **316,** 1735–1738 (2007).
7. Lovenduski, N. S., Gruber, N. & Doney, S. C. Toward a mechanistic understanding of the decadal trends in the Southern Ocean carbon sink. *Glob. Biogeochem. Cycles* **22,** GB3016 (2008).
8. Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. R. & Schwarzkopf, F. U. The response of the Antarctic Circumpolar Current to recent climate change. *Nature Geosci.* **1,** 864–869 (2008).
9. Watson, A. J. & Orr, J. C. in *Carbon Dioxide Fluxes in the Global Ocean* (eds Field, J., Fasham, M., Platt, T. & Zeitzschel, B.) 123–141 (Springer, 2003).
10. Orr, J. C. *et al.* Estimates of anthropogenic carbon uptake from four three-dimensional global ocean models. *Glob. Biogeochem. Cycles* **15,** 43–60 (2001).
11. Mikaloff-Fletcher, S. E. *et al.* Inverse estimates of anthropogenic CO_2 uptake, transport, and storage by the ocean. *Glob. Biogeochem. Cycles* **20,** GB2002 (2006).
12. Sabine, C. L. *et al.* The oceanic sink for anthropogenic CO_2. *Science* **305,** 367–371 (2004).
13. Sarmiento, J. L., Orr, J. C. & Siegenthaler, U. A perturbation simulation of CO_2 uptake in an ocean general circulation model. *J. Geophys. Res.* **97,** 3621–3645 (1992).
14. Caldeira, K. & Duffy, P. B. The role of the Southern Ocean in uptake and storage of anthropogenic carbon dioxide. *Science* **287,** 620–622 (2000).
15. Marshall, J., Shuckburgh, E., Jones, H. & Hill, C. Estimates and implications of surface eddy diffusivity in the southern ocean derived from tracer transport. *J. Phys. Oceanogr.* **36,** 1806–1821 (2006).
16. Thompson, D. W. J. & Solomon, S. Interpretation of recent Southern Hemisphere climate change. *Science* **296,** 895–899 (2002).
17. Marshall, G. J. Trends in the southern annular mode from observations and reanalyses. *J. Clim.* **16,** 4134–4143 (2003).
18. Miller, R. L., Schmidt, G. A. & Shindell, D. T. Forced annular variations in the 20th century Intergovernmental Panel on Climate Change Fourth Assessment Report model. *J. Geophys. Res.* **111,** D18101 (2006).
19. Marshall, J., Hill, C., Perelman, L. & Adcroft, A. Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling. *J. Geophys. Res.* **102,** 5733–5752 (1997).
20. Marshall, J., Adcroft, A., Hill, C., Perelman, L. & Heisey, C. A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.* **102,** 5753–5766 (1997).
21. Key, R. M. *et al.* A global ocean carbon climatology: results from Global Data Analysis Project (GLODAP). *Glob. Biogeochem. Cycles* **18,** GB4031 (2004).
22. Sokolov, S. & Rintoul, S. R. On the relationship between fronts of the Antarctic Circumpolar Current and surface chlorophyll concentrations in the Southern Ocean. *J. Geophys. Res.* **112,** C07030 (2007).
23. Gruber, N. *et al.* Oceanic sources, sinks, and transport of atmospheric CO_2. *Glob. Biogeochem. Cycles* **23,** GB1005 (2009).
24. Mazloff, M. *The Southern Ocean Meridional Overturning Circulation as Diagnosed from an Eddy Permitting State Estimate.* PhD thesis, Massachusetts Inst. Technol. (2008).
25. Mazloff, M. R., Heimbach, P. & Wunsch, C. An eddy permitting Southern Ocean state estimate. *J. Phys. Oceanogr.* (submitted).
26. Wunsch, C. & Heimbach, P. Practical global ocean state estimation. *Physica D* **230,** 197–208 (2007).
27. Najjar, R., Sarmiento, J. & Toggweiler, J. R. Downward transport and fate of organic matter in the ocean: simulations with a general circulation model. *Glob. Biogeochem. Cycles* **6,** 45–76 (1992).
28. Garcia, H. E., Locarnini, R. A., Boyer, T. P. & Antonov, J. I. in *NOAA Atlas NESDIS 64* (ed. Levitus, S.) 396 (US Government Printing Office, 2006).
29. Moore, J. K., Abbott, M. R. & Richman, J. G. Location and dynamics of the Antarctic Polar Front from satellite sea surface temperature data. *J. Geophys. Res.* **104,** 3059–3073 (1999).

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

Acknowledgements T.I. and M.W. are supported by the US National Aeronautics and Space Administration (NASA) grant NNX08AL72G and the US National Oceanic and Atmospheric Administration grant NA08OAR4320893. The Southern Ocean State Estimate is supported by NASA and US National Oceanographic Partnership Program contracts to the Massachusetts Institute of Technology. The NASA Advanced Supercomputing division and the San Diego Supercomputer Center provided computing and data storage resources. J. L. Sarmiento and N. Lovenduski provided comments that improved the manuscript.

Author Contributions T.I. designed and performed numerical simulations of the Southern Ocean carbon cycle and analysed the model output; M.W. performed model–data comparison and calculations of carbon transport; M.M. developed the Southern Ocean State Estimate; all authors contributed to the interpretation of the results and writing of the manuscript.

Author Information 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 T.I. (ito@atmos.colostate.edu).

METHODS

Model development. We optimized the Massachusetts Institute of Technology ocean general circulation model[19,20] with a lateral resolution of 1/6°, which obeys the Navier–Stokes equations and conserves mass, heat and salt, to physical observations in a weighted least-squares sense[24,25]. We used a cost function to compare the model output with *in situ* observations (ARGO, CTD, SEaOS, and XBTs), altimetric observations (ENVISAT, GEOSAT, Jason), and other data sets (for example sea surface temperature). Reduction of the model–observation discrepancy was achieved by systematically adjusting the control variables (prescribed atmospheric state and initial conditions) using the adjoint method[26]. Costs associated with control-variable perturbations ensured a physically realistic solution. The Southern Ocean has been very thoroughly observed in the past several years, and the state estimate is now largely consistent with this collection of data[24]. We simulated the ocean carbon cycle and air–sea fluxes of CO_2 in the offline mode using 5-d-averaged state-estimate fields and a simple biogeochemistry scheme[27]. The biogeochemistry model was based on the Ocean Carbon-Cycle Model Intercomparison Project scheme[27], in which biological carbon uptake is parameterized using the linear relaxation of surface phosphate to the monthly mean climatology of the World Ocean Atlas 2005[28]. The model transports five tracers, including dissolved inorganic carbon (DIC), alkalinity, phosphate, dissolved organic phosphate and oxygen. The top 75 m of the modelled phosphate was conditionally restored towards monthly climatology values if the model concentration was greater than that of climatology. Constant stoichiometric ratios were used to calculate the elemental ratio of organic material between phosphate, carbon and oxygen, 1:117:−170, and a constant rain ratio of 0.07 was used to calculate the calcium carbonate formation. One-third of carbon uptake was allocated to the sinking particulate pool, whose vertical dissolution profile is given using a simple power-law function.

We determined anthropogenic components of carbon concentrations and fluxes by subtracting the natural run from the contemporary run. The two runs were initialized in 1995 using two separate initial conditions based on the Global Ocean Data Analysis Project data set[21], which consists of the natural and contemporary components of DIC. Then the model was 'spun up' for 10 yr, to the end of 2004; initial model drifts decreased after a few years of the spin-up integration. We focused on the intra-annual variability over the 2-yr period from January 2005 to December 2006. These calculations required massive computational resources and were made possible by the San Diego Supercomputer Center's DataStar supercomputer and NASA's Columbia and Pleiades supercomputers.

Model validation. We performed extensive model validation to evaluate uncertainties in the simulated ACO_2 distribution. Direct comparison of simulated tracer fields with *in situ* measurements from CLIVAR repeat hydrography lines A16S and P16S demonstrated significant skill in reproducing the observed pattern of water mass distribution and tracer properties. The representations of hydrographic temperature–salinity (*T*–*S*) structure and water mass distribution were well reproduced by the model. It is crucial to represent upper-ocean water masses such as SAMW (potential-density range, $26.5\,\mathrm{kg\,m^{-3}} < \sigma_0 < 27.1\,\mathrm{kg\,m^{-3}}$) to reproduce transient tracers such as anthropogenic CO_2 and chlorofluorocarbons. Close examination of the observed *T*–*S* profiles showed somewhat tighter gradients in *T*–*S* space, suggesting that the model may be overemphasizing the degree of diffusion and mixing. The model–data discrepancy in the *in situ* ocean temperature and salinity was quite small over most of the Southern Ocean. Model–data comparison of biogeochemical tracers such as DIC and alkalinity ensured that the model captured the magnitude and gradients of observed *in situ* distributions; however, these tracers showed a model–data discrepancy greater than that of temperature and salinity. The sources of uncertainty include errors in the initial conditions, simplified parameterization of biological processes, and circulation errors including front position and turbulent tracer mixing.

Errors in DIC and alkalinity can produce errors in the buffer factor and the estimates of anthropogenic carbon. The model was initialized using the Global Ocean Data Analysis Project data set[21] representing the observationally determined three-dimensional distribution of anthropogenic CO_2, and its inventory. The error associated with this initial condition was relatively small; the inventory error was on the order of 7% and the local concentration error was on the order of 20% (ref. 30). Simplified parameterization of biological carbon uptake forces the surface nutrients to remain close to the climatological levels, which may underestimate the variability of biological carbon sources and sinks. Comparison with *in situ* measurements from CLIVAR lines A16S and P16S indicate that the root-mean-squared model–data discrepancies for DIC and alkalinity are respectively 17.8 μM and 17.4 μequiv. $\mathrm{kg^{-1}}$ at the surface layer, and that the discrepancies decrease in the interior ocean. These tracers control the buffer (Revelle) factor, regulating the ability of the sea water to absorb anthropogenic carbon from the atmosphere. Incomplete representation of DIC and alkalinity affects the simulated anthropogenic carbon fluxes through errors in the buffer factor. We estimated the magnitude of this uncertainty to be less than 16%, on the basis of carbonate chemistry calculations.

30. Matsumoto, K. & Gruber, N. How accurate is the estimation of anthropogenic carbon in the ocean? An evaluation of the Delta C* method. *Glob. Biogeochem. Cycles* 19, GB3014 (2005).

LETTERS

Endogenous non-retroviral RNA virus elements in mammalian genomes

Masayuki Horie[1]*, Tomoyuki Honda[1,2]*, Yoshiyuki Suzuki[3], Yuki Kobayashi[3], Takuji Daito[1], Tatsuo Oshida[4], Kazuyoshi Ikuta[1], Patric Jern[5], Takashi Gojobori[3], John M. Coffin[5] & Keizo Tomonaga[1,6]

Retroviruses are the only group of viruses known to have left a fossil record, in the form of endogenous proviruses, and approximately 8% of the human genome is made up of these elements[1,2]. Although many other viruses, including non-retroviral RNA viruses, are known to generate DNA forms of their own genomes during replication[3–5], none has been found as DNA in the germline of animals. Bornaviruses, a genus of non-segmented, negative-sense RNA virus, are unique among RNA viruses in that they establish persistent infection in the cell nucleus[6–8]. Here we show that elements homologous to the nucleoprotein (N) gene of bornavirus exist in the genomes of several mammalian species, including humans, non-human primates, rodents and elephants. These sequences have been designated endogenous Borna-like N (EBLN) elements. Some of the primate EBLNs contain an intact open reading frame (ORF) and are expressed as mRNA. Phylogenetic analyses showed that EBLNs seem to have been generated by different insertional events in each specific animal family. Furthermore, the EBLN of a ground squirrel was formed by a recent integration event, whereas those in primates must have been formed more than 40 million years ago. We also show that the *N* mRNA of a current mammalian bornavirus, Borna disease virus (BDV), can form EBLN-like elements in the genomes of persistently infected cultured cells. Our results provide the first evidence for endogenization of non-retroviral virus-derived elements in mammalian genomes and give novel insights not only into generation of endogenous elements, but also into a role of bornavirus as a source of genetic novelty in its host.

Bornaviruses are the only animal RNA viruses that achieve a highly cell-associated life cycle within the nuclear envelope[6–9], and can therefore provide not only new models of RNA virus replication, but also insight into dynamics of RNA molecules in eukaryote cells. In an effort to understand whether bornaviruses mimic host factors to maintain persistent infection in the nucleus, we searched human protein databases for sequences with similarity to BDV proteins. This search identified two hypothetical human proteins (GeneID LOC340900 and LOC55096), each of which has significant sequence similarity to BDV N (Fig. 1a and Supplementary Table 1). BDV N is a major structural protein, which tightly encapsidates the viral RNA to form the nucleocapsid. The LOC340900 sequence encodes a protein of comparable length (366 residues) to BDV N (370 residues), whereas LOC55096 seems to contain several frameshift mutations relative to BDV N, resulting in a shorter ORF length (Fig. 1a). Both LOC340900 and LOC55096 showed an overall 41% sequence identity and 58% similarity to BDV N and 72% identity to each other. The close relationship between BDV N and the homologous genes was further demonstrated by the alignment of transcription regulatory sequences on either side of BDV N (Fig. 1b). The S and T motifs in flanking sequences of both putative human proteins were well conserved with those of BDV (Fig. 1b). In addition, a poly-A sequence appears after the T1-like motif in the 3′ flanking region of LOC55096 (Fig. 1b). The homology of the human genes to BDV N was also confirmed by a permutation test (Supplementary Fig. 1). These findings indicated that both human genes may be endogenous elements related to BDV N gene, and therefore we designated them EBLNs (LOC340900, EBLN-1 and LOC55096, EBLN-2).

To investigate the presence of EBLN sequences in other animal species, we conducted tblastn searches using BDV N as a query in eukaryote and whole-genome shotgun databases at NCBI. Sequences with blast *E*-values of 10^{-10} or lower were identified as EBLNs. We found two additional human elements (EBLN-3 and -4) as well as a number of related sequences in various mammalian species, including marsupials (Supplementary Table 2). Orthologous genes to human EBLNs were identified in the genomes of non-human anthropoid primates, including chimpanzee, gorilla, orang-utan, and macaque (Supplementary Table 2). We also detected primate EBLNs in the genomes of the suborder Strepsirrhini, including the mouse lemur and Garnett's galago. Furthermore, two species of the Afrotheria, African elephant and cape hyrax, and four rodents were found to have EBLNs with *E*-values of less than 10^{-20} (Supplementary Table 2). An EBLN locus with a high level of similarity to BDV N was also identified in the thirteen-lined ground squirrel (TLS) genome (Supplementary Fig. 2a). Like the human EBLNs, the TLS EBLN contained a 3′ poly-A sequence, as well as S and T signal motifs, in its 3′ flanking region (Supplementary Fig. 2b). Almost all EBLN fragments, except for EBLN-1 and the TLS gene, contained several stop codons in the predicted coding sequences, or lacked the identifiable flanking sequences. In addition, we found that all anthropoid EBLNs, except for EBLN-4, are expressed as mRNAs in some human and monkey-derived cell lines (Supplementary Fig. 3). A previous study reported the interaction of human EBLN-2 with other cellular proteins, such as AP1S1, TUSC2/FUS1 and FANCC (ref. 10) (Supplementary Table 1), indicating that anthropoid EBLNs may encode functional proteins.

To investigate whether other mammalian species contain EBLN-related sequences in their genomes further, we conducted Southern blot hybridization under low-stringency conditions using human, murine and TLS EBLN as probes (Fig. 1c and d). Along with the clear signals in primate genomes, we detected reproducible faint positive bands in murine and shrew genomes when using a human EBLN probe (Fig. 1c, dots). The signals were also observed using a mouse

[1]Department of Virology, Research Institute for Microbial Diseases (BIKEN), Osaka University, Osaka 565-0871, Japan. [2]Japan Society for the Promotion of Science (JSPS), Chiyoda-ku, Tokyo 102-8472, Japan. [3]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan. [4]Department of Life Science and Agriculture, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido 080-8555, Japan. [5]Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, Massachusetts 02111, USA. [6]PRESTO, Japan Science and Technology Agency (JST), Chiyoda-ku, Tokyo 102-0075, Japan.
*These authors contributed equally to this work.

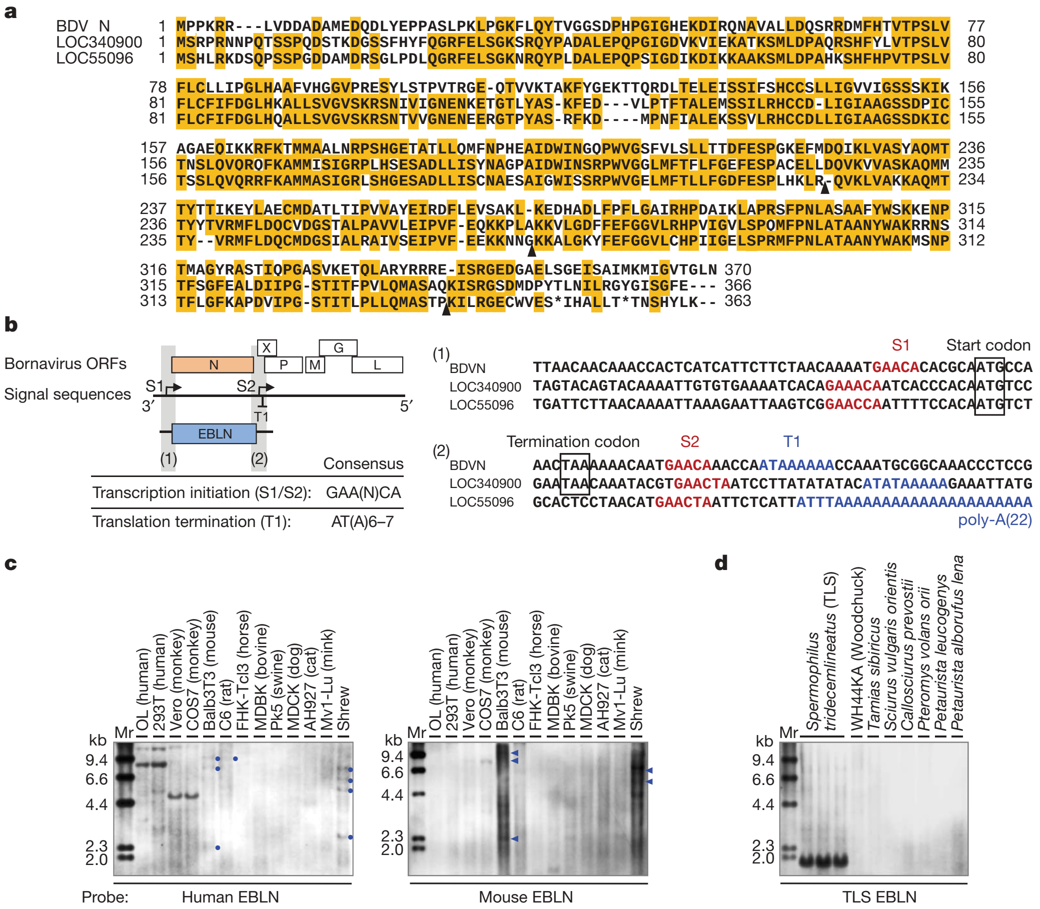

a
BDV N 1 MPPKRR---LVDDADAMEDQDLYEPPASLPKLPGKFLQYTVGGSDPHPGIGHEKDIRQNAVALLDQSRRDMFHTVTPSLV 77
LOC340900 1 MSRPRNNPQTSSPQDSTKDGSSFHYFQGRFELSGKSRQYPADALEPQPGIGDVKVIEKATKSMLDPAQRSHFYLVTPSLV 80
LOC55096 1 MSHLRKDSQPSSPGDDAMDRSGLPDLQGRFELSGKNRQYPLDALEPQPSIGDIKDIKKAAKSMLDPAHKSHFHPVTPSLV 80
78 FLCLLIPGLHAAFVHGGVPRESYLSTPVTRGE-QTVVKTAKFYGEKTTQRDLTELEISSIFSHCCSLLIGVVIGSSSKIK 156
81 FLCFIFDGLHKALLSVGVSKRSNIVIGNENKETGTLYAS-KFED---VLPTFTALEMSSILRHCCD-LIGIAAGSSDPIC 155
81 FLCFIFDGLHQALLSVGVSKRSNTVVGNENEERGTPYAS-RFKD----MPNFIALEKSSVLRHCCDLLIGIAAGSSDKIC 155
157 AGAEQIKKRFKTMMAALNRPSHGETATLLQMFNPHEAIDWINGQPWVGSFVLSLLTTDFESPGKEFMDQIKLVASYAQMT 236
156 TNSLQVQRQFKAMMISIGRPLHSESADLLISYNAGPAIDWINSRPWVGGLMFTFLFGEFESPACELLDQVKVVASKAQMM 235
156 TSSLQVQRRFKAMMASIGRLSHGESADLLISCNAESAIGWISSRPWVGELMFTLLFGDFESPLHKLR-QVKLVAKKAQMT 234
237 TYTTIKEYLAECMDATLTIPVVAYEIRDFLEVSAKL-KEDHADLFPFLGAIRHPDAIKLAPRSFPNLASAAFYWSKKENP 315
236 TYYTVRMFLDQCVDGSTALPAVVLEIPVF-EQKKPLAKKVLGDFFEFGGVLRHPVIGVLSPQMFPNLATAANYWAKRRNS 314
235 TY--VRMFLDQCMDGSIALRAIVSEIPVF-EEKKNNGKKALGKYFEFGGVLCHPIIGELSPRMFPNLATAANYWAKMSNP 312
316 TMAGYRASTIQPGASVKETQLARYRRRE-ISRGEDGAELSGEISAIMKMIGVTGLN 370
315 TFSGFEALDIIPG-STITFPVLQMASAQKISRGSDMDPYTLNILRGYGISGFE--- 366
313 TFLGFKAPDVIPG-STITLPLLQMASTPKILRGECWVES*IHALLT*TNSHYLK-- 363
b
Bornavirus ORFs
N
X
P
M
G
L
Signal sequences
S1
S2
3′
5′
T1
EBLN
(1)
(2)
Consensus
Transcription initiation (S1/S2): GAA(N)CA
Translation termination (T1): AT(A)6–7
(1)
S1
Start codon
BDVN TTAACAACAAACCACTCATCATTCTTCTAACAAAATGAACACACGCAATGCCA
LOC340900 TAGTACAGTACAAAATTGTGTGAAAATCACAGAAACAATCACCCACAATGTCC
LOC55096 TGATTCTTAACAAAATTAAAGAATTAAGTCGGAACCAATTTTCCACAATGTCT
(2)
Termination codon
S2
T1
BDVN AACTAAAAAACAATGAACAAACCAATAAAAAACCAAATGCGGCAAACCCTCCG
LOC340900 GAATAACAAATACGTGAACTAATCCTTATATATACATATAAAAAGAAATTATG
LOC55096 GCACTCCTAACATGAACTAATTCTCATTATTTAAAAAAAAAAAAAAAAAAAAAA
poly-A(22)
c
kb
9.4
6.6
4.4
2.3
2.0
Mr
OL (human)
293T (human)
Vero (monkey)
COS7 (monkey)
Balb3T3 (mouse)
C6 (rat)
FHK-Tcl3 (horse)
MDBK (bovine)
Pk5 (swine)
MDCK (dog)
AH927 (cat)
Mv1-Lu (mink)
Shrew
Probe:
Human EBLN
Mouse EBLN
d
Spermophilus tridecemlineatus (TLS)
WH44KA (Woodchuck)
Tamias sibiricus
Sciurus vulgaris orientis
Callosciurus prevostii
Pteromys volans orii
Petaurista leucogenys
Petaurista alborufus lena
TLS EBLN

Figure 1 | Bornavirus N-like elements in mammalian genomes. **a**, Alignment between predicted amino acid sequences of BDV N and two human bornavirus N-like elements. Black arrowheads indicate predicted frameshift sites in LOC55096. **b**, Sequence alignments of transcription signal sites (S1/S2 and T1) at both the 5′ and 3′ ends of the bornavirus N ORF. A schematic representation of bornavirus genome structure is shown. **c**, **d**, Low-stringency Southern blot hybridizations of DNA from various mammalian species using human EBLN-1 and mouse EBLN chr.11 (**c**) and TLS EBLN (**d**) as probes. Dots and arrowheads on the right side of the murine and shrew lanes in panel **c** indicate the positions of reproducible positive signals. Mr, Molecular marker.

EBLN probe (Fig. 1c, arrowheads), indicating that the faint bands are most likely to be EBLN-related sequences. In fact, EBLN-like sequences, albeit with *E*-values greater than 10^{-10}, were found in the Eurasian shrew genome in our tblastn searches. On the other hand, except for TLS, no positive band was detected by the TLS probe in the genomes of several different squirrel species, such as woodchuck (*Marmota* spp.), the closest species to the TLS (*Spermophilus* spp.) (Fig. 1d)[11], indicating that the ground squirrels are likely to be the only host species of EBLN within the squirrel family. The BDV N probe detected many faint and smear bands that include the signals detected by EBLN-specific probes in both selected mammalian species and the squirrel families (Supplementary Fig. 4), indicating that EBLN-related fragments are more widely distributed in the mammalian genome.

We next performed a comprehensive phylogenetic analysis using nucleotide sequences of all EBLNs with *E*-values less than 10^{-20} (Fig. 2 and Supplementary Fig. 5). In addition to EBLNs, we included avian bornaviruses (ABVs)[12] and an exogenous reptile bornavirus (RBV) sequence, which was detected in a cDNA library from a *Bitis gabonica* (Gaboon viper) venom gland[13] (Supplementary Fig. 6). As shown in Fig. 2, the anthropoid and murine EBLNs are clustered phylogenetically within each host order. By contrast, EBLNs from other species, including African elephant, cape hyrax and guinea pig, form branches independent from the evolutionary lineage of their hosts, indicating that these EBLNs had most likely invaded each species via independent integration events. Interestingly, the TLS EBLNs form a tight cluster more closely related to modern exogenous bornaviruses than to those of other animals. Considering that a closely related species does not contain EBLNs, the integration of squirrel EBLN could have been a very recent event. A phylogenetic analysis using all primate EBLNs, including marmoset (Supplementary Fig. 7), showed that the integration events leading to the primate EBLNs occurred in the Haplorrhini at least before the split between rhesus macaque and marmoset.

To investigate whether current bornaviruses are able to be copied into DNA to produce EBLN-like elements, we first performed PCR analyses using DNA of persistently BDV-infected cells. As shown in Fig. 3a and Supplementary Table 3, BDV DNA was clearly detected in some cell lines by a primer set targeted to the BDV N region. To understand which viral RNA species serve as template for the DNA form of BDV, we used several primers within the BDV genome for amplification. The results showed that primer sets straddling the boundaries of BDV transcription units could not amplify BDV-specific DNA (Fig. 3b and c), indicating that the DNA is transcribed from mRNAs of BDV. We detected BDV-specific DNA in the brains of persistently BDV-infected mice (Supplementary Fig. 8), indicating that BDV can produce DNA forms *in vitro* and *in vivo*. We next performed Alu-PCR to investigate whether BDV DNA detected in the infected cells exists as integrated or extrachromosomal DNA. As shown in Fig. 3d and Supplementary Fig. 9, an Alu-specific PCR product was detected in BDV-infected cells only when using an N-specific forward primer about 30 days post-infection. This observation indicated that although BDV DNA in infected cells may be mainly extrachromosomal, the N gene is integrated into the host genome during persistent infection.

We further characterized the BDV DNA insertions and flanking cellular sequences by using Alu-PCR and inverse PCR (Supplementary Fig. 10)[14]. Integration sites were present on various chromosomes

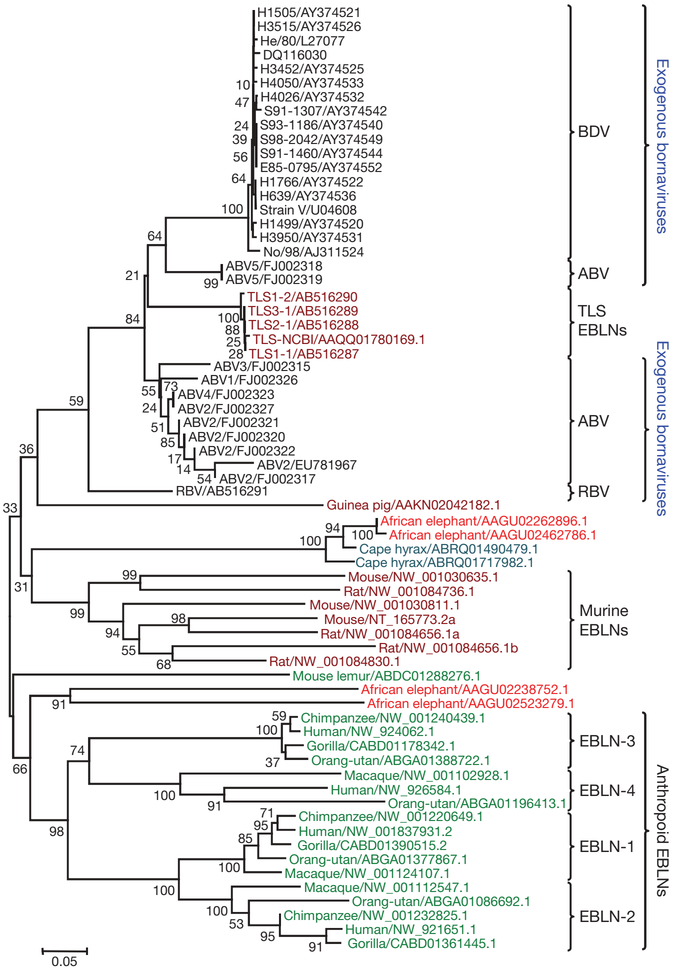


Figure 2 | Phylogenetic tree of exogenous bornaviruses and mammalian EBLNs. The bootstrap probability is indicated for each interior branch. The scale bar indicates the number of amino acid substitutions per site. Animals belonging to the same order are indicated by the same colour. Strain and sequence accession numbers are given for each sequence.

(Fig. 4). Similar to some mammalian EBLNs, many BDV DNA insertions contained a 3′ poly-A sequence (Fig. 4b and c). In addition, integrations of truncated BDV N DNA were also found in some clones. No apparent consensus sequences were found at the sites, although target site duplications (TSDs) were detected in some clones from the inverse PCR (Fig. 4c). We also found deletions, as well as sequence rearrangement, of host genome adjacent to BDV DNA insertions (Fig. 4c). These results indicate that modern BDV is able to produce DNA forms leading to insertion of EBLN-like elements into its host's genome.

This report is the first to provide evidence of endogenous sequences derived from a non-retroviral RNA virus in mammalian species. Phylogenetic analyses demonstrate that the oldest primate EBLN observed must have appeared in an ancestor of primates after the separation between Strepsirrhini and Haplorrhini, implying that bornaviruses have coexisted with primates for an evolutionary history stretching at least 40 million years. Thus, bornaviruses are the first non-retroviral RNA virus whose existence in prehistoric times has been confirmed. To date, the evolution/origin of RNA viruses is a major puzzle in the relationship between viruses and mammalian hosts, because simple molecular clock calculations using an average rate of nucleotide substitutions estimate the origin of RNA viruses to be a very recent event[15–17]. Despite replication during tens of millions of years as exogenous viruses, the amino acid sequences of current BDV N seem surprisingly conserved relative to EBLNs. This conservation demonstrates the inapplicability of simple molecular clocks to RNA virus evolution. Discovery of EBLNs in several

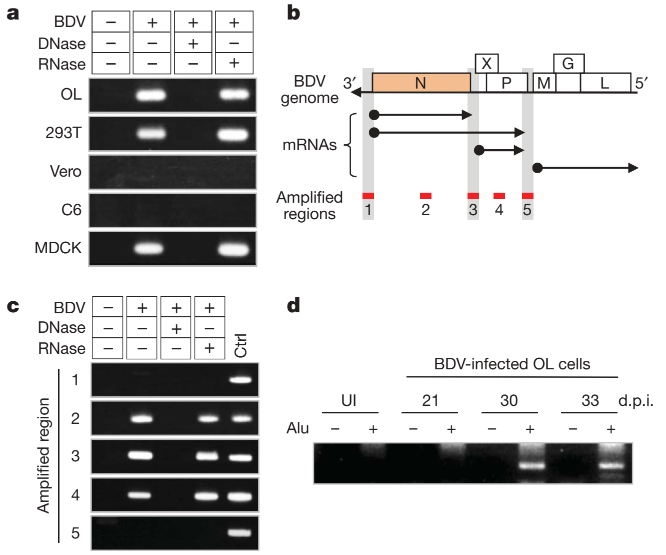


Figure 3 | Reverse transcription and integration of BDV RNA in mammalian cells. **a**, PCR amplification of BDV N-specific cDNA in BDV-infected cells. OL and 293T, human; Vero, monkey; C6, rat; MDCK, dog. **b**, Schematic representation of the bornavirus genome and mRNAs. Regions for the PCR amplification are indicated by red bars. **c**, Region-dependent amplification of BDV cDNA in infected OL cells. The numbers on the left side of the panels correspond to the amplification regions in panel **b**. Ctrl indicates the results of RT–PCR using RNA from BDV-infected OL cells. **d**, Integration of BDV DNA. Genomic DNA was isolated from BDV-infected OL cells at the indicated days after infection, and Alu-PCR was performed with (+) or without (−) the Alu primer. UI, genomic DNA from uninfected cells; d.p.i., days post-infection.

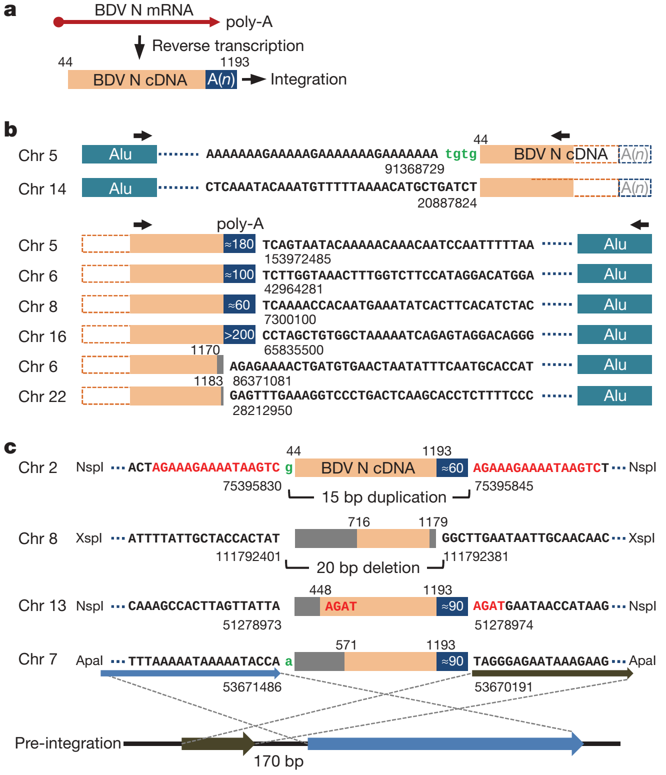


Figure 4 | Structures of BDV N integration events in OL cells. **a**, Structure of BDV N cDNA. The numbering corresponds to nucleotide positions in the BDV genome. The BDV N transcript runs from nucleotide positions 44 to 1193. **b**, **c**, Structures of BDV N integrations detected by Alu-PCR (**b**) and inverse PCR (**c**). Grey rectangles in the N cDNA indicate truncated regions. Black lettering, host genome sequences; green lettering, inserted nucleotides; red lettering, predicted TSDs. The blue box indicates the position and length of the poly-A sequence. The pre-integration form of chromosome 7 is shown in panel **c**.

mammalian species will help shed light on the evolutional history of RNA viruses and their hosts.

The sequence characteristics of both EBLNs and BDV DNA insertions in host genomes indicate that the reverse transcriptase activity encoded by retrotransposons, such as long interspersed nucleotide elements (LINEs), is likely to be involved in the reverse transcription and integration of bornavirus mRNAs, although some clones showed no apparent TSDs (ref. 18). LINE-1s (L1) are abundant retrotransposons, whose enzymes are able to sometimes target cellular mRNAs and produce processed pseudogenes in mammalian genomes[19–21]. The organization of sequences flanking EBLN-2 is consistent with the action of L1. The sequence shows the presence of an AluSx element immediately downstream of the 3′ poly-A tail of EBLN-2 (Supplementary Fig. 11). The key observation is that the EBLN-2/AluSx element is flanked by a perfect 9-bp TSD. Because the AluSx itself is not flanked by TSDs and the 3′ end of Alu is known to be recognized by L1 during target-primed reverse transcription, the presumed EBLN-2/AluSx chimaera element was most likely created and integrated by the L1 machinery. Thus, it is likely that EBLNs are processed pseudogenes derived from ancient bornavirus infections. At present, the reasons why bornaviruses but not other non-retroviral RNA viruses, and why only N and not other genes, have been preserved in mammalian genomes as endogenous elements are not clear. There are several possibilities. First, bornaviruses may have greater access to the germline. Second, the BDV N mRNA, like some cellular RNAs, may have features that, by chance, make it a favourable template for L1-mediated reverse transcription[22,23]. Third, the predominant transcription of BDV N mRNA in infected cells may also favour its association with the L1 replication machinery. The selectivity for BDV N mRNA implies a role for specific structural features, perhaps in conjunction with one or more of the other possibilities. Our data also raise the possibility that, like some endogenous retroviruses, EBLNs may have some function in their host species. An analysis of the non-synonymous to synonymous substitution ratios among anthropoid EBLNs indicates functional, albeit weak, evolutionary conservation. This finding implicates bornaviruses as a new source of genetic innovation in their hosts. Further studies will be needed to explore this possibility.

METHODS SUMMARY

Homology searches (blastp, tblastn) were conducted using the amino acid sequence of BDV N H1499 (International Nucleotide Sequence Database accession number AY374520) as a query and the genomic sequences of 234 eukaryotes as a database at the genomic blast server at the National Center for Biotechnology and Information, NCBI. Sequence hits with *E*-values less than 10^{-10} were collected together with neighbouring hits, if any, with higher *E*-values and combined according to their alignment pattern with BDV N. The resulting amino acid sequence was examined for the presence of a BDV_P40 domain (Pfam accession number PF06407.3) using HMMPFAM. The sequence was identified as a putative EBLN when the domain was detected with the *E*-value of less than 10^{-10}.

The putative EBLN amino acid sequences that were identified with *E*-value of less than 10^{-20} in both tblastn and HMMPFAM were used for the phylogenetic analysis with N sequences of various exogenous bornaviruses. The multiple alignments of EBLN and BDV N amino acid sequences were made according to the alignment pattern of EBLN sequences to BDV N in the tblastn results. The phylogenetic tree was constructed using the neighbour-joining method[24] and the evolutionary distance measured as the proportion of difference (p distance) with the pairwise deletion option in MEGA (version 4.0)[25]. The reliability of interior branches in the phylogenetic tree was assessed by the bootstrap method with 1,000 resamplings.

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

Received 2 September; accepted 17 November 2009.

1. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annu. Rev. Genet.* **42,** 709–732 (2008).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. Zhdanov, V. M. Integration of viral genomes. *Nature* **256,** 471–473 (1975).
4. Klenerman, P., Hengartner, H. & Zinkernagel, R. M. A non-retroviral RNA virus persists in DNA form. *Nature* **390,** 298–301 (1997).
5. Geuking, M. B. *et al.* Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* **323,** 393–396 (2009).
6. Tomonaga, K., Kobayashi, T. & Ikuta, K. Molecular and cellular biology of Borna disease virus infection. *Microbes Infect.* **4,** 491–500 (2002).
7. de la Torre, J. C. Molecular biology of Borna disease virus and persistence. *Front. Biosci.* **7,** d569–d579 (2002).
8. Lipkin, W. I. & Briese, T. in *Fields Virology* 5th edn (eds Knipe, D. M. & Howley, P. M.) 1829–1851 (Lippincott Williams & Wilkins, 2007).
9. Chase, G. *et al.* Borna disease virus matrix protein is an integral component of the viral ribonucleoprotein complex that does not interfere with polymerase activity. *J. Virol.* **81,** 743–749 (2007).
10. Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89 (2007).
11. Mercer, J. M. & Roth, V. L. The effects of Cenozoic global change on squirrel phylogeny. *Science* **299,** 1568–1572 (2003).
12. Kistler, A. L. *et al.* Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. *Virol. J.* **5,** 88 (2008).
13. Francischetti, I. M., My-Pham, V., Harrison, J., Garfield, M. K. & Ribeiro, J. M. Bitis gabonica (Gaboon viper) snake venom gland: toward a catalog for the full-length transcripts (cDNA) and proteins. *Gene* **337,** 55–69 (2004).
14. Hui, E. K., Wang, P. C. & Lo, S. J. Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. *Cell. Mol. Life Sci.* **54,** 1403–1411 (1998).
15. Holmes, E. C. Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77,** 3893–3897 (2003).
16. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Rev. Genet.* **9,** 267–276 (2008).
17. Korber, B., Theiler, J. & Wolinsky, S. Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science* **280,** 1868–1871 (1998).
18. Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genet.* **31,** 159–165 (2002).
19. Maestre, J., Tchenio, T., Dhellin, O. & Heidmann, T. mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* **14,** 6333–6338 (1995).
20. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
21. Kazazian, H. H. Jr. Mobile elements: drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
22. Zhang, Z., Carriero, N. & Gerstein, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **20,** 62–67 (2004).
23. Pavlicek, A. & Jurka, J. in *Genomic disorders* (eds Lupski, J. R. & Stankiewicz, P.) 57–72 (Humana Press, 2006).
24. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
25. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24,** 1596–1599 (2007).

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

Acknowledgements We thank A. Kawahara for helping the capture of the wild shrews (*Sorex unguiculatus* and *Sorex gracillimus*) at Kiritappu wetland, Hokkaido, Japan. We thank I. Francischetti for provision of Gaboon viper (*Bitis gabonica)* venom gland tissue and a cDNA library, D. Vaughan for thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*) brain and liver tissues, and K. Maeda, T. Miyazawa and N. Ohtaki for providing culture cell lines from several mammalian species. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Grants-in-aid for Scientific Research on Priority Areas (Infection and Host Responses; Matrix of Infection Phenomena) (K.T.), PRESTO (RNA and Biofunctions) from Japan Science and Technology Agency (JST) (K.T.), a Health Labour Sciences Research Grants for Research on Measures for Intractable Diseases (H20 nanchi ippan 035) from the Ministry of Health, Labor and Welfare of Japan (K.T.), research grant R37 CA 089441 from the National Cancer Institute (J.M.C.) and a fellowship from the Wenner-Gren Foundation (P.J.). J.M.C. was a Research Professor of the American Cancer Society with support from the George Kirby Foundation.

Author Contributions K.T. designed research; M.H., T.H., T.D. and K.T. conducted experiments using virus and culture systems; T.O. collected samples; Y.S., Y.K. and T.G. performed phylogenetic analysis; M.H., T.H., Y.S., K.I., P.J., T.G., J.M.C. and K.T. analysed data; and M.H., Y.S., P.J., J.M.C. and K.T. wrote the manuscript. All authors discussed the results.

Author Information The TLS EBLN and RBV sequences reported here have been deposited in the DDBJ/EMBL/GenBank and the accession numbers are shown in Figure 2. 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 K.T. (tomonaga@biken.osaka-u.ac.jp).

METHODS

Permutation test. A permutation test was conducted to examine the homology of human EBLNs to the N gene of BDV, taking into account their base composition. The nucleotide sequence of each EBLN was aligned with that of the BDV N gene (strain CRP3A: accession number AY114161) using CLUSTAL W. Gaps were eliminated from the alignment, and the proportion of identical sites (q) was computed. Nucleotide sequences of both the EBLN and the BDV N gene were randomly permuted using pseudorandom numbers, and the q value was computed as indicated above. The permutation process was repeated 10,000 times, and the distribution of the q value between two unrelated sequences of the same base composition as the original EBLN and the N gene was obtained. The probability (p) of observing the q value equal to or greater than the original value in the comparison of unrelated sequences was obtained from the distribution.

Tissue samples. Tissues from three weanling thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*) born in May 2008 (four generations from wild stock) were provided from the Ground Squirrel Captive Breeding Colony at the University of Wisconsin Oshkosh, USA. Immediately after decapitation, brain and liver were rapidly dissected, cut into 5 mm cubes, immersed in chilled methanol, and stored frozen in liquid nitrogen until use. Shrew tissues (brain and liver) were isolated from wild-captured long-clawed shrews (*Sorex unguiculatus*) in Hokkaido, Japan. The shrews were captured under sampling permission of the government of Hokkaido. Immediately after capture, tissue samples were fixed in RNAlater (Ambion) and stored frozen until use. Gaboon viper (*Bitis gabonica*) venom gland tissue was obtained as frozen samples from the Laboratory of Malaria and Vector Research at National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA. Ethanol-fixed tissues from Siberian flying squirrels (*Pteromys volans orii*) and Eurasian red squirrels (*Sciurus vulgaris orientis*) were obtained from the Department of Life Science and Agriculture, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido, Japan.

DNA isolation. Total DNA from cultured cells was isolated using QIAamp DNA Blood Mini kit (Qiagen). Two monkey cell lines, Vero and COS7, used in this study are derived from African green monkey. High molecular mass DNA was extracted by using a Blood and Cell Culture DNA Mini kit (Qiagen). Genomic DNAs of shrews, ground squirrels and the Gaboon viper were prepared from tissue samples using a phenol/chloroform extraction method or the Blood and Cell Culture DNA Mini kit. To minimize the risks of contamination, DNA extraction was performed in UV-irradiated safety cabinet with UV-irradiated pipettes, tubes and filter tips.

DNA samples. Genomic DNAs from chipmunks (*Tamias sibiricus*), Japanese giant flying squirrels (*Petaurista leucogenys*) and red and white giant flying squirrels (*Petaurista alborufus lena*) were obtained from the Department of Life Science and Agriculture, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido, Japan.

Southern blot hybridization. Genomic DNA (5 μg) was digested with appropriate restriction endonucleases (TaKaRa). After electrophoresis in a 0.9% agarose gel, DNA was transferred onto positively charged Nylon membranes (Roche) and baked at 120 °C for 30 min. The membrane was prehybridized in DIG Easy Hyb (Roche) at 32 °C for 30 min. Human and TLS EBLN and BDV N probes were labelled by DIG-High Prime (Roche). Hybridization was performed in DIG Easy Hyb containing 25 ng ml^{-1} probe at 32 °C overnight. The membrane was washed twice with 2× SSC, 0.1% SDS at room temperature for 5 min, and then washed twice with 0.5× SSC, 0.1% SDS at 50 °C for 15 min. For chemiluminescence detection, Anti-DIG-alkaline phosphatase, Fab (Roche) and CDP-Star (Roche) were used according to the manufacturer's instructions. The low-stringency condition can theoretically detect sequences having at least 75% identity with each probe.

F-PERT assay. F-PERT (fluorescent product-enhanced reverse transcriptase) assay was performed as described previously[26]. Briefly, cells were lysed in disruption buffer (40 mM Tris-HCl, pH 8.1; 50 mM KCl; 20 mM dithiothreitol; 0.2% NP-40) and the protein concentration was measured. For the reverse transcription reaction, 1 μg of the cellular protein in 10 μl disruption buffer and an equal volume of 2× RT mix (100 mM KCl; 20 mM Tris-HCl pH 8.3; 11 mM $MgCl_2$; 1 mM dATP, dCTP, dGTP and dTTP; 0.4 μM reverse primer: 5′-CACAGGTCAAACCTCCTAG GAATG-3′, 0.2% NP-40; 20 mM dithiothreitol; 0.8 U μl^{-1} RNasin (Promega); 314 ng μl^{-1} calf thymus DNA (Sigma) and 1.5 ng MS2 RNA (Roche)) were mixed and incubated at 48 °C for 30 min. cDNA was mixed with forward primer: 5′-TCCTGCTCAACTTCCTGTCGAG-3′, reverse primer, probe: 5′-(FAM)-TC TTTAGCGAGACGCTACCATGGCTA-(TAMRA)-3′ and 2× TaqMan Universal PCR Master Mix (Applied Biosystems). Real-time PCR was carried out in an ABI 7900HT Fast Real-Time PCR System using the following parameters: 95 °C 10 min, then 50 cycles consisting of 94 °C for 30 s and 64 °C for 1 min. SuperScript III reverse transcriptase (Invitrogen) was used as standard control.

Virus infection. The BDV strains, huP2br, He/80 and recombinant BDV expressing GFP (rBDV-5′ GFP), were used in this study. Virus stock was prepared from the supernatants of BDV-infected cells. Confluent BDV-infected cells were washed with 20 mM HEPES, pH 7.5 and incubated with 5 ml of 20 mM HEPES (pH 7.5) containing 250 mM $MgCl_2$ and 1% FCS for 1.5 h at 37 °C. Supernatants were harvested and centrifuged at 2,500g for 5 min. The resulting supernatants were used for virus stock. The infectious titre was determined by focus forming assay as described previously[27]. The cell lines used in this study were cultured in Dulbecco's modified Eagle's medium (DMEM)-containing 10% fetal bovine serum (FBS). Newborn Balb/c mice (Oriental kobo) were inoculated intracranially with 200 focus forming units of BDV stock per animal within 24 h after birth. Infected animals were sacrificed at 21 days post-infection. The brains were collected for further analyses. All animal experiments conformed to the guide for the care and use of laboratory animals in the Research Institute for Microbial Diseases, Osaka University, Japan.

Alu-PCR analysis. Integration of BDV sequences into host genomes was detected by using primers specific to human Alu repeats and to BDV N region. First round amplification was performed in a final volume of 25 μl containing 0.5 U Ex Taq (TAKARA), 1× Ex Taq buffer, 0.2 mM dNTP, BDV N-specific primer, Alu primer and 100 ng of high molecular mass genome DNA. As control, PCR without the Alu primer was also performed. The condition of first PCR was as follows: denature for 5 min, 20 cycles of 94 °C for 30 s, 53 °C for 30 s, 72 °C for 4 min, followed by an extended elongation at 72 °C for 10 min. The second round PCR reaction was carried out with 1 μl of the first reaction using BDV N-specific nested primers. The reaction was run as follows: denature for 5 min, 40 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 20 s with the final extension at 72 °C for 3 min. The sequence information for primers used in Alu-PCR is available on request.

Amplification of virus–host junction. Virus–host junctions were amplified by using Alu-PCR and inverse PCR methods. Alu-PCR analysis was performed as described previously[28]. Briefly, the first round PCR reaction was carried out with 100 ng of high molecular mass genome DNA in a final volume of 25 μl containing 0.5 U Ex Taq, 0.2 mM dNTP, 2 μM BDV-specific primer and 0.2 μM Alu primer under the following conditions: denaturing at 94 °C for 1 min, 10 cycles of 94 °C for 30 s, 59 °C for 30 s, 70 °C for 3 min, followed by an extended elongation at 70 °C for 10 min. After amplification, 0.5 U of uracil DNA glycosylase (New England Biolabs) was added into the tubes and incubated at 37 °C for 30 min. After heating at 94 °C for 10 min to break DNA strands at apurinic dUTP sites, the next amplification primers, Tag- and BDV-specific primers, were added. Second round PCR was performed as follows: after denaturing at 94 °C for 2 min, 20 cycles of touch-down PCR in which the annealing temperature was decreased one degree every other cycle from 65 °C to 56 °C. The remaining 20 cycles were run with the annealing temperature at 55 °C, followed by an extended elongation at 72 °C for 3 min. One microlitre of the second round PCR products was further amplified with Tag- and BDV-specific primers as follows: after denaturing for 2 min, 25 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 3 min with the final extension at 72 °C for 3 min. Amplified DNA was electrophoresed, extracted and then sequenced.

Inverse PCR was described elsewhere[29]. Briefly, 1 μg genomic DNA was digested with an appropriate restriction enzyme, including ApaI, BamHI, EcoRI, NspI, PstI or XspI, for 3 h. Digested DNA was purified with QIAquick PCR Purification kit (Qiagen) and diluted with T4 DNA ligase buffer to a final DNA concentration of 1 ng μl^{-1}, and then T4 DNA ligase (New England Biolabs) was added to a final concentration of 4 U μl^{-1}. After ligation at 16 °C for 16 h, ligated DNA was isolated using a QIAquick PCR Purification kit. Five microlitres of the eluate were used for nested PCR. First round PCR was conducted in a 50 μl final volume containing 1 U TaKaRa Ex Taq, 0.2 mM dNTP and 0.2 μM BDV-specific primer set with the following program: after denaturing at 94 °C for 2 min, 20 cycles of 94 °C for 30 s, 70 °C for 30 s (temperature was decreased one degree every other cycle), 72 °C for 4 min and 20 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 4 min with the final extension at 72 °C for 3 min. Second round PCR was performed with 1 μl of the first reaction. The reaction condition was 94 °C for 2 min, 25 cycles of 94 °C for 30 s, 58 °C for 30 s, 72 °C for 4 min with the final extension at 72 °C for 3 min. PCR products were electrophoresed and DNA was extracted from the desired bands and sequenced. Sequence information for primers used in this study is available on request.

26. Lovatt, A. *et al.* High throughput detection of retrovirus-associated reverse transcriptase using an improved fluorescent product enhanced reverse transcriptase assay and its comparison to conventional detection methods. *J. Virol. Methods* **82**, 185–200 (1999).

27. Ohtaki, N. *et al.* Downregulation of an astrocyte-derived inflammatory protein, S100B, reduces vascular inflammatory responses in brains persistently infected with Borna disease virus. *J. Virol.* **81**, 5940–5948 (2007).

28. Minami, M., Poussin, K., Brechot, C. & Paterlini, P. A novel PCR technique using Alu-specific primers to identify unknown flanking sequences from the human genome. *Genomics* **29**, 403–408 (1995).

29. Wo, Y. Y., Peng, S. H. & Pan, F. M. Enrichment of circularized target DNA by inverse polymerase chain reaction. *Anal. Biochem.* **358**, 149–151 (2006).

LETTERS

Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma

R. Eric Davis[1]*, Vu N. Ngo[1]*, Georg Lenz[1]*, Pavel Tolar[3], Ryan M. Young[1], Paul B. Romesser[1,4], Holger Kohlhammer[1], Laurence Lamy[1], Hong Zhao[1], Yandan Yang[1], Weihong Xu[1], Arthur L. Shaffer[1], George Wright[5], Wenming Xiao[6], John Powell[6], Jian-kang Jiang[7], Craig J. Thomas[7], Andreas Rosenwald[8], German Ott[9], Hans Konrad Muller-Hermelink[8], Randy D. Gascoyne[10], Joseph M. Connors[10], Nathalie A. Johnson[10], Lisa M. Rimsza[11,12], Elias Campo[13], Elaine S. Jaffe[2], Wyndham H. Wilson[1], Jan Delabie[14], Erlend B. Smeland[15], Richard I. Fisher[12,16], Rita M. Braziel[12,17], Raymond R. Tubbs[12,18], J. R. Cook[12,18], Dennis D. Weisenburger[19], Wing C. Chan[19], Susan K. Pierce[3] & Louis M. Staudt[1]

A role for B-cell-receptor (BCR) signalling in lymphomagenesis has been inferred by studying immunoglobulin genes in human lymphomas[1,2] and by engineering mouse models[3], but genetic and functional evidence for its oncogenic role in human lymphomas is needed. Here we describe a form of 'chronic active' BCR signalling that is required for cell survival in the activated B-cell-like (ABC) subtype of diffuse large B-cell lymphoma (DLBCL). The signalling adaptor CARD11 is required for constitutive NF-κB pathway activity and survival in ABC DLBCL[4]. Roughly 10% of ABC DLBCLs have mutant CARD11 isoforms that activate NF-κB[5], but the mechanism that engages wild-type CARD11 in other ABC DLBCLs was unknown. An RNA interference genetic screen revealed that a BCR signalling component, Bruton's tyrosine kinase, is essential for the survival of ABC DLBCLs with wild-type CARD11. In addition, knockdown of proximal BCR subunits (IgM, Ig-κ, CD79A and CD79B) killed ABC DLBCLs with wild-type CARD11 but not other lymphomas. The BCRs in these ABC DLBCLs formed prominent clusters in the plasma membrane with low diffusion, similarly to BCRs in antigen-stimulated normal B cells. Somatic mutations affecting the immunoreceptor tyrosine-based activation motif (ITAM) signalling modules[6] of CD79B and CD79A were detected frequently in ABC DLBCL biopsy samples but rarely in other DLBCLs and never in Burkitt's lymphoma or mucosa-associated lymphoid tissue lymphoma. In 18% of ABC DLBCLs, one functionally critical residue of CD79B, the first ITAM tyrosine, was mutated. These mutations increased surface BCR expression and attenuated Lyn kinase, a feedback inhibitor of BCR signalling. These findings establish chronic active BCR signalling as a new pathogenetic mechanism in ABC DLBCL, suggesting several therapeutic strategies.

DLBCL is a heterogeneous diagnostic category consisting of molecularly distinct subtypes that differ in gene expression, oncogenic aberrations and clinical outcome[7,8]. The ABC DLBCL subtype relies on constitutive NF-κB signalling to block apoptosis, but the germinal-centre B-cell-like (GCB) subtype does not[9]. Recurrent CARD11 mutations in ABC DLBCL provided genetic evidence that NF-κB signalling is central to its pathogenesis[5]. However, most ABC DLBCLs have wild-type CARD11 yet nonetheless rely on CARD11 to activate NF-κB signalling[4,9].

In normal B cells, CARD11 is engaged during antigenic stimulation of BCR signalling. Antigen specificity of the BCR is provided by surface immunoglobulin, but signalling is mediated by two associated proteins, CD79A (Ig-α) and CD79B (Ig-β)[10]. The CD79A–CD79B heterodimer is a scaffold for the assembly and membrane expression of the BCR and also initiates downstream signalling to the NF-κB, phosphatidylinositol-3-OH kinase, extracellular signal-regulated kinase (ERK) mitogen-activated protein (MAP) kinase and NF-AT pathways. Engagement of the BCR by antigen induces Src-family kinases to phosphorylate tyrosines in the ITAM motifs of CD79A and CD79B. The tyrosine kinase Syk is activated by binding to the phosphorylated ITAMs, triggering a signalling cascade that involves the tyrosine kinase Bruton's tyrosine kinase (BTK), phospholipase Cγ and protein kinase Cβ (PKC-β). PKC-β phosphorylates CARD11, causing it to recruit BCL10 and MALT1 into a multiprotein 'CBM' complex that activates IκB kinase (IKK), thereby initiating NF-κB signalling.

A potential role for BCR signalling in ABC DLBCLs with wild-type CARD11 was revealed by an RNA interference screen. Two short hairpin RNAs (shRNAs) targeting the BCR pathway component BTK were highly toxic for an ABC DLBCL line with wild-type CARD11 (OCI-Ly10) but not for one with mutant CARD11 (OCI-Ly3), nor for GCB DLBCL and multiple myeloma lines (Fig. 1a and Supplementary Fig. 1). In subsequent survival assays, a BTK shRNA was toxic for four ABC DLBCL lines with wild-type CARD11 but not for OCI-Ly3 or six GCB DLBCL lines (Fig. 1b). BTK kinase activity

[1] Metabolism Branch, [2] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [3] Laboratory of Immunogenetics, National Institute for Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20852, USA. [4] Howard Hughes Medical Institute – National Institutes of Health Research Scholars Program, Bethesda, Maryland 20892, USA. [5] Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [6] Bioinformatics and Molecular Analysis Section, Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, USA. [7] NIH Chemical Genomics Center, National Human Genome Research Institute, National Institutes of Health, 9800 Medical Center Drive, Rockville, Maryland 20850, USA. [8] Department of Pathology, University of Würzburg, 97080 Würzburg, Germany. [9] Department of Clinical Pathology, Robert-Bosch-Krankenhaus, and Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, 70376 Stuttgart, Germany. [10] British Columbia Cancer Agency, Vancouver, British Columbia, Canada V5Z 4E6. [11] Department of Pathology, University of Arizona, Tucson, Arizona 85724, USA. [12] Southwest Oncology Group, 24 Frank Lloyd Wright Drive, Ann Arbor, Michigan 48106, USA. [13] Hospital Clinic, University of Barcelona, 08036 Barcelona, Spain. [14] Pathology Clinic, Rikshospitalet University Hospital, N-0310 Oslo, Norway. [15] Institute for Cancer Research, Rikshospitalet University Hospital and Center for Cancer Biomedicine, Faculty Division of the Norwegian Radium Hospital, University of Oslo, N-0310 Oslo, Norway. [16] James P. Wilmot Cancer Center, University of Rochester School of Medicine, Rochester, New York 14642, USA. [17] Oregon Health and Science University, Portland, Oregon 97239, USA. [18] Cleveland Clinic Pathology and Laboratory Medicine Institute, Cleveland, Ohio 44195, USA. [19] Departments of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA.
*These authors contributed equally to this work.

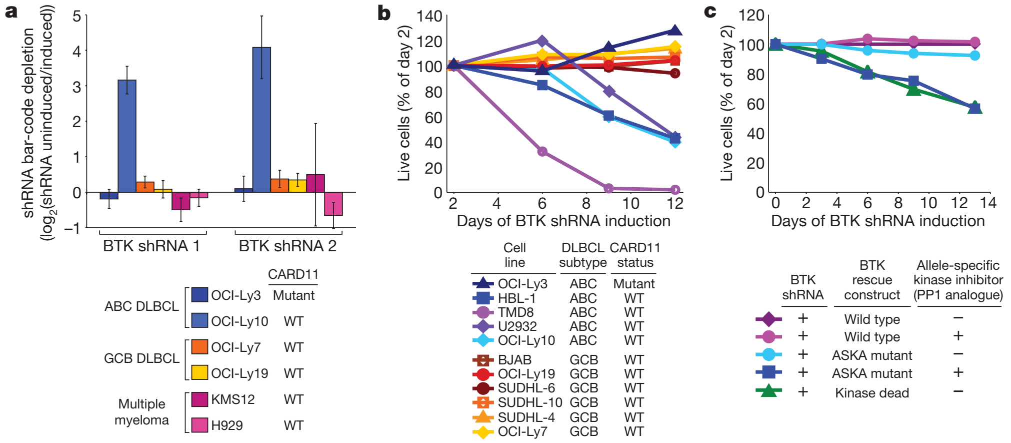


Figure 1 | BTK is a critical kinase for survival of ABC DLBCL cells. **a**, RNA interference screen in lymphoma and multiple myeloma cell lines. An shRNA library targeting 442 kinases was screened in the indicated cell lines as described[4]. Shown is the selective toxicity of two BTK shRNAs after three weeks in culture. Results are shown as means ± s.d. for four independent transductions. WT, wild type. **b**, Selective toxicity of a BTK shRNA for ABC DLBCLs with wild-type CARD11. DLBCL cell lines were infected with a retrovirus that expresses BTK shRNA 1 together with green fluorescent protein (GFP). Shown is the fraction of GFP-positive cells relative to the GFP-positive fraction on day 2. **c**, BTK kinase activity is required for survival of ABC DLBCL cells. OCI-Ly10 cells were transduced with cDNAs encoding wild-type or mutant BTK (kinase-dead allele or analogue-sensitive kinase allele (ASKA)[29]). Wild-type but not kinase-dead BTK rescued cells with endogenous BTK knockdown. The ASKA isoform-specific kinase inhibitor 1-NM-PP1 (2 mM) killed cells bearing the BTK ASKA allele.

was required for the rescue of ABC DLBCL lines from the toxicity of BTK knockdown (Fig. 1c).

The role of BTK in BCR signalling prompted us to investigate the reliance of ABC DLBCLs on other BCR pathway components. A CD79A shRNA killed all four ABC DLBCL lines with wild-type CARD11 but not the line with mutant CARD11 or the GCB DLBCL lines (Fig. 2a). In contrast, a CARD11 shRNA killed all ABC DLBCL lines, and a control shRNA was non-toxic. In the line

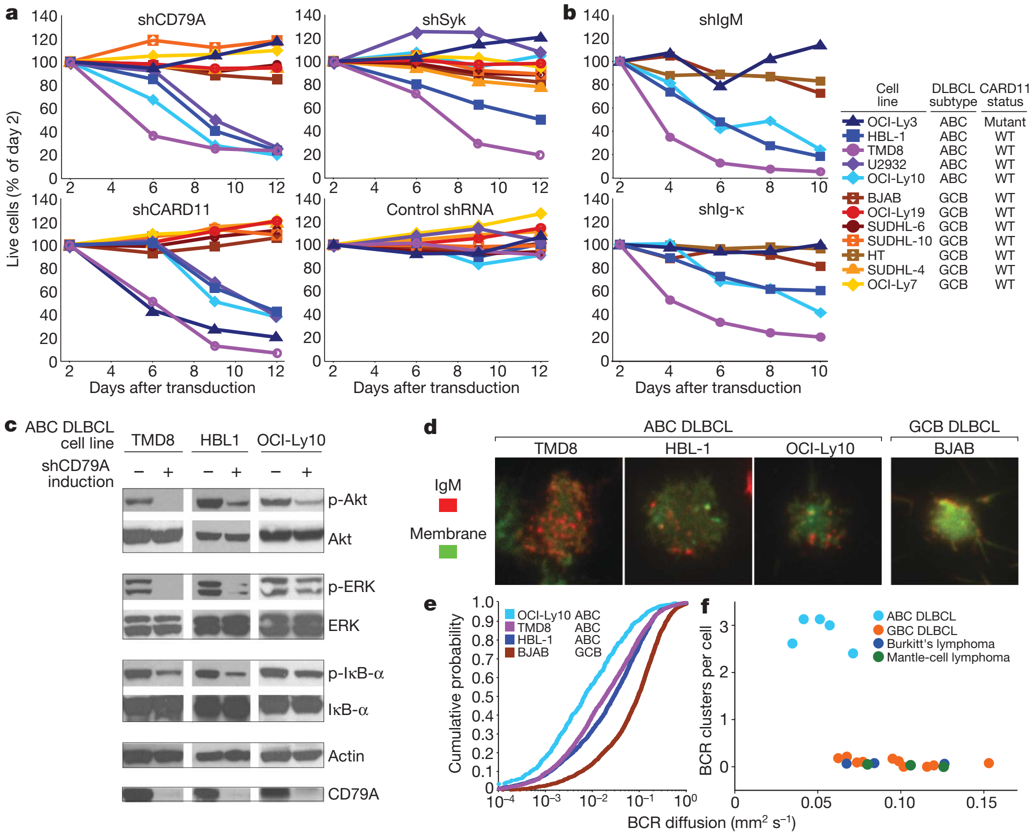


Figure 2 | Chronic active BCR signalling in ABC DLBCL lines. **a**, Survival of DLBCL cell lines after shRNA-mediated knockdown of BCR signalling components CD79A, Syk and CARD11. **b**, Knockdown of immunoglobulin heavy or light chain is toxic for ABC DLBCLs with chronic active BCR signalling. **c**, Phosphoproteins in multiple signalling pathways depend on chronic active BCR signalling. The indicated ABC DLBCL cell lines were transduced with an shRNA targeting CD79A and phosphorylated, or total proteins were assessed by western blotting before and after shRNA induction for 48 h. **d**, Clustering of IgM in the plasma membrane was observed only in ABC DLBCL lines with chronic active BCR signalling, using TIRF microscopy. Plasma membrane density was revealed by membrane dye R18. **e**, Decreased diffusion of surface IgM in ABC DLBCL lines with chronic active BCR signalling compared with the GCB DLBCL line, as quantified by TIRF microscopy. **f**, Immobile BCR clusters are characteristic of lines representing ABC DLBCL but not other lymphoma types.

HBL-1, the knockdown of surface CD79A expression by different shRNAs caused a proportional decrease in surface IgM, implying that the toxicity of CD79A knockdown was due to a loss of surface BCR (Supplementary Fig. 2a). CD79B shRNAs were also toxic to ABC DLBCLs, and the degree of CD79B knockdown was proportional to the decrease in surface BCR and to toxicity (Supplementary Fig. 2b, c). To investigate the role of the immunoglobulin receptor, we developed shRNAs targeting IgM and Ig-κ (Supplementary Fig. 3). These shRNAs were also selectively toxic to ABC DLBCLs with wild-type CARD11, establishing a direct role for immunoglobulin in this signalling (Fig. 2b).

The NF-κB pathway is activated by BCR signalling in ABC DLBCLs because knockdown of BTK, CD79A, CD79B and CARD11 decreased the expression of NF-κB target genes and inhibited IKK (Supplementary Fig. 4). BCR signalling also activates the phosphatidylinositol-3-OH kinase and ERK MAP kinase pathways in these cells, because CD79A knockdown inhibited phosphorylation of Akt and ERK in addition to IκB-α (Fig. 2c).

A subsequent focused shRNA screen suggested that other BCR signalling components contribute to chronic active BCR signalling, including Syk, BLNK, phospholipase Cγ2 and PKC-β (Supplementary Fig. 5). A Syk shRNA killed two ABC DLBCL lines with wild-type CARD11 (HBL-1 and TMD8) but not two others (OCI-Ly10 and U2932), and also had no effect on OCI-Ly3 or GCB DLBCL lines (Fig. 2a), despite comparable knockdown (Supplementary Fig. 6a). Not only was OCI-Ly10 insensitive to Syk knockdown but it also died with ectopic expression of wild-type but not kinase-dead Syk (Supplementary Fig. 6b). A previous study with a Syk inhibitor, R406, concluded that most DLBCLs rely on tonic BCR signalling[11]. However, R406 killed Syk-independent GCB and ABC DLBCL lines (including OCI-Ly10), suggesting that its toxicity in these lines may be due to inhibition of other kinases and not BCR signalling (Supplementary Fig. 6c).

We next used total internal reflection fluorescence (TIRF) microscopy to reveal BCRs on the surface of lymphoma lines. In normal mouse B cells, TIRF microscopy revealed that antigen exposure causes BCRs to form clusters with decreased diffusion, leading to BCR signalling[12]. All five ABC DLBCL lines had prominent BCR clusters that were not present in 16 other lines derived from GCB DLBCL, Burkitt's lymphoma or mantle-cell lymphoma (Fig. 2d, f). BCR clusters were also observed in biopsies from three patients with ABC DLBCL (Supplementary Fig. 7a). Moreover, the BCRs in ABC DLBCLs diffused less rapidly than those in other lymphoma lines (Fig. 2e, f). We observed a correlation between BCR clusters and phosphotyrosine accumulation in ABC DLBCL lines, suggesting that these structures may be actively signalling (Supplementary Fig. 7b).

Taken together, these findings establish a continuing requirement for proximal BCR signalling in ABC DLBCLs with wild-type CARD11. Because these lines also depend on CARD11, like antigen-activated normal B cells, we refer to this phenomenon as 'chronic active' BCR signalling. We wish to distinguish chronic active BCR signalling from 'tonic' BCR signalling. Tonic BCR signalling promotes cell survival in all mature mouse B cells[13,14], but mice deficient in CBM components have relatively normal numbers of mature follicular B cells[15]. It therefore seems likely that CARD11 is not essential for tonic BCR signalling but is required for chronic active BCR signalling. Moreover, chronic active BCR signalling is characterized by BCR clustering, which is not observed in resting mouse B cells that depend on tonic BCR signalling[12].

To provide genetic evidence of BCR signalling in the pathogenesis of ABC DLBCL, we resequenced genes in the BCR pathway in DLBCL cell lines and biopsies. We identified missense mutations affecting the first tyrosine of the CD79B ITAM motif in two cell lines, HBL-1 (Y196F) and TMD8 (Y196H) (Fig. 3a). Both lines were heterozygous for this mutation, but more than 90% of the CD79B messenger RNA in HBL-1 was derived from the mutant allele (data not shown). These mutations prompted us to resequence the CD79B ITAM region in 225 DLBCL biopsies. In 18% (29 out of 161) of ABC DLBCLs, the first ITAM tyrosine was replaced by a variety of amino acids as a result of point mutations; in one case, this residue was removed by a three-base-pair deletion (Fig. 3a, b). Less common were missense mutations in other ITAM residues and deletions that disrupted all or part of the motif. Of 64 GCB DLBCLs, only one had a mutation affecting the first ITAM tyrosine and one other had a different ITAM mutation (L199Q). Overall, the frequency of CD79B ITAM mutations was significantly higher in ABC DLBCL (21.1%) than in GCB DLBCL (3.1%) ($P = 8.9 \times 10^{-4}$). CD79B ITAM mutations were not present in 20 Burkitt's lymphoma and 16 gastric mucosa-associated lymphoid tissue (MALT) lymphoma biopsies. In six cases of ABC DLBCL, analysis of non-malignant tissue established that the CD79B mutations were somatically acquired by the malignant cells (Supplementary Fig. 8).

The CD79A ITAM region of the ABC DLBCL line OCI-Ly10 has a splice-donor-site mutation[16] causing an 18-amino-acid deletion that removes most of the ITAM, including the second tyrosine. Though OCI-Ly10 was heterozygous for this mutation, more than 90% of the CD79A mRNA was mutated (data not shown). One ABC DLBCL biopsy had a similar splice-site mutation and another had mutations that deleted the entire CD79A ITAM (Fig. 3a). CD79A mutations were rare among ABC DLBCLs, occurring in 2.9% (2 out of 68) of biopsies.

In mouse B cells, mutations in the CD79A or CD79B ITAM tyrosine residues elevate surface BCR expression by inhibiting receptor internalization[17]. Indeed, GCB DLBCL cells reconstituted with CD79A or CD79B mutants derived from ABC DLBCLs had more surface IgM expression than cells with wild-type isoforms, but this was not true of CD79 ITAM mutations that were not observed in samples from patients (Fig. 3c). Similarly, ABC DLBCL cells reconstituted with mutant CD79B had higher surface BCR expression than those reconstituted with wild-type CD79B (Fig. 3d). Interruption of chronic active BCR signalling with the kinase inhibitor dasatinib (see later) increased surface BCR expression in ABC DLBCL cells with wild-type but not mutant CD79B (Fig. 3d). Hence, one function of the CD79 mutations is to maintain surface BCR expression in the face of chronic active BCR signalling.

We speculated that the CD79B mutations might be genetically selected in ABC DLBCLs for their ability to circumvent negative regulatory circuits that attenuate BCR signalling. Whereas several Src-family tyrosine kinases can initiate BCR signalling, Lyn is unique in mediating negative feedback on BCR signalling[18]. Indeed, Lyn-deficient mice succumb to an autoimmune disease that has been traced to BCR hyperactivity[19]. Lyn is required for BCR internalization[20,21], suggesting that CD79 mutations might elevate surface BCR expression by inhibiting Lyn. To test this, we knocked down endogenous CD79B expression in HBL-1 and TMD8 cells, both of which harbour a CD79B mutation, and complemented them with exogenous wild-type or mutant CD79B complementary DNAs. Immunoprecipitation of Lyn followed by an *in vitro* kinase assay showed greater Lyn kinase activity in cells reconstituted with wild-type CD79B (Fig. 3e). These data suggest a model in which CD79B mutations are selected in ABC DLBCLs to attenuate negative autoregulation by Lyn during chronic active BCR signalling.

The CD79 mutants are not loss-of-function mutants because they prevented death of ABC DLBCL cells caused by knockdown of endogenous CD79 isoforms (Supplementary Fig. 9). However, the CD79 mutants were not functionally superior to their wild-type counterparts in this assay (Supplementary Fig. 9) and did not spontaneously activate NF-κB when introduced into GCB DLBCL cells, unlike CARD11 mutants[5] (data not shown). We therefore propose that the CD79 ITAM mutations may be selected early in the genesis of the malignant clone, perhaps to allow it to respond abnormally well to a self or foreign antigen (Supplementary Fig. 10). In this regard it is notable that mutations that impair CD79A or CD79B ITAM function in mouse B cells lead to exaggerated antigenic responses[17,22,23]. Future research should investigate the potential role of antigenic stimulation

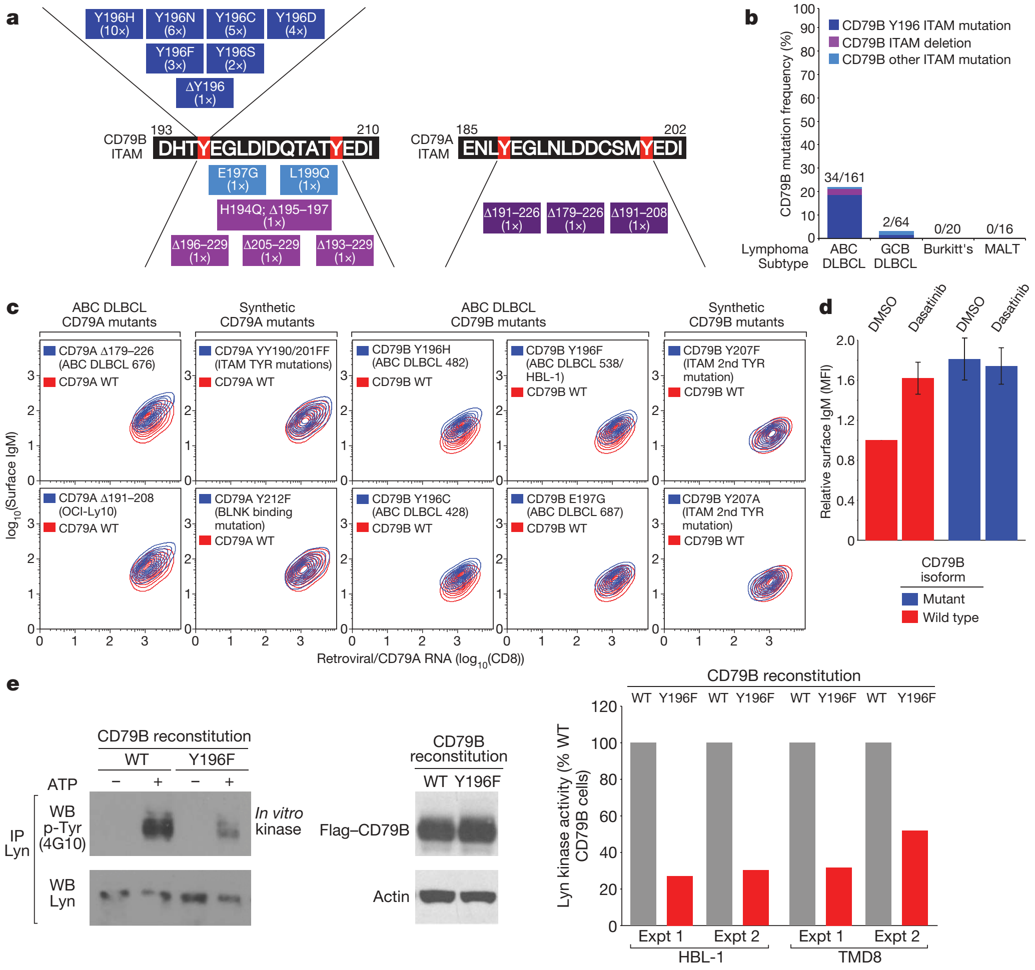


Figure 3 | CD79A and CD79B ITAM mutations in ABC DLBCL. **a**, CD79B and CD79A ITAM mutations in DLBCL biopsies and lines (case number in parenthesis). **b**, CD79B ITAM mutation frequencies in lymphoma biopsies. **c**, Mutant CD79A and CD79B isoforms increase surface IgM. The GCB DLBCL line BJAB was reconstituted with either wild-type or mutant CD79A/B proteins. Surface IgM is depicted relative to CD79 RNA levels, estimated with bicistronic expression of CD8. 'Synthetic' mutants were not observed in patient samples. **d**, CD79B mutations prevent downmodulation of surface BCR by BCR signalling. The ABC DLBCL line HBL-1 was reconstituted with wild-type or Y196H mutant CD79B and treated for 24 h with dimethylsulphoxide (DMSO) or dasatinib, a BCR signalling inhibitor. Surface IgM (mean fluorescence intensity; MFI) is depicted relative to the levels in cells with wild-type CD79B treated with DMSO. Results are shown as means ± s.e.m. for two experiments. **e**, CD79B mutations inhibit Lyn kinase activity in ABC DLBCLs. The indicated ABC DLBCL lines were reconstituted with wild-type or Y196F mutant CD79B. Lyn kinase activity in immunoprecipitates (IP) was estimated by densitometric analysis of western blots (WB) as phospho-Lyn (using anti-phosphotyrosine antibody 4G10) relative to total Lyn.

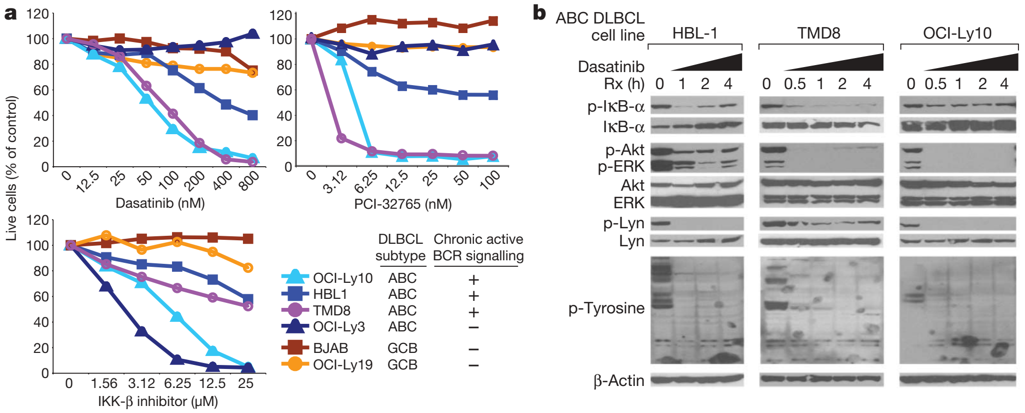


Figure 4 | Therapeutic strategies to target chronic active BCR signalling. **a**, Viability of DLBCL lines assessed by assay with 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) after four days of treatment with various doses of dasatinib, the BTK inhibitor PCI-32765 (compound 13 in ref. 25) or an IKK-β inhibitor[27]. **b**, Effect of dasatinib on phosphoprotein levels in ABC DLBCL cells. Three ABC DLBCL lines were treated with 50 nM dasatinib for the indicated durations and analysed by western blotting. Rx, treatment.

in chronic active BCR signalling and in the spontaneous BCR clustering that characterizes ABC DLBCLs. BCR clustering does not depend on the CD79B mutations (Supplementary Fig. 11), suggesting that other mechanisms contribute to this aspect of chronic active BCR signalling.

We considered therapeutic strategies to exploit chronic active BCR signalling in ABC DLBCL. Dasatinib, a kinase inhibitor approved for the treatment of chronic myelogenous leukaemia, inhibits Src-family kinases and BTK[24]. Dasatinib killed ABC DLBCL lines that rely on chronic active BCR signalling but not the BCR-independent line OCI-Ly3 or GCB DLBCL lines (Fig. 4a). A selective BTK inhibitor, PCI-32765 (ref. 25), was also selectively toxic to cell lines with chronic active BCR signalling (Fig. 4a). By contrast, all ABC DLBCL lines were sensitive to an IKK-β inhibitor. In BCR-dependent lines, dasatinib decreased the phosphorylation of IκB-α, Akt, ERK and Lyn, as well as total protein tyrosine phosphorylation and IKK activity (Fig. 4b and Supplementary Fig. 12). Dasatinib toxicity may therefore be due to NF-κB inhibition, which causes apoptosis, and Akt/mTOR inhibition, which causes 'metabolic catastrophe'[26]. Indeed, rapamycin, an mTOR inhibitor, synergized with an IKK-β inhibitor in killing ABC DLBCL lines with chronic active BCR signalling (Supplementary Fig. 13). Our studies suggest that the position of molecular lesions in the BCR and NF-κB signalling pathways could be used to guide the therapy of ABC DLBCL. ABC DLBCLs with wild-type CARD11 and chronic active BCR signalling might respond to a BTK inhibitor, such as PCI-32765, and possibly to inhibitors of Src-family kinases, PKC-β or Syk, in some cases. By contrast, CARD11-mutant tumours would need to be treated with agents that target downstream components of the NF-κB pathway such as IKK[27]. A precise delineation of which ABC DLBCL cases depend on chronic active BCR signalling awaits the development of predictive biomarkers and the results of clinical trials involving BCR signalling inhibitors.

METHODS SUMMARY

Cell lines possessing the ecotropic retroviral receptor and the tetracycline repressor were generated and used in RNA interference library screening, shRNA toxicity assays and cDNA complementation studies as described[4]. DLBCL cell lines were assigned to the ABC or GCB subtypes by gene expression profiling[4] (Supplementary Fig. 14). shRNA screening results are given in Supplementary Tables 1 and 3, and shRNA sequences are listed in Supplementary Tables 2 and 3. Specific shRNA-mediated mRNA and protein knockdown was documented (Fig. 2c and Supplementary Figs 6a and 15). IKK reporter lines were engineered to express an IκB-α–*Photinus* luciferase fusion and *Renilla* luciferase[27]. TIRF imaging of the BCR was based on techniques described previously[12].

Tumour biopsies were obtained before treatment from patients with *de novo* DLBCL[28], gastric MALT lymphoma and Burkitt's lymphoma. All samples were studied in accordance with a protocol approved by the National Cancer Institute Institutional Review Board.

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

Received 18 April; accepted 4 November 2009.

1. Klein, G. & Klein, E. Conditioned tumorigenicity of activated oncogenes. *Cancer Res.* **46,** 3211–3224 (1986).
2. Bahler, D. W. & Levy, R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. *Proc. Natl Acad. Sci. USA* **89,** 6770–6774 (1992).
3. Refaeli, Y. *et al.* The B cell antigen receptor and overexpression of MYC can cooperate in the genesis of B cell lymphomas. *PLoS Biol.* **6,** e152 (2008).
4. Ngo, V. N. *et al.* A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441,** 106–110 (2006).
5. Lenz, G. *et al.* Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319,** 1676–1679 (2008).
6. Reth, M. Antigen receptor tail clue. *Nature* **338,** 383–384 (1989).
7. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).
8. Staudt, L. M. & Dave, S. The biology of human lymphoid malignancies revealed by gene expression profiling. *Adv. Immunol.* **87,** 163–208 (2005).
9. Davis, R. E., Brown, K. D., Siebenlist, U. & Staudt, L. M. Constitutive nuclear factor κB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **194,** 1861–1874 (2001).
10. Dal Porto, J. M. *et al.* B cell antigen receptor signaling 101. *Mol. Immunol.* **41,** 599–613 (2004).
11. 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).
12. Tolar, P., Hanna, J., Krueger, P. D. & Pierce, S. K. The constant region of the membrane immunoglobulin mediates B cell-receptor clustering and signaling in response to membrane antigens. *Immunity* **30,** 44–55 (2009).
13. Lam, K. P., Kuhn, R. & Rajewsky, K. *In vivo* ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death. *Cell* **90,** 1073–1083 (1997).
14. Kraus, M., Alimzhanov, M. B., Rajewsky, N. & Rajewsky, K. Survival of resting mature B lymphocytes depends on BCR signaling via the Igα/β heterodimer. *Cell* **117,** 787–800 (2004).
15. Thome, M. CARMA1, BCL-10 and MALT1 in lymphocyte development and activation. *Nature Rev. Immunol.* **4,** 348–359 (2004).
16. Gordon, M. S., Kanegai, C. M., Doerr, J. R. & Wall, R. Somatic hypermutation of the B cell receptor genes *B29* (*Igβ*, CD79b) and *mb1* (*Igα*, CD79a). *Proc. Natl Acad. Sci. USA* **100,** 4126–4131 (2003).
17. Gazumyan, A., Reichlin, A. & Nussenzweig, M. C. Igβ tyrosine residues contribute to the control of B cell receptor signaling by regulating receptor internalization. *J. Exp. Med.* **203,** 1785–1794 (2006).
18. Gauld, S. B. & Cambier, J. C. Src-family kinases in B-cell development and signaling. *Oncogene* **23,** 8001–8006 (2004).
19. Chan, V. W., Meng, F., Soriano, P., DeFranco, A. L. & Lowell, C. A. Characterization of the B lymphocyte populations in Lyn-deficient mice and the role of Lyn in signal initiation and down-regulation. *Immunity* **7,** 69–81 (1997).
20. Niiro, H. *et al.* The B lymphocyte adaptor molecule of 32 kilodaltons (Bam32) regulates B cell antigen receptor internalization. *J. Immunol.* **173,** 5601–5609 (2004).
21. Ma, H. *et al.* Visualization of Syk-antigen receptor interactions using green fluorescent protein: differential roles for Syk and Lyn in the regulation of receptor capping and internalization. *J. Immunol.* **166,** 1507–1516 (2001).
22. Kraus, M., Saijo, K., Torres, R. M. & Rajewsky, K. Ig-α cytoplasmic truncation renders immature B cells more sensitive to antigen contact. *Immunity* **11,** 537–545 (1999).
23. Torres, R. M. & Hafen, K. A negative regulatory role for Ig-α during B cell development. *Immunity* **11,** 527–536 (1999).
24. Hantschel, O. *et al.* The Btk tyrosine kinase is a major target of the Bcr-Abl inhibitor dasatinib. *Proc. Natl Acad. Sci. USA* **104,** 13283–13288 (2007).
25. Pan, Z. *et al.* Discovery of selective irreversible inhibitors for Bruton's tyrosine kinase. *ChemMedChem* **2,** 58–61 (2007).
26. Jin, S., DiPaola, R. S., Mathew, R. & White, E. Metabolic catastrophe as a means to cancer cell death. *J. Cell Sci.* **120,** 379–383 (2007).
27. Lam, L. T. *et al.* Small molecule inhibitors of IκB-kinase are selectively toxic for subgroups of diffuse large B cell lymphoma defined by gene expression profiling. *Clin. Cancer Res.* **11,** 28–40 (2005).
28. Lenz, G. *et al.* Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **359,** 2313–2323 (2008).
29. Blethrow, J., Zhang, C., Shokat, K. M. & Weiss, E. L. Design and use of analog-sensitive protein kinases. *Curr. Protocols Mol. Biol.* Unit 18.11, doi:10.1002/0471142727.mb1811s66 (2004).

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

Acknowledgements We thank L. Honigberg for PCI-32765, L. Dang for IKK-β inhibitors, S. Ansher for dasatinib, and S. Tohda for the TMD8 cell line. This research was supported by the Intramural Research Program of the National Institutes of Health, the National Cancer Institute, the Center for Cancer Research, the National Institute of Allergy and Infectious Disease, and the National Human Genome Research Institute. P.B.R. was a Howard Hughes Medical Institute-National Institutes of Health Research Scholar.

Author Contributions R.E.D., V.N.N., G.L., P.T., R.M.Y., P.B.R., H.K., L.L. and A.L.S. designed and performed experiments. H.Z., Y.Y. and W.X. performed experiments. G.W., W.X. and J.P. analysed data. J.J. and C.J.T. synthesized reagents. A.R., G.O., H.K.M.-H., R.D.G., J.M.C., N.A.J., L.M.R., E.C., E.S.J., W.H.W., J.D., E.B.S., R.I.F., R.M.B., R.R.T., J.R.C., D.D.W. and W.C.C. supplied samples from patients and reviewed pathological and clinical data. S.K.P. supervised research. L.M.S. designed and supervised research and wrote the manuscript.

Author Information Gene expression profiling data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE18817. 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 L.M.S. (lstaudt@mail.nih.gov).

METHODS

Cell lines. All cell lines were maintained in a humidified 5% CO_2 incubator at 37 °C. Cell lines were grown in RPMI 1640 medium supplemented with glutamine, 2-mercaptoethanol, penicillin/streptomycin and 10% fetal bovine serum, except for OCI-Ly3 and OCI-Ly10, which were maintained in Iscove's modified Dulbecco's medium supplemented with 2-mercaptoethanol, penicillin/streptomycin and 20% heparinized human plasma. All cell lines had previously been modified to express an ecotropic retroviral receptor and a fusion protein of the Tet repressor and the blasticidin resistance gene, as described previously[4].

Retroviral transductions. Retroviral supernatants were prepared as described previously[5]. In brief, Lipofectamine 2000 (Invitrogen) was used to transfect 293T producer cells with a plasmid mixture for *gag* and *pol*, a mutant ecotropic *env*, and each particular retrovirus. After two days, supernatant was passed through a 0.45-μM filter, mixed with Polybrene (8 μg ml^{-1}) and used for centrifugal transduction of target cells expressing the ecotropic retroviral receptor. In some instances a second infection was performed, using fresh supernatant collected three days after transfection of producer cells.

shRNA library screening. shRNA library screening was performed as described[4]. The shRNA library was constructed in a pMSCV-based retroviral vector (pRSMX_Puro) with two expression cassettes, one for constitutive expression of a selectable marker (conferring puromycin resistance) and the other for inducible expression of shRNA, after release of binding of the bacterial tetracycline repressor by doxycycline (20 ng ml^{-1}).

For the focused screen of BCR pathway genes (Supplementary Fig. 5), the same procedure was applied except that the abundance of each shRNA was enumerated with an Illumina GAII sequencer. Comparison was made between an shRNA-uninduced sample from day 0 and an shRNA-induced sample from day 24 of culture. The magnitude and standard error of the shRNA effect were calculated by using a logistic regression model to estimate the relative probability that an shRNA was found in the shRNA-induced versus uninduced samples. This model included a normalization factor that was a constant for all genes from a given shRNA pool in a given experiment and including normal random effects representing gene by experiment interactions.

Other shRNA vectors and shRNA sequences. Modifications were made to the puromycin resistance gene cassette in the pMSCV-based retroviral shRNA vector, including fusion of the puromycin resistance gene to enhanced green fluorescent protein (EGFP) (pRSMX_PuroGFP) as described previously[4] or replacement of the puromycin resistance gene with the *Escherichia coli* gene encoding inosine 5′-monophosphate dehydrogenase (IMPDH) to allow selection in mycophenolic acid. shRNA sequences used in this study are presented in Supplementary Tables 2 and 3.

Expression vectors and cDNA mutagenesis or modification. Retroviral vectors for inducible cDNA expression were either pBMN-based (http://www.stanford.edu/group/nolan/plasmid_maps/pmaps.html) or pMSCV-based with the cDNA expressed from a doxycycline-inducible cytomegalovirus (CMV) promoter in which a binding site for the bacterial tetracycline repressor is inserted at the transcription start site (derived from pCDNA4/TO (Invitrogen)). All mutagenesis was performed with the QuikChange kit from Stratagene and verified by dye termination sequencing. Modification of the CD79A and CD79B cDNAs for the surface expression experiments shown in Fig. 4c–e included insertion of a Flag peptide coding sequence just downstream of the signal peptide cleavage site, and fusion to EGFP at the 3′ end, after removal of the stop codon and addition of a six-amino-acid spacer as described previously[30]. 3′ GFP fusions were also constructed for Syk (Supplementary Fig. 6b).

shRNA toxicity and complementation assays. The toxicity of individual shRNA sequences was chiefly tested with the PuroGFP vector as described previously[4]. In brief, after infection with a retrovirus expressing shRNA and GFP, FACS was used to determine the fraction of live cells that were GFP-positive two days after transduction. Doxycycline was then added to induce shRNA expression, and the fraction of GFP-positive live cells was determined at various intervals during subsequent culture. Parallel cultures were prepared with a vector expressing a control shRNA. The GFP-positive fraction from the test shRNA cultures was normalized both to the GFP-positive fraction from the control shRNA culture on the same day and to the GFP-positive fraction on day 2. For CD79B (Supplementary Fig. 2b, c), adequate knockdown by shRNA required the selection of shRNA- and GFP-expressing single-cell clones by limiting dilution. Toxicity assays of these clones were performed by mixing the clones with untransduced cells and determining an initial GFP-positive fraction before the addition of doxycycline to induce shRNA expression. The fraction of GFP-positive cells at various time points was normalized to this initial value.

For complementation studies with CD79A (Supplementary Fig. 9), OCI-Ly10 cells were infected with retroviruses expressing wild-type or mutant CD79A coding regions along with Lyt2 (mouse CD8). Subsequently, cells were infected with a retrovirus expressing a CD79A shRNA (targeting the 3′ untranslated region (UTR)) and GFP and the fraction of Lyt2-positive/GFP-positive cells was monitored over time as above.

For complementation studies with CD79B (Supplementary Fig. 9), HBL-1 cells were infected with a retrovirus expressing a CD79B shRNA (targeting the 3′ UTR) and GFP, selected in puromycin, and single-cell cloned. A clone was selected with the best doxycycline-inducible knockdown of CD79B as assessed by FACS. This clone was infected with a retrovirus expressing a wild-type or mutant CD79B coding region along with Lyt2 (mouse CD8). These GFP-positive cells were mixed with unmodified HBL-1 cells and the fraction of GFP-positive/Lyt2-positive cells was monitored over time, as above.

To analyse the influence of the CD79B mutations on chronic active BCR signalling in ABC DLBCLs (Fig. 4d, e), HBL1 or TMD8 cells were infected with a retroviral vector with doxycycline-inducible expression of an shRNA directed at the CD79B 3′ UTR-directed and an IMPDH drug selection marker. After selection in mycophenolic acid, single-cell clones were assayed for doxycycline-inducible knockdown of CD79B by FACS. Clones were subsequently infected with a retrovirus expressing wild-type or mutant (Y196F) CD79B coding-region cDNAs from the MSCV LTR together with a hygromycin resistance gene. After selection in hygromycin, the CD79B shRNA was induced for three to five days before assay.

BTK ASKA assay. BTK was modified by a 'gatekeeper' T474A mutation, as well as a second S538A mutation used previously to construct a BTK ASKA mutant[31]. Wild-type, kinase-dead (K430R) and ASKA forms of BTK were introduced as 3′ GFP fusions without a selectable marker, and then the mixed (GFP-negative and GFP-positive) population was infected with a retrovirus expressing an shRNA targeting the BTK 3′ UTR and selected with puromycin. The starting proportion of GFP-positive cells was determined by FACS with bead quantification, and then shRNA expression was induced with doxycycline in the presence or absence of the ASKA inhibitor 1NM-PP1 (2 μM; Sigma) for continued culture and periodic quantification. The increase in GFP-positive cells in the various cultures was normalized to the increase observed for wild-type BTK without 1NM-PP1.

FACS assays for protein level. For quantification of most surface markers, FACS was performed by staining unfixed cells on ice. For quantification of CD79A, cells were fixed in 1.6% paraformaldehyde, permeabilized in methanol and stained with an antibody recognizing a peptide sequence in the intracellular domain (BD Biosciences). For analysis of total and surface-expressed levels of exogenous mutant CD79A or CD79B in the BJAB model system, and also their effect on surface IgM (Fig. 4c), BJAB cells were first prepared to express an shRNA targeting either endogenous gene, using selected single clones in the case of CD79B, and then infected with bicistronic expression vectors containing 5′ Flag-tagged and 3′ GFP-tagged versions (described above) of the respective genes and an IRES-Lyt2 (murine CD8a) cassette. After induced knockdown of the endogenous proteins, three-colour FACS assays were performed in which GFP fluorescence was used as a measure of total exogenous protein, an Alexa-647 antibody against Lyt2 (BD Biosciences, excited with a He/Ne laser) was used as a marker of exogenous mRNA, and a phycoerythrin-conjugated antibody was used to detect surface expression of IgM (direct) or Flag (by secondary detection of unlabelled mouse anti-Flag (M2; Sigma).

Western blotting. Control or doxycycline-treated cells were lysed for 30 min in lysis buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1% Triton X-100, 1% Nonidet P40, 2 mM EDTA) supplemented with a cocktail of protease inhibitors (Roche) and phosphatase inhibitors (Sigma). Lysates were cleared by centrifugation at 15,000*g* at 4 °C for 10 min, and protein concentrations were determined by bicinchoninic acid protein assay (Pierce). Lysates (80–100 μg) were subjected to electrophoresis through a 4–12% Bis-Tris gel (Invitrogen) and immobilized on the nitrocellulose membranes. Proteins were detected by using the following antibodies: CD79A, Syk, IκB-α (Santa Cruz Biotechnology), BTK (ref. 32; provided by D. Stewart), phosphotyrosine (4G10; Millipore), Akt, ERK, p-Akt (S473), p-ERK (p44-T202, p42-Y204), β-actin, p-Lyn (Y507) and p-IκB-α (S32) (Cell Signaling Technology).

Lyn immunoprecipitation and *in vitro* kinase assay. Cells were suspended at 10^7 cells ml^{-1} in PBS and lysed by mixing 1:1 (v/v) with RIPA buffer (0.5% Triton X-100, 0.5% deoxycholate, 0.05% SDS) in lysis buffer. Cells were lysed for 10 min on ice and then clarified by microcentrifugation (12,000*g*) for 20 min at 4 °C to yield a postnuclear supernatant. Lyn was immunoprecipitated from 0.5–1 ml of detergent extracts by incubation with 2 μg of anti-Lyn mouse monoclonal antibody H6 (Santa Cruz Biotechnology) for 1–2 h on ice and then rotated for 45 min with 35 μl of ImmunoPure Immobilized Protein A (Pierce) at 4 °C. Immunoprecipitates were washed twice with lysis buffer without detergent and then once with kinase assay buffer (20 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 1 mM Na_3VO_4). After being washed, Lyn immunoprecipitates were subjected to *in vitro* kinase assays by the addition of 200 μl of kinase assay buffer containing either 1 mM ATP or no ATP. Samples were then incubated at 37 °C for 15 min.

Reactions were quenched by the addition of 50 μl of 5 × non-reducing SDS sample buffer followed immediately by boiling.

MTT assays. Cells in 96-well plates were treated with dasatinib, rapamycin and/or IKK-β inhibitor for four or eight days with replacement of fresh medium supplemented with 1 × concentration of drugs every two days. At harvesting, cells were treated with MTT (Sigma) for 0.5–1 h and the coloured substrate produced was solubilized in propan-2-ol containing 1% hydrochloric acid. The absorbance of the coloured supernatant at a wavelength of 570 nm was measured by a spectrophotometer with subtraction of background absorbance at 630 nm.

TIRF microscopy. TIRF imaging of the BCR was based on described previously techniques[12]. In brief, surface IgM was labelled by staining with anti-IgM-Cy5 Fab (Jackson Research) for 15 min on ice in Hank's balanced salt solution containing 1% calf serum. Cells were then incubated with 1 mM R18 dye (Invitrogen) at room temperature (21–23 °C) for 1 min, washed and allowed to adhere to coverslips coated with planar lipid bilayers consisting of 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (Avanti Polar Lipids). Cells were imaged live at 37 °C, or alternatively were fixed with 2% paraformaldehyde, permeabilized with 0.1% Triton X-100 and stained with anti-phosphotyrosine antibodies (4G10–FITC; Upstate). Co-localization of the BCR and phosphotyrosine signal was quantified by Pearson correlation coefficient with the use of ImageJ (National Institutes of Health). For measurements of single-molecule diffusion of the BCR, surface IgM was labelled with 2 ng ml^{-1} anti-IgM-Cy3 Fab for 10 min at room temperature, and tracks of single BCR molecules were recorded at 35 ms resolution. Images were bandpass-filtered and the position of the BCR in each point of the track was determined by two-dimensional Gaussian fitting with the use of Matlab (Mathworks) scripts. Short-range diffusion coefficients from individual BCR tracks were calculated as described[12].

IKK reporter assay. As described previously for OCI-Ly3 (ref. 4), stable clones of TMD8 were constructed with separate vectors to express a fusion between IκB-α and *Photinus* luciferase (from pGL3; Promega) as the reporter, and *Renilla* luciferase (from phRL-TK; Promega) for normalization. Clones responsive to an IKK-β small-molecule inhibitor were identified and used with the OCI-Ly3 reporter clone for short-term (4 h) incubation with dasatinib (Supplementary Fig. 12) or, after infection with pRSMX_Puro and selection, shRNA induction for one to four days (Supplementary Fig. 4c). After development with the Dual-Glo luciferase assay system (Promega), the ratio of IκB-α–*Photinus* to *Renilla* luminescence was normalized to that in untreated or uninduced cultures.

Patient samples. Tumour biopsy specimens were obtained before treatment from 223 patients with *de novo* DLBCL that had previously been analysed by gene expression profiling[28], 16 patients with MALT lymphoma, and 20 patients with Burkitt's lymphoma. All samples were studied in accordance with a protocol approved by the National Cancer Institute Institutional Review Board.

PCR amplification and sequencing. Genomic DNA from cell lines and patient samples was extracted with the DNeasy Tissue kit (Qiagen) in accordance with the manufacturer's instructions. Long-distance PCR for CD79A and CD79B was performed with the LA PCR kit (TaKaRa Bio Inc.) using the following conditions: 94 °C for 5 min followed by 30 cycles of denaturation (30 s at 94 °C), annealing (30 s at 60 °C) and extension (6–7 min at 72 °C), and then final extension for 10 min at 72 °C. PCR primers used were as follows: for CD79A, CD79A_1_f (5′-TCCACTCACAGCCTGAAGCATAC-3′) and CD79A_1_r (5′- GGTTAGGAGGTGGGGCAGTTTAG-3′); for CD79B, CD79B_1_f (5′-GGTGCAGTTACACGTTTTCCTCC-3′) and CD79B_1_r (5′-TGGTTGCGGGAGAGGAATGATG-3′).

The PCR products were revealed by electrophoresis on a 0.8% agarose gel and ethidium bromide staining. The templates were purified with the QIAquick PCR Purification Kit (Qiagen) and subsequently sequenced (BigDye sequencing system; Applied Biosystems). Mutations were confirmed on independent PCR products.

Reverse transcriptase PCR and TA cloning. Total RNA (1 mg) from ABC DLBCL cell lines was transcribed with the GeneAmp RNA PCR Core Kit (Applied Biosystems) in accordance with the manufacturer's instructions. cDNA was amplified using the following conditions: 94 °C for 10 min followed by 40 cycles of denaturation (30 s at 94 °C), annealing (30 s at 58 °C) and extension (1 min at 72 °C), and then final extension for 10 min at 72 °C. The templates were purified with the QIAquick Gel Extraction Kit (Qiagen) and subsequently TA-cloned with the TOPO TA Cloning Kit (Invitrogen) in accordance with the manufacturer's instructions. Between 12 and 28 clones were picked, bacterial cultures were grown, and plasmid DNA was isolated and subsequently sequenced.

PCR primers. PCR primers used were as follows: for CD79a, CD79a_2_f (5′-GCAACTCAAACTAACCAACCCACTG-3′) and CD79a_2_r (5′-CACTAACGAGGCTGCTACAATCAG-3′); for CD79b, CD79b_2_f (5′-ATGGGATTCAGCACCTTGGC-3′) and CD79b_2-r (5′-CCTCATAGGTGGCTGTCTGGTC-3′).

Gene-expression profiling. Total RNA (Trizol reagent; Invitrogen) was prepared from HBL-1 cells after incubation with 25 μM MLN120B (Millennium Pharmaceuticals) for 2, 3, 4, 6, 8, 12, 16 and 24 h. In addition, HBL-1 cells were infected with retroviral vectors expressing various shRNAs in a doxycycline-inducible fashion (Supplementary Fig. 4a, b and Supplementary Table 2), selected with puromycin, treated with doxycycline for 24 or 48 h and then collected for total RNA. Uninduced cultures were prepared in parallel.

Gene expression was measured with whole-genome Agilent 4 × 44K gene expression arrays (Agilent), in accordance with the manufacturer's protocol. Signals from either untreated or uninduced HBL-1 cells (labelled with Cy3) were compared with signals from the respective MLN120B-treated or doxycycline-induced cells (labelled with Cy5). For each sample, 2 mg of total RNA was used for the preparation of fluorescent probes.

A gene was selected as an NF-κB target gene in HBL-1 cells if MLN120B decreased the expression of the gene by more than 0.65log$_2$ (1.57-fold) at four or more time points. This NF-κB target gene signature was subsequently applied to the gene expression data after the induction of shRNAs directed against BTK, CARD11, Syk and CD79A.

30. Tolar, P., Sohn, H. W. & Pierce, S. K. The initiation of antigen-induced B cell antigen receptor signaling viewed in living cells by fluorescence resonance energy transfer. *Nature Immunol.* 6, 1168–1176 (2005).
31. Jiang, S. *et al.* Chemical genetic transcriptional fingerprinting for selectivity profiling of kinase inhibitors. *Assay Drug Dev. Technol.* 5, 49–64 (2007).
32. Stewart, D. M., Kurman, C. C. & Nelson, D. L. Production of monoclonal antibodies to Bruton's tyrosine kinase. *Hybridoma* 14, 243–246 (1995).

Vol 463 | 7 January 2010 | doi:10.1038/nature08644

Identification of sister chromatids by DNA template strand sequences

Ester Falconer[1], Elizabeth A. Chavez[1], Alexander Henderson[1], Steven S. S. Poon[1,2], Steven McKinney[2], Lindsay Brown[3], David G. Huntsman[3] & Peter M. Lansdorp[1,4]

It is generally assumed that sister chromatids are genetically and functionally identical and that segregation to daughter cells is a random process. However, functional differences between sister chromatids regulate daughter cell fate in yeast[1] and sister chromatid segregation is not random in *Escherichia coli*[2]. Differentiated sister chromatids, coupled with non-random segregation, have been proposed to regulate cell fate during the development of multicellular organisms[3]. This hypothesis has not been tested because molecular features to reliably distinguish between sister chromatids are not obvious. Here we show that parental 'Watson' and 'Crick' DNA template strands can be identified in sister chromatids of murine metaphase chromosomes using CO-FISH (chromosome orientation fluorescence *in situ* hybridization[4]) with unidirectional probes specific for centromeric and telomeric repeats. All chromosomes were found to have a uniform orientation with the 5′ end of the short arm on the same strand as T-rich major satellite repeats. The invariable orientation of repetitive DNA was used to differentially label sister chromatids and directly study mitotic segregation patterns in different cell types. Whereas sister chromatids appeared to be randomly distributed between daughter cells in cultured lung fibroblasts and embryonic stem cells, significant non-random sister chromatid segregation was observed in a subset of colon crypt epithelial cells, including cells outside positions reported for colon stem cells[5]. Our results establish that DNA template sequences can be used to distinguish sister chromatids and follow their mitotic segregation *in vivo*.

Major satellite repeats have a uniform head-to-tail orientation on mouse chromosomes relative to the centromere[6,7]. To determine whether this polarity is fixed relative to chromosome ends, we hybridized unidirectional probes specific for major satellite and telomere repeats to single-stranded metaphase chromosomes using CO-FISH (Fig. 1a). For the CO-FISH procedure, cells are treated with BrdU for one round of DNA replication resulting in BrdU incorporation exclusively into the newly formed DNA[4,8]. After treatment with Hoechst 33258 (a DNA dye) and ultraviolet irradiation, nicks are created exclusively at sites of BrdU incorporation, which are then used to remove newly formed DNA by exonuclease treatment and DNA denaturation. The resulting single-stranded chromosomes (containing template DNA only) are hybridized with strand-specific probes (Fig. 1a).

Notably, all chromosomes except the Y chromosome showed a uniform orientation of major satellite relative to telomeric repeats (Fig. 1b). On each chromosome, the 5′ end of the short arm (characterized by C-rich telomere repeats) is adjacent to T-rich major satellite repeat sequences, and the 3′ end of the short arm (characterized by G-rich telomere repeats) is adjacent to A-rich major satellite repeat sequences. All template strands (except those in chromosomes 4 and 18)[9] show mutually exclusive staining with fluorescently labelled peptide-nucleic acid (PNA) probes specific for either A-rich or T-rich major satellite DNA (Fig. 1b, c and Supplementary Figs 1 and 9). Because the orientation of major satellite DNA relative to telomeric DNA is fixed, probes hybridized to major satellite repeats were used to arbitrarily define Watson (red fluorescence, Fig. 1d) and Crick (green fluorescence, Fig. 1d) DNA template strands. A similar chromosomal polarity was observed in *Mus spretus* fibroblasts, with the 5′ end of the short arm adjacent to T-rich minor satellite repeats in most chromosomes (Supplementary Fig. 2). As CO-FISH can differentially label sister chromatids, we adapted the CO-FISH technique to allow us to directly follow chromatid segregation *in vivo* (Fig. 1e).

Non-random segregation of DNA strands in mammalian cells was first reported using indirect pulse-chase experiments with nucleotide analogues in dividing murine intestinal crypt epithelial cells[10]. To study directly the pattern of sister chromatid segregation in such cells, we injected adult mice for 12 h at 1-h intervals with BrdU before the collection of colon tissue, which was fixed, sectioned and subjected to CO-FISH with major satellite probes. Only a minority of cells in colon crypts were actively dividing, as shown by BrdU incorporation (Fig. 2a, right and inset). These BrdU-positive cells showed discrete, non-overlapping red and green fluorescent signals (herein referred to as CO-FISH signals) from the strand-specific probes (Fig. 2b, white arrowheads) indicating successful generation of single-stranded chromosomes. In contrast, most non-mitotic cells showed overlapping red and green fluorescence from the major satellite probes hybridizing to both strands of double-stranded chromosomes (Fig. 2b, yellow arrowhead, Fig. 2c). Cell pairs showing apparent template strand asymmetry were found at different positions within the colon crypt, including high within the crypt axis (Fig. 2c, d and Supplementary Fig. 3). Sister nuclei showing reciprocal, asymmetric CO-FISH fluorescence are compatible with non-random distribution of sister chromatids containing either Watson or Crick DNA template strands (Figs 1e, 2e, Supplementary Fig. 4 and Supplementary Movie 1). We confirmed that CO-FISH signals in mitotic colon cells from mice subjected to 12 h of BrdU treatment were exclusively derived from cells after only one round of DNA replication (Supplementary Fig. 5)[11]. Of note, DNA template strand asymmetry was also observed in colon tissue sections of *M. spretus* using probes specific for minor satellite repeats (Supplementary Fig. 6).

Chromosomes aligned at the metaphase plate *in vivo* displayed what appeared to be a polar arrangement of Watson and Crick sister chromatids (Fig. 2f and Supplementary Movie 2). Furthermore, major satellite DNA template strands appeared to be clustered after mitosis (Fig. 2g), and often had a marked 'mirror-image' asymmetry with territories of red and green fluorescence in one daughter cell

[1]Terry Fox Laboratory, [2]Molecular Oncology and Breast Cancer Program, B.C. Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada. [3]Centre for Translational and Applied Genomics, B.C. Cancer Agency, Vancouver, British Columbia V6H 3Z6, Canada. [4]Division of Hematology, Department of Medicine, University of British Columbia, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada.

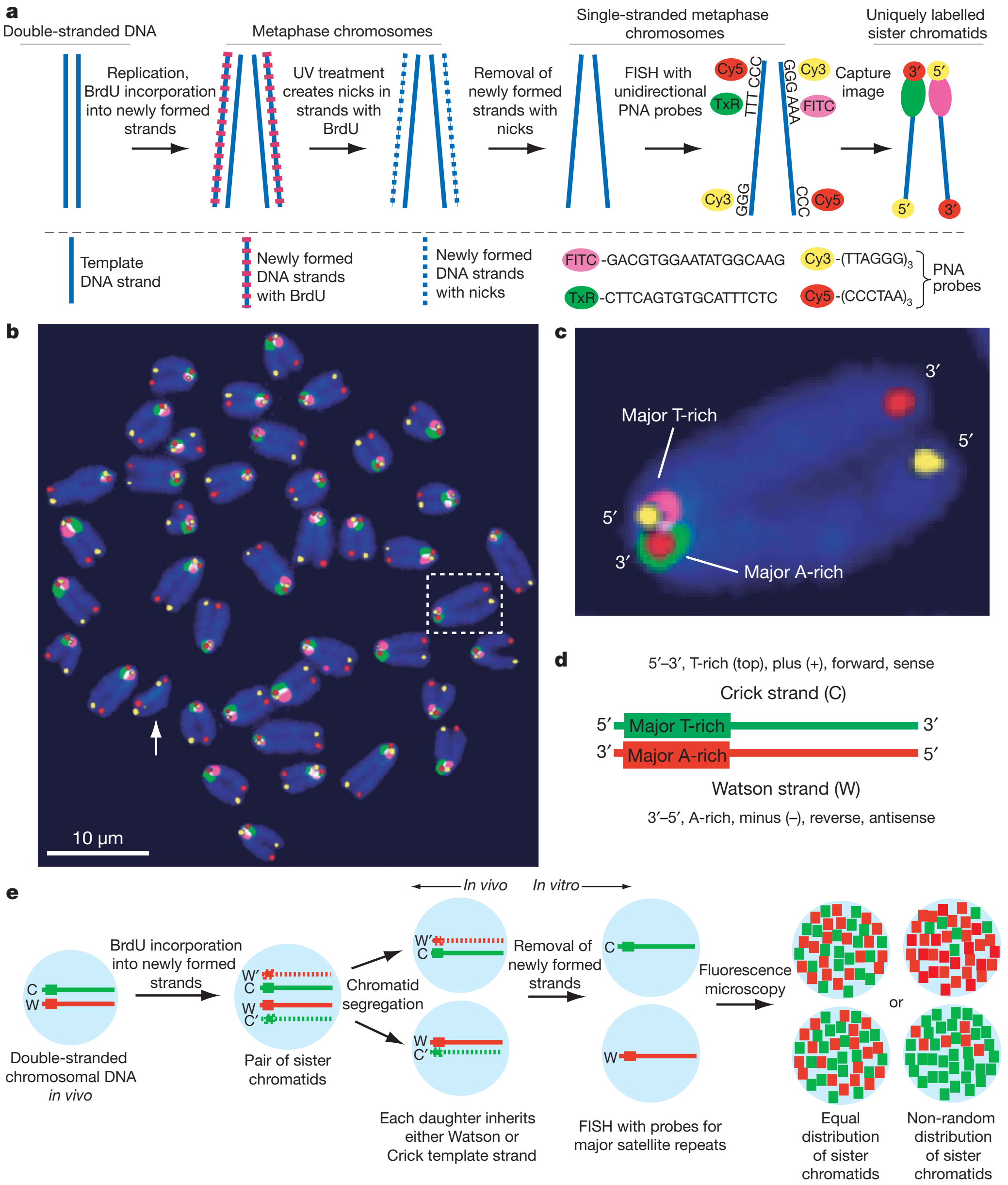

a
Double-stranded DNA
Metaphase chromosomes
Single-stranded metaphase chromosomes
Uniquely labelled sister chromatids
Replication, BrdU incorporation into newly formed strands
UV treatment creates nicks in strands with BrdU
Removal of newly formed strands with nicks
FISH with unidirectional PNA probes
Cy5
TxR
TTT CCC
GGG AAA
Cy3
FITC
Capture image
3′ 5′
Cy3
GGG
CCC
Cy5
5′ 3′
Template DNA strand
Newly formed DNA strands with BrdU
Newly formed DNA strands with nicks
FITC-GACGTGGAATATGGCAAG
TxR-CTTCAGTGTGCATTTCTC
Cy3-(TTAGGG)$_3$
Cy5-(CCCTAA)$_3$
PNA probes
b
10 μm
c
Major T-rich
3′
5′
5′
3′
Major A-rich
d
5′–3′, T-rich (top), plus (+), forward, sense
Crick strand (C)
5′ Major T-rich 3′
3′ Major A-rich 5′
Watson strand (W)
3′–5′, A-rich, minus (–), reverse, antisense
e
In vivo
In vitro
C
W
Double-stranded chromosomal DNA in vivo
BrdU incorporation into newly formed strands
W′
C
W
C′
Pair of sister chromatids
Chromatid segregation
W′
C
W
C′
Each daughter inherits either Watson or Crick template strand
Removal of newly formed strands
C
W
FISH with probes for major satellite repeats
Fluorescence microscopy
or
Equal distribution of sister chromatids
Non-random distribution of sister chromatids

Figure 1 | Highly conserved orientation of telomeric and major satellite DNA in murine chromosomes revealed by four-colour CO-FISH. a, Schematic diagram of the CO-FISH procedure. UV, ultraviolet. **b**, Pseudo-colour CO-FISH image of murine metaphase chromosomes. Note that major satellite repeats on all chromosomes except the Y chromosome (arrow, no major satellite DNA) have the same orientation. **c**, Magnification of the boxed chromosome shown in **b**. **d**, Definition of Watson and Crick DNA template strands based on the uniform orientation of major satellite DNA. **e**, The relative distribution of Watson and Crick major satellite fluorescence can be used to study sister chromatid segregation patterns *in vivo*.

mirrored by territories of the opposite colour in the other daughter cell (Fig. 2g and Supplementary Movies 3 and 4). These observations indicate that pericentric regions of several chromosomes cluster in at least some post-mitotic colon cells on the basis of parental DNA template strand sequences. To exclude major rearrangements in nuclear architecture by our CO-FISH procedure, we performed three-colour CO-FISH with both major satellite probes and a telomeric probe (Supplementary Fig. 7). Telomeric signals were observed at expected positions adjacent to centric regions (the terminus of the short chromatid arms) and adjacent to the division plane (the terminus of the long chromatid arms) in support of the notion that the CO-FISH procedure does not grossly alter the general morphology and positioning of segregating chromosomes.

Our qualitative observations suggested that sister chromatids of most chromosomes are segregating non-randomly in a subset of dividing colon epithelial cells. To test whether our observations could nevertheless be explained by chance, we quantified the relative Watson and Crick fluorescence in each daughter cell using dedicated software (see Supplementary Fig. 8 and Supplementary Methods for details). The measured fluorescence was converted to a relative fluorescence ratio (Fig. 3a) based on the reasoning that the total fluorescence from both daughters is the outcome of redistributing a fixed number of DNA template strands from a mother cell to the two daughter cells (Fig. 1e). Reciprocal ratios of Watson and Crick fluorescence are in agreement with the expected distribution of chromatids between daughter cells.

We compared the measured CO-FISH fluorescence signals from sectioned colon and preparations of isolated colon cells to two cultured cell types not expected to show non-random segregation patterns: pluripotent embryonic stem (ES) cells and lung fibroblasts (Fig. 3a and Supplementary Data Table). To avoid selection bias for asymmetry, every cell pair with clear non-overlapping CO-FISH signals was analysed (Fig. 3b). Although this impartial acquisition of data ensures that the measured sister chromatid segregation patterns are not influenced by cell selection, the results will include all recently divided cells, which may complicate data analysis if chromatid segregation patterns differ between cell types. Nevertheless, cell pairs from colon section and isolated colon cells showed a broader distribution of Watson and Crick fluorescence, compared to cultured ES cells and lung fibroblasts (Fig. 3b, grey boxes), reflecting a higher frequency of sister chromatid asymmetry. Up to 50% of cell pairs from all cell types showed an excellent reciprocal ratio of measured fluorescence values between daughter cells, with Watson fluorescence distribution ratios mirrored within 5% by complementary Crick fluorescence distribution

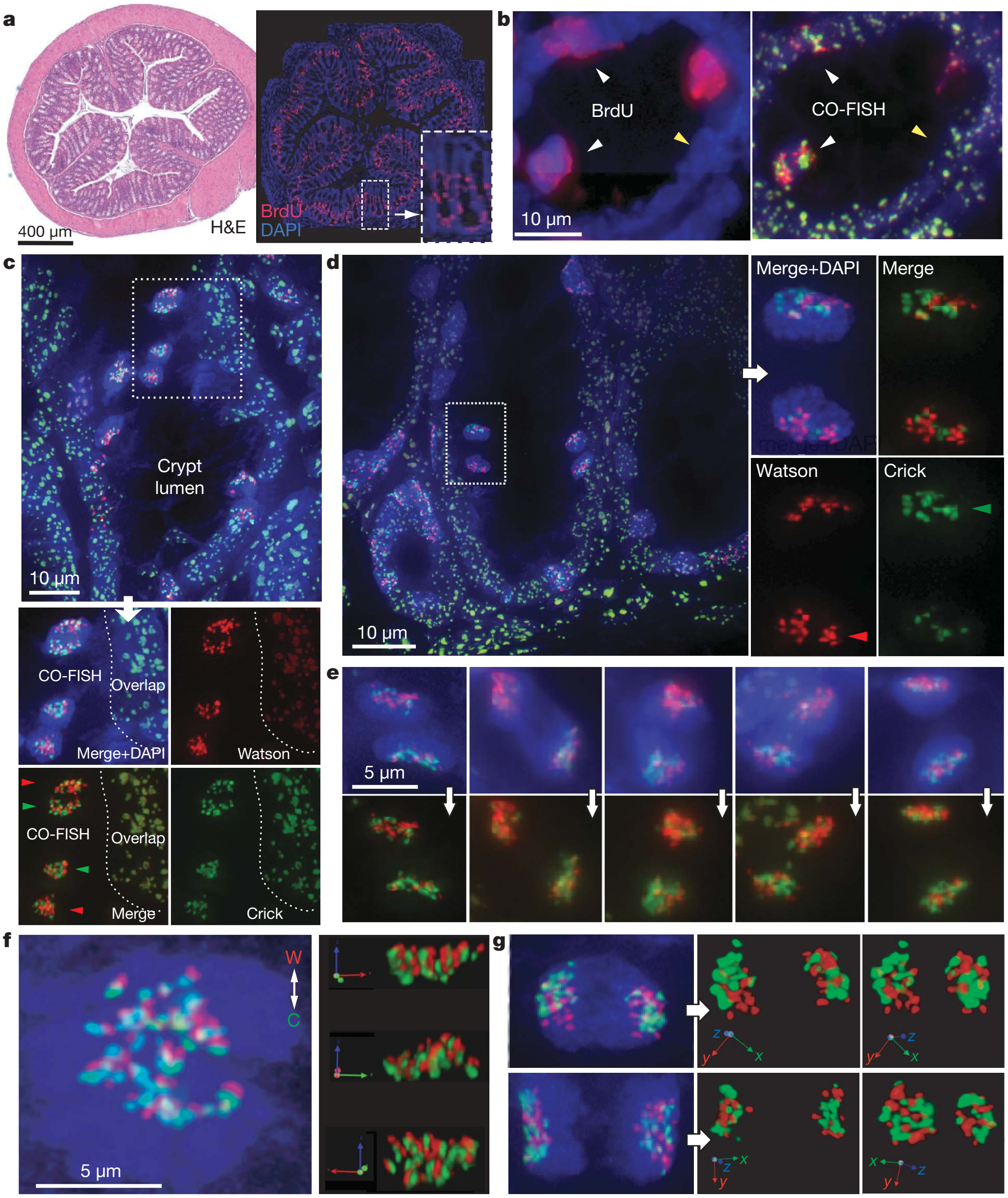


Figure 2 | CO-FISH to study sister chromatid segregation patterns. **a**, Low magnification of adjacent colon sections stained with haematoxylin and eosin (H&E; left), and with DAPI and an anti-BrdU antibody (right). **b**, High magnification of a section stained for BrdU (left) that was subsequently subjected to CO-FISH (right). BrdU-labelled cells show non-overlapping red and green fluorescence (white arrowheads), non-mitotic cells without BrdU show overlapping probe signals (yellow arrowhead). **c**, Example of CO-FISH (non-overlapping) signals in pairs of post-mitotic cells in colon crypts. **d**, Post-mitotic cell pairs relatively high in colon crypt with asymmetric CO-FISH fluorescence. **e**, Examples of asymmetric CO-FISH fluorescence in paired colon cells. **f**, Non-random alignment of sister chromatids at metaphase (right: different projections from Supplementary Movie 2). **g**, Mirror-image symmetry and clustered CO-FISH fluorescence in paired daughter cells (see also Supplementary Movies 3 and 4).

ratios (Fig. 3b, filled squares, and Supplementary Data Table). Cell pairs showing reciprocal fluorescence outside this arbitrary cutoff (Fig. 3b, open circles) most probably reflect noise in CO-FISH measurements due to loss of DNA, non-specific fluorescence and other causes.

To test whether the measured asymmetry in colon cells was non-random, we superimposed our observed fluorescence distributions of cell pairs within the arbitrary 5% reciprocal cutoff value to 95% and 99% confidence intervals calculated from simulated random segregations, representing the range of fluorescence values expected by chance (see Supplementary Figs 9–11 and Supplementary Materials for full discussion of simulated random segregation and statistical analysis). The distribution of Crick template strand fluorescence from sectioned colon tissue and isolated colon cells was outside the 95% or 99% confidence intervals calculated for random sister chromatid segregation (Fig. 3c; $P < 0.05$ open arrowheads, $P < 0.01$ solid arrowheads). This includes a higher frequency of cell pairs with extreme asymmetry, as well as a lower frequency of cell pairs with a symmetrical distribution, than predicted by simulated random segregation. Although fewer cell pairs with extreme asymmetry were present or preserved in colon cell suspensions, the results were nevertheless significant ($P < 0.01$). In contrast, in ES cells and lung fibroblasts the measured fluorescence intensity values were within the 95% and 99% confidence intervals calculated for random segregation. The one exception was in lung fibroblasts at the symmetrical 55% fluorescence value, suggesting a skewing of segregation towards a 50:50 distribution of chromatids (Fig. 3c, bottom, arrowhead). These results support the conclusion that the observed asymmetry of DNA template strand fluorescence in paired colon cells results from non-random segregation of sister chromatids rather than from rare random segregation events. We consider it unlikely that this conclusion is flawed by errors in our methods or fluorescence measurements. The inevitable measurement noise from various sources is not expected to affect adjacent daughter cells in opposite ways (skewing for red fluorescence in one daughter and for green fluorescence in the other). On the other hand, we cannot exclude that BrdU incorporation itself somehow affected sister chromatid segregation and further studies are needed to confirm our findings. Of note, we did not observe 100% asymmetric segregation of sister chromatids in any pair of mitotic colon cells. Most likely, a subset of colon cells selectively segregates sister chromatids from most but not all chromosomes. Alternatively, a small number of specific chromatids could be selectively captured in a larger proportion of cells. The strand-specific probes in this study are unable to detect minor deviations from random sister chromatid segregation or detect selective segregation of a few or single chromosomes[12].

Our results provide the first direct data supporting non-random segregation of DNA template strands in mammalian cells *in vivo*. Non-random segregation of sister chromatids has previously been observed in *E. coli*[2] and has been suggested from indirect measurements in various eukaryotic cells[13,14]. Neither the mechanism nor the function of selective sister chromatid segregation is known at present. To enable non-random segregation, sister chromatid centromeres as well as the two centrosomes of the mitotic spindle must have distinct marks or properties that enable specific connections (Fig. 4). Asymmetry at centrosomes[15,16] could result in differences in the timing, the number or the dynamic behaviour of microtubules radiating from each pole. Alternatively, such differences could result from proteins enriched at a specific pole (Fig. 4a). The adenomatous polyposis coli (APC) tumour suppressor protein could be an example of the latter given its involvement in several cellular processes including chromosome segregation and spindle assembly[17–21]. How sister chromatid centromeres are distinguished is equally enigmatic, but probably depends on differences in (peri)centric chromatin, perhaps by differences in the loading[22] or retention[23] of (peri)centric proteins or strand-specific replication[24], methylation[25] or transcription of centromeric DNA[26,27]. Centromeric RNA is known to regulate the assembly of centromeres[28] and strands of major satellite DNA are differentially transcribed during murine development[29]. Chromatin differences between sister chromatids could either be directly recognized by factors at asymmetric spindles (Fig. 4a) or favour selective

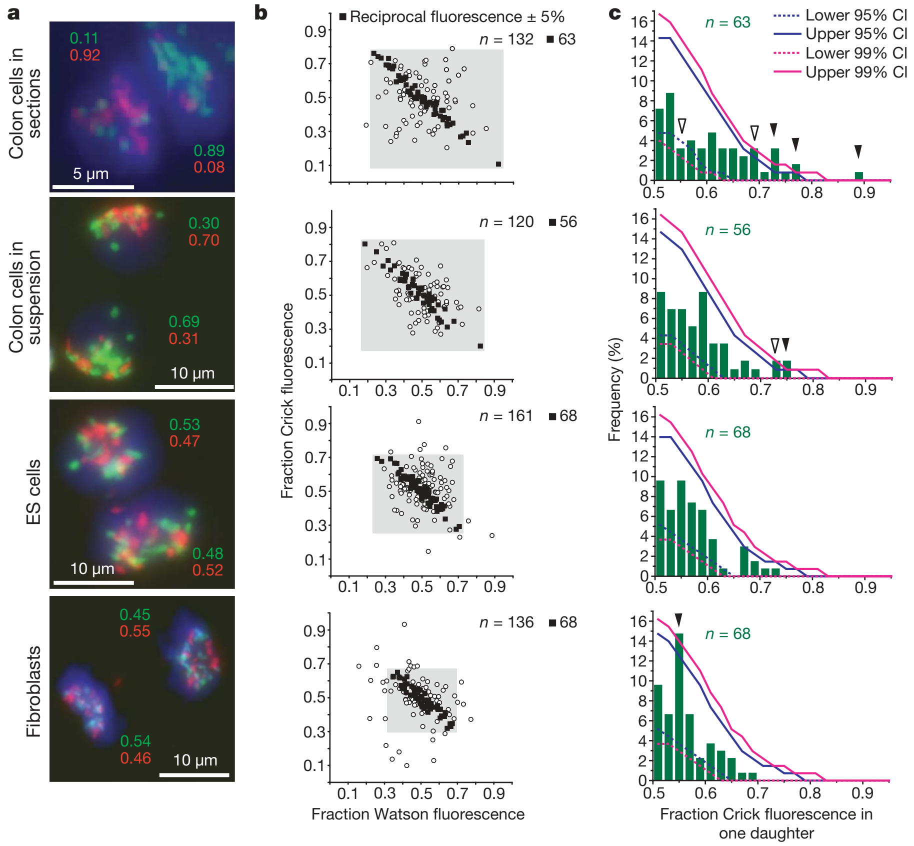


Figure 3 | Measurements of Watson and Crick DNA template strand fluorescence in post-mitotic cells. a, Examples of fluorescence measured in the indicated cell types. **b**, For *n* cell pairs, the ratio of Watson and Crick fluorescence in one of the daughter cells (arbitrary selection) is plotted. Solid black squares show cells with reciprocal Watson and Crick fluorescence ratios ± 5%, whereas open circles represent cells with Watson/Crick fluorescence ratios outside this arbitrary cutoff. **c**, The observed Crick fluorescence distributions in selected individual cells (*n* = black squares in **b**) were compared to fluorescence distribution values obtained by simulated random segregation. The observed frequency (*y* axis) of Crick fluorescence (*x* axis, green histograms) in one daughter cell (with the brightest Crick fluorescence) is plotted. The upper and lower 95% and 99% confidence intervals (CI, solid and dashed blue and magenta lines, respectively) represent the range of fluorescence distributions expected by chance. The values measured in colon tissue sections as well as colon cell suspensions fall outside the range for simulated random segregation ($P < 0.05$: open arrowheads; $P < 0.01$: solid arrowheads).

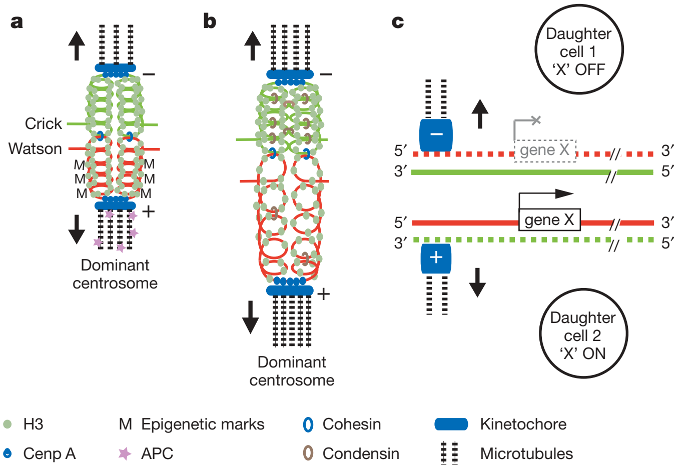


Figure 4 | Models for the mechanism and function of asymmetric sister chromatid segregation. Only the template strand of double-stranded DNA in sister chromatids is shown. **a**, Uneven distribution of epigenetic marks (M) between sister chromatid centromeres could result in asymmetric nucleation of microtubules or selective capture of microtubules coming from the 'dominant' centrosome[15,16]. **b**, Differences in higher-order chromatin structure could alter the elastic properties of (peri)centric chromatin[30] and select specific sister chromatids by microtubules originating from the dominant centrosome. **c**, Regulation of cell fate by selective segregation of sister chromatids that differ in epigenetic marks at centromeres and selected genes.

attachment to microtubules by changes in elastic properties[30] (Fig. 4b). We propose that the observed non-random segregation of sister chromatids contributes to cell fate decisions as predicted by the 'silent sister' hypothesis[3] (Fig. 4c). Further studies will test the predictions of this hypothesis that chromatin differences between sister chromatids contribute to differences in gene expression between cells, and thus regulate cell fate in asymmetrically dividing cells.

METHODS SUMMARY

CO-FISH analysis. Metaphase chromosomes were prepared from mouse ES cells (C57BL/6J background) incubated with 40 μm BrdU for 12 h before collection using standard cytogenetic procedures. Cytospin preparations of binucleated cells ($3\,\mu g\,ml^{-1}$ of cytochalasin B for 2 h before collection) were prepared from cultured ES cells and cultured adult lung fibroblasts. For CO-FISH, BrdU-treated cells or chromosomes were treated with pepsin, RNase A, Hoechst 33258 and irradiated with ultraviolet light[8]. Nicked DNA was removed by denaturation after digestion with exonuclease III, and remaining DNA (template) strands were hybridized to directly labelled fluorescent PNA probes specific for C- and G-rich telomere repeats and T- and A-rich major satellite DNA. Slides were counterstained with 4′,6-diamidino-2-phenylindole (DAPI). For *in vivo* studies, C57BL/6J mice (2–3 months old) received intraperitoneal injections of BrdU at 12 × 1-h intervals. Colon tissue was formaldehyde-fixed and paraffin-embedded using standard procedures. Deparaffinized tissue sections (6 μm) were used for CO-FISH as described earlier, with further treatment at pH 6.0 and 80 °C for 45 min. Suspensions of viable colon cells were prepared with collagenase, then dropped onto slides after hypotonic treatment and fixation with methanol/acetic acid to control for possible artefacts from tissue sectioning.

Image analysis. Fluorescence images captured using a digital camera mounted on a fluorescence microscope were combined and processed to provide pseudo-colour images using Adobe Photoshop software. For tissue sections, a stack of images at 0.2–0.5 μm intervals was acquired and projected into a single plane. In some cases deconvolution software (SoftWoRx, Applied Precision) was used to create projection images. For image and data analysis, the fluorescence intensities within individual nuclei of paired daughter cells were measured using in-house software, and a combination of Bootstrap inference and Monte Carlo simulations was applied to build a random-segregation model and calculate 95% and 99% confidence intervals of expected fluorescence distribution profiles.

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

Received 16 June; accepted 6 November 2009.
Published online 16 December 2009.

1. Klar, A. J. Differentiated parental DNA strands confer developmental asymmetry on daughter cells in fission yeast. *Nature* **326**, 466–470 (1987).
2. White, M. A., Eykelenboom, J. K., Lopez-Vernaza, M. A., Wilson, E. & Leach, D. R. Non-random segregation of sister chromosomes in *Escherichia coli. Nature* **455**, 1248–1250 (2008).
3. Lansdorp, P. M. Immortal strands? Give me a break. *Cell* **129**, 1244–1247 (2007).
4. Meyne, J. & Goodwin, E. H. Strand-specific fluorescence *in situ* hybridization for determining orientation and direction of DNA sequences. *Methods Mol. Biol.* **33**, 141–145 (1994).
5. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449**, 1003–1007 (2007).
6. Garagna, S. *et al.* Pericentromeric organization at the fusion point of mouse Robertsonian translocation chromosomes. *Proc. Natl Acad. Sci. USA* **98**, 171–175 (2001).
7. Lin, M. S. & Davidson, R. L. Centric fusion, satellite DNA, and DNA polarity in mouse chromosomes. *Science* **185**, 1179–1181 (1974).
8. Bailey, S. M., Goodwin, E. H. & Cornforth, M. N. Strand-specific fluorescence *in situ* hybridization: the CO-FISH family. *Cytogenet. Genome Res.* **107**, 14–17 (2004).
9. Alves, P. & Jonasson, J. New staining method for the detection of sister-chromatid exchanges in BrdU-labelled chromosomes. *J. Cell Sci.* **32**, 185–195 (1978).
10. Potten, C. S., Hume, W. J., Reid, P. & Cairns, J. The segregation of DNA in epithelial stem cells. *Cell* **15**, 899–906 (1978).
11. Schneider, E. L., Sternberg, H. & Tice, R. R. *In vivo* analysis of cellular replication. *Proc. Natl Acad. Sci. USA* **74**, 2041–2044 (1977).
12. Armakolas, A. & Klar, A. J. Cell type regulates selective segregation of mouse chromosome 7 DNA strands in mitosis. *Science* **311**, 1146–1149 (2006).
13. Bell, C. D. Is mitotic chromatid segregation random? *Histol. Histopathol.* **20**, 1313–1320 (2005).
14. Karpowicz, P. *et al.* The germline stem cells of *Drosophila melanogaster* partition DNA non-randomly. *Eur. J. Cell Biol.* **88**, 397–408 (2009).
15. Wang, X. *et al.* Asymmetric centrosome inheritance maintains neural progenitors in the neocortex. *Nature* **461**, 947–955 (2009).
16. Yamashita, Y. M., Mahowald, A. P., Perlin, J. R. & Fuller, M. T. Asymmetric inheritance of mother versus daughter centrosome in stem cell division. *Science* **315**, 518–521 (2007).
17. Etienne-Manneville, S. & Hall, A. Cdc42 regulates GSK-3β and adenomatous polyposis coli to control cell polarity. *Nature* **421**, 753–756 (2003).
18. Hanson, C. A. & Miller, J. R. Non-traditional roles for the Adenomatous Polyposis Coli (APC) tumor suppressor protein. *Gene* **361**, 1–12 (2005).
19. Kaplan, K. B. *et al.* A role for the Adenomatous Polyposis Coli protein in chromosome segregation. *Nature Cell Biol.* **3**, 429–432 (2001).
20. Kita, K., Wittmann, T., Nathke, I. S. & Waterman-Storer, C. M. Adenomatous polyposis coli on microtubule plus ends in cell extensions can promote microtubule net growth with or without EB1. *Mol. Biol. Cell* **17**, 2331–2345 (2006).
21. Yamashita, Y. M., Jones, D. L. & Fuller, M. T. Orientation of asymmetric stem cell division by the APC tumor suppressor and centrosome. *Science* **301**, 1547–1550 (2003).
22. Jansen, L. E., Black, B. E., Foltz, D. R. & Cleveland, D. W. Propagation of centromeric chromatin requires exit from mitosis. *J. Cell Biol.* **176**, 795–805 (2007).
23. Thorpe, P. H., Bruno, J. & Rothstein, R. Kinetochore asymmetry defines a single yeast lineage. *Proc. Natl Acad. Sci. USA* **106**, 6673–6678 (2009).
24. Lew, D. J., Burke, D. J. & Dutta, A. The immortal strand hypothesis: how could it work? *Cell* **133**, 21–23 (2008).
25. Luo, S. & Preuss, D. Strand-biased DNA methylation associated with centromeric regions in *Arabidopsis. Proc. Natl Acad. Sci. USA* **100**, 11133–11138 (2003).
26. Kanellopoulou, C. *et al.* Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19**, 489–501 (2005).
27. Murchison, E. P., Partridge, J. F., Tam, O. H., Cheloufi, S. & Hannon, G. J. Characterization of Dicer-deficient murine embryonic stem cells. *Proc. Natl Acad. Sci. USA* **102**, 12135–12140 (2005).
28. Bouzinba-Segard, H., Guais, A. & Francastel, C. Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proc. Natl Acad. Sci. USA* **103**, 8709–8714 (2006).
29. Rudert, F., Bronner, S., Garnier, J. M. & Dolle, P. Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. *Mamm. Genome* **6**, 76–83 (1995).
30. Bouck, D. C. & Bloom, K. Pericentric chromatin is an elastic component of the mitotic spindle. *Curr. Biol.* **17**, 741–748 (2007).

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

Acknowledgements K. Lisaingo helped acquire images for this study, J. Tan helped with image analysis and C. J. S. Smith assisted with *in vivo* studies. We thank S. Selig, T. de Lange and R. C. Wang for help with CO-FISH protocols. S. Aparicio, L. Lefebvre, L. F. Lansdorp, B. M. Lansdorp and members of the Lansdorp laboratory are thanked for discussions and for critically reading the manuscript. This study was funded in part by grants from the Canadian Institutes of Health Research (RMF-92093), the Michael Smith Foundation for Health Research, the Canadian Cancer Society Research Institute and the Terry Fox Foundation. We are grateful to the British Columbia Cancer Agency, the Canada Foundation for Innovation, the British Columbia Knowledge Development Fund, the British Columbia Cancer Foundation, the Blossom Fund of the University of British Columbia and the Mahon family for funding the infrastructure that enabled this work.

Author Contributions E.F. helped with the design of the experiments, image acquisition, data analysis, interpretation of results and writing of the paper. E.A.C. performed most of the CO-FISH experiments. A.H. performed most of the mouse work. L.B. acquired some images for this study. S.S.S.P. performed analysis of digital data and helped with statistical analysis that was performed by S.M. D.G.H. helped with the design of the study and interpretation of results. P.M.L. conceived the study, helped with image acquisition, interpretation of results and writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.M.L. (plansdor@bccrc.ca).

METHODS

Preparation of cells for CO-FISH analysis. Undifferentiated wild-type murine embryonic stem cells (C2, C57BL/6NTac background) and R1-derived wild-type embryonic stem cells ($C1^{+/+}$, 129S1 background)[31] were obtained from A. Nagy (Samuel Lunenfeld Research Institute, Toronto) and cultured on gelatin-coated plastic culture dishes in DMEM containing 20% FCS in the presence of 100 ng ml^{-1} of leukaemia inhibitory factor as described[32]. Murine embryonic fibroblasts were grown in DMEM-FCS. BrdU (Invitrogen) was added to semi-confluent cultures at a final concentration of 40 μM for 12 h before collection. Binucleated cells were prepared by adding cytochalasin B (Sigma-Aldrich, 3 μg ml^{-1}) to cultures 2 h before collection of cells by trypsinization. Cells were spun onto microscope slides using a Shandon Cytospin 4 (Shandon Scientific) and fixed with 3:1 methanol/acetic acid. For preparation of metaphase cells colcemid (Sigma-Aldrich, 0.1 μg ml^{-1}) was added for 1 h before collection. Trypsinized cells were treated with 0.075 M KCl (Stem Cell Technologies, Inc.) for 10 min before fixation with 3:1 methanol/acetic acid using standard cytogenetic procedures. Fixed cells were stored at −20 °C. To obtain metaphase spreads for CO-FISH, cells were dropped onto wet microscope slides and dried overnight at room temperature.

CO-FISH. Slides were rehydrated in PBS, pH 7.4 (Stem Cell Technologies) for 15 min and fixed for 2 min in 4% formaldehyde in PBS followed by three washes in PBS for 5 min each. Slides were treated with freshly prepared pepsin (P7000, Sigma-Aldrich) at 1 mg ml^{-1} in acidified water (pH 2.0) at 37 °C for 10 min followed by two washes for 2 min each in PBS, a rinse in 2×SSC and treatment with RNase A (0.1 mg ml^{-1} in PBS) for 10 min at 37 °C. Slides were washed twice with PBS for 5 min each, and stained with 100 μl Hoechst 33258 (Sigma-Aldrich) at 1 mg ml^{-1} under parafilm for 15 min at room temperature. The slides were rinsed with 2×SSC, transferred to a tray, covered with a glass coverslip and irradiated with ultraviolet light for 30 min in a UV Stratalinker 1800 (calculated dose 5.4×10^3 J m^{-2}). BrdU-substituted DNA strands were digested with 50 μl exonuclease III (New England Biolabs) at 3,000 U ml^{-1} in buffer supplied by the manufacturer (50 mM Tris-HCl, 5 mM $MgCl_2$ and 5 mM dithiothreitol (DTT), pH 8.0) at 37 °C for 10 min under a coverslip. Slides were rinsed three times for 5 min each in 2×SSC before denaturation in 70% formamide in 2×SSC for 1 min at 70 °C and dehydration in ice cold 70%, 90% and 100% ethanol for 2 min each. Cells were rehydrated in PBS for 10 min, fixed in 4% formaldehyde in PBS for 2 min, washed three times for 5 min each in PBS, dehydrated in ethanol again and air dried. Hybridization mixture (20 μl) was added to each slide, covered with a coverslip, and cells were denatured on a hot plate at 80 °C for 2 min. The hybridization mixture consisted of 10 mM Tris-HCl, pH 7.6, 1 mM $MgCl_2$, 70% formamide, 0.25% blocking reagent (New England Nuclear), 0.5 μg ml^{-1} Cy5-labelled $(CCCTAA)_3$ PNA and 0.5 μg ml^{-1} Cy3-labelled $(TTAGGG)_3$ PNA (specific for the G- and C-rich telomeres, respectively), 1 μg ml^{-1} fluorescein-labelled GACGTGGAATATGGCAAG PNA specific for the T-rich strand of mouse major satellite DNA[33] and 1 μg ml^{-1} TexasRed-labelled CTTCAGTGTGCATTTCTC PNA specific for the A-rich strand of mouse major satellite DNA. Fluorescently labelled PNA probes were obtained from Applied Biosystems, Panage Inc. or Biosynthesis Inc. without noticeable differences in results. After hybridization for 1 h at room temperature, slides were washed twice for 15 min each in 70% formamide, 10 mM Tris-HCl, 1% BSA, and for 3 × 5 min in TNT (0.1 M Tris-HCl, 0.15 M NaCl, 0.08% Tween-20, pH 7.5). After dehydration in ethanol, slides were air dried and counterstained with DAPI at 200 ng ml^{-1} in DABCO antifade solution[34].

CO-FISH on paraffin-embedded tissue sections. C57BL/6J mice (2–3 months old) were injected intraperitoneally with BrdU for 12 h at 1-h intervals as described[35]. For metaphase analysis, BrdU was injected for 0, 8, 12 or 16 h, with an extra injection of colcemid to arrest cells at metaphase 1 h before tissue collection. Colon tissue was fixed overnight in 4% formaldehyde in PBS, and embedded in paraffin using standard procedures. Tissue sections (6 μm) were baked overnight at 60 °C, deparaffinized in xylene three times for 15 min each at room temperature before dehydration in 100% ethanol (twice for 10 min). The previously described CO-FISH protocol was used for metaphase spreads from cultured cells with the following modifications. Slides were treated in 10 mM citric acid buffer, pH 6.0 at 80 °C for 45 min, washed at room temperature in PBS twice for 5 min each, and water for 5 min followed by pepsin and RNase treatment as described earlier. After incubation with Hoechst and treatment with ultraviolet irradiation as earlier, slides were rinsed with 2×SSC for 5 min and denatured in 70% formamide, 2×SSC for 2 min at 72 °C, dehydrated, air dried and rehydrated in PBS for 10 min before treatment with RNase A and exonuclease III as above. Slides were washed twice for 5 min each in 2×SSC, denatured in 70% formamide, 2×SSC for 1 min at 70 °C, dehydrated in ethanol and air dried. For hybridization, 40 μl of hybridization mixture containing 1 μg ml^{-1} Cy5-labelled GACGTGGAATATGGCAAG or Cy5-labelled GAAGGACCTGGAATATGG PNA specific for the T-rich Crick strand of mouse major satellite DNA[33] and 1 μg ml^{-1} Cy3-labelled CTTGCCATATTCCACGTC specific for the A-rich Watson strand of mouse major satellite DNA was used. After denaturation for 3 min at 80 °C and hybridization overnight at room temperature, slides were washed and counterstained with DAPI at 10 ng ml^{-1} in PBS for 5 min, rinsed three times for 5 min in PBS, dehydrated in ethanol and covered under DABCO antifade solution for fluorescence microscopy.

Isolation of paired cells from colon. C57BL/6J mice (2–3 months old) were injected intraperitoneally with BrdU for 12 h at 1-h intervals and paired cells from colon were isolated by modification of a published method[35]. In brief, colon tissues were dissected from BrdU-treated mice and placed in ice-cold PBS. Faeces were cleared by flushing the colon with a syringe filled with ice-cold PBS. Colons were subsequently cut longitudinally and minced into 1-cm pieces then incubated in 15 ml of predigestion solution (5 mM EDTA, 1 mM DTT, 1× PBS) for 30 min at 37 °C. The resulting cell suspension was centrifuged for 5 min at 350*g* at 20 °C. Supernatant was aspirated and the cell pellet was resuspended in 15 ml of digestion solution (prepared by mixing 50 mg of collagenase type XI (Sigma-Aldrich) and 100 mg of dispase II (Sigma-Aldrich) into 100 ml of PBS) for 90 min at 37 °C. After incubation, the cell suspension was centrifuged for 5 min at 350*g* at 20 °C. Supernatant was aspirated and discarded. Cell pellet was resuspended in PBS, vortexed for 20 s and passed through a 100-μm cell strainer (BD Falcon). Cells were treated with 0.075 M KCl for 10 min at 37 °C before fixation with 3:1 methanol/acetic acid. Fixed cells were stored at −20 °C, then spun onto microscope slides and subject to CO-FISH analysis as above.

Fluorescence microscopy, image acquisition and selection. Fluorescence signals were captured on an Axioplan microscope (Zeiss) equipped with filters for DAPI, FITC, Cy3, Cy5 and Texas Red (Chroma Technology and Semrock) using an Axiocam MRm digital camera controlled by Metasystems ISIS software (Altlussheim). Alternatively, images were acquired on a Coolsnap HQ digital camera attached to an inverted microscope (IX70 Olympus) fitted to an imaging system (DeltaVision RT, Applied Precision) equipped with similar filter sets. Grey-scale (12 bit) images at the wavelengths of interest were acquired through a high-numerical-aperture ×63/1.4 or ×60/1.4 oil immersion lens. For tissue sections, a stack of images at 0.15–0.5 μm intervals were acquired to cover the entire thickness of the section. Fluorescence signals in individual image planes were projected onto a single image plane using ISIS software (Metasystems) or SoftWoRx (Applied Precision) software before or after deconvolution.

Acquisition of image stacks was limited to informative cell pairs defined as cell pairs in which both nuclei appeared to be intact and did not overlap with neighbouring nuclei. To avoid ascertainment bias, image stacks from every informative cell pair were acquired. Because only a few informative cell pairs were present on individual slides, images were acquired from several slides to generate sufficient data for statistical analysis. Details of quantitative image analysis and statistical analysis are provided as Supplementary Information.

31. Ding, H. *et al.* Regulation of murine telomere length by *Rtel*: an essential gene encoding a helicase-like protein. *Cell* **117**, 873–886 (2004).

32. Gertsenstein, M., Lobe, C. & Nagy, A. ES cell-mediated conditional transgenesis. *Methods Mol. Biol.* **185**, 285–307 (2002).

33. Hörz, W. & Altenburger, W. Nucleotide sequence of mouse satellite DNA. *Nucleic Acids Res.* **9**, 683–696 (1981).

34. Johnson, G. D. *et al.* Fading of immunofluorescence during microscopy: a study of the phenomenon and its remedy. *J. Immunol. Methods* **55**, 231–242 (1982).

35. Allen, J. W. & Latt, S. A. Analysis of sister chromatid exchange formation *in vivo* in mouse spermatogonia as a new test system for environmental mutagens. *Nature* **260**, 449–451 (1976).

LETTERS

High-performance genetically targetable optical neural silencing by light-driven proton pumps

Brian Y. Chow[1,2]*, Xue Han[1,2]*, Allison S. Dobry[1,2], Xiaofeng Qian[1,2], Amy S. Chuong[1,2], Mingjie Li[1,2], Michael A. Henninger[1,2], Gabriel M. Belfort[2], Yingxi Lin[2], Patrick E. Monahan[1,2] & Edward S. Boyden[1,2]

The ability to silence the activity of genetically specified neurons in a temporally precise fashion would provide the opportunity to investigate the causal role of specific cell classes in neural computations, behaviours and pathologies. Here we show that members of the class of light-driven outward proton pumps can mediate powerful, safe, multiple-colour silencing of neural activity. The gene archaerhodopsin-3 (Arch)[1] from *Halorubrum sodomense* enables near-100% silencing of neurons in the awake brain when virally expressed in the mouse cortex and illuminated with yellow light. Arch mediates currents of several hundred picoamps at low light powers, and supports neural silencing currents approaching 900 pA at light powers easily achievable *in vivo*. Furthermore, Arch spontaneously recovers from light-dependent inactivation, unlike light-driven chloride pumps that enter long-lasting inactive states in response to light. These properties of Arch are appropriate to mediate the optical silencing of significant brain volumes over behaviourally relevant timescales. Arch function in neurons is well tolerated because pH excursions created by Arch illumination are minimized by self-limiting mechanisms to levels comparable to those mediated by channelrhodopsins[2,3] or natural spike firing. To highlight how proton pump ecological and genomic diversity may support new innovation, we show that the blue–green light-drivable proton pump from the fungus *Leptosphaeria maculans*[4] (Mac) can, when expressed in neurons, enable neural silencing by blue light, thus enabling alongside other developed reagents the potential for independent silencing of two neural populations by blue versus red light. Light-driven proton pumps thus represent a high-performance and extremely versatile class of 'optogenetic' voltage and ion modulator, which will broadly enable new neuroscientific, biological, neurological and psychiatric investigations.

We screened type I microbial opsins (see Supplementary Table 1) from archaebacteria, bacteria, plants and fungi for light-driven hyperpolarizing capability[5]. Mammalian codon-optimized genes were synthesized, cloned into green fluorescent protein (GFP)-fusion expression vectors, and transfected into cultured neurons. We measured opsin photocurrents and cell capacitance-normalized photocurrent densities under stereotyped illumination conditions (Fig. 1a, black and grey bars, respectively), as well as opsin action spectra (photocurrent as a function of wavelength; Supplementary Table 2). From this information, we estimated the photocurrent density for each opsin at its own spectral peak (Fig. 1a, white bars). For comparison purposes, we included an earlier-characterized microbial opsin, the *Natronomonas pharaonis* halorhodopsin (Halo/NpHR)—a light-driven inward chloride pump capable of modest hyperpolarizing currents[6–9]. Archaerhodopsin-3 from *H. sodomense* (Arch/aR-3), proposed to be a proton pump[1], generated large photocurrents in the screen, as did two other proton pumps, the *Leptosphaeria maculans* opsin (Mac/LR/Ops)[4] and cruxrhodopsin-1 (ref. 10) (albeit less than that of Arch; Fig. 1a). All light-driven chloride pumps assessed had lower screen photocurrents than these light-driven proton pumps.

Arch is a yellow–green light-sensitive (Fig. 1b) opsin that seems to express well on the neural plasma membrane (Fig. 1c; see Supplementary Notes on Arch expression levels and enhancing Arch membrane trafficking). Arch-mediated currents exhibited excellent kinetics of light-activation and post-light recovery. After illumination, Arch currents rose with a 15–85% onset time of 8.8 ± 1.8 ms (mean ± standard error (s.e.) reported throughout, unless otherwise indicated; $n = 16$ neurons), and after light cessation, Arch currents fell with an 85–15% offset time of 19.3 ± 2.9 ms. Under continuous yellow illumination, Arch photocurrent declined (Fig. 1d, e), as did the photocurrents of all of the opsins in our screen. However, unlike all of the halorhodopsins we screened (including products of halorhodopsin site-directed mutagenesis aimed at improving kinetics; Supplementary Table 3), which after illumination remained inactivated for long periods of time (for example, tens of minutes, with accelerated recovery requiring more blue light[6,11]), Arch spontaneously recovered function in seconds (Fig. 1d, e), more like the light-gated cation channel channelrhodopsin-2 from *Chlamydomonas reinhardtii* (ChR2)[2,3]. The magnitude of Arch-mediated photocurrents was large. At low light irradiances of 0.35 and 1.28 mW mm^{-2} (Fig. 1f, left), neural Arch currents were 120 and 189 pA, respectively; at higher light powers (for example, at which Halo currents saturate), Arch currents continued to increase, approaching 900 pA at effective irradiances of 36 mW mm^{-2}, well within the reach of typical *in vivo* experiments (Fig. 1f, right; see Methods for how effective irradiances were calculated). The high dynamic range of Arch may enable the use of light sources (for example, light-emitting diodes (LEDs), lasers) that are safe and effective for optical control *in vivo*[12,13].

Several lines of evidence supported the idea that Arch functioned as an outward proton pump when expressed in neurons. Removing the endogenous ions that commonly subserve neural inhibition, Cl$^-$ and K$^+$, from physiological solutions did not alter photocurrent magnitude ($P > 0.4$ comparing either K$^+$-free or Cl$^-$-free solutions to regular solutions, *t*-test; Fig. 2a). In solutions lacking Na$^+$, K$^+$, Cl$^-$ and Ca^{2+}, photocurrents were still no different from those measured in normal solutions ($P > 0.4$; $n = 4$ neurons tested without these four charge carriers). The reversal potential appeared to be less than −120 mV (Fig. 2b), also consistent with Arch being a proton pump.

We assessed the voltage swings driven by illumination of current-clamped Arch-expressing cultured neurons. As effective irradiance increased from 7.8 mW mm^{-2} to 36.3 mW mm^{-2} (Fig. 1f), voltage-clamped neurons exhibited peak currents that increased from

[1]The MIT Media Laboratory, Synthetic Neurobiology Group, and Department of Biological Engineering, [2]Department of Brain and Cognitive Sciences and MIT McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

*These authors contributed equally to this work.

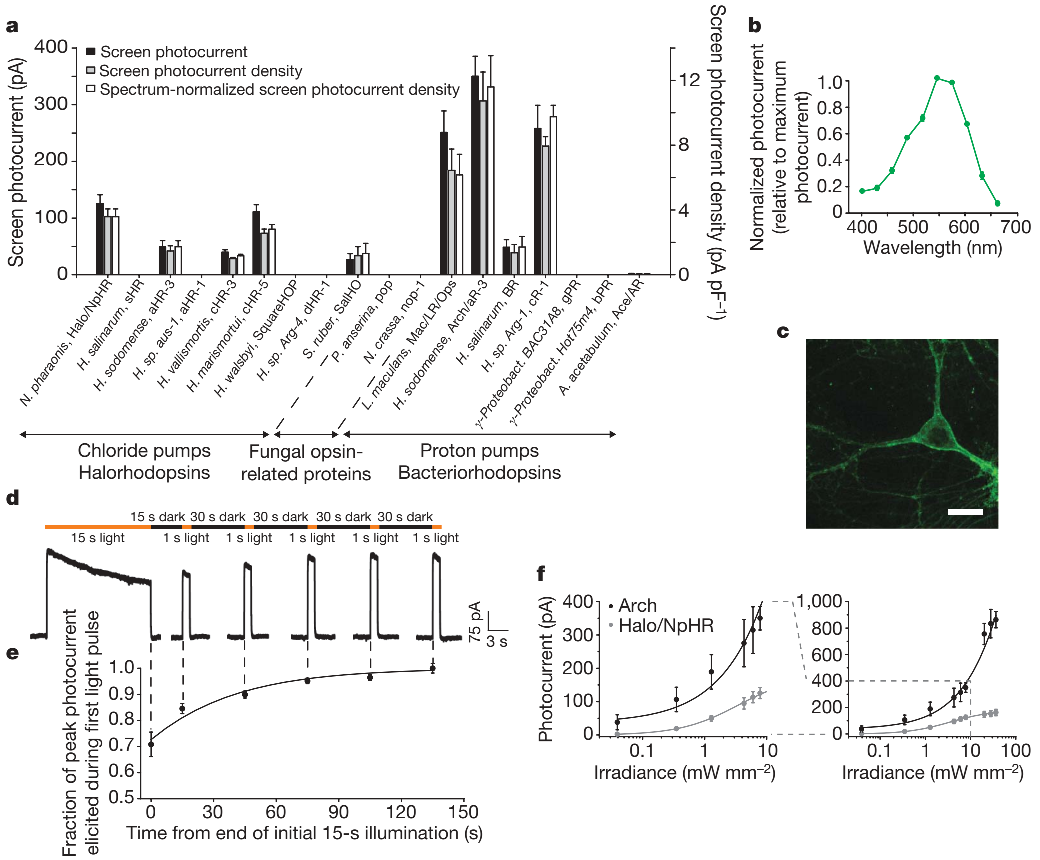


Figure 1 | Optical neural silencing by light-driven proton pumping, revealed by a cross-kingdom functional molecular screen. **a**, Screen data showing outward photocurrents (left y axis, black bars), photocurrent densities (right y axis, grey bars), and action spectrum-normalized photocurrent densities (right y axis, white bars), measured by whole-cell patch-clamp of cultured neurons under screening illumination conditions (575 ± 25 nm, 7.8 mW mm^{-2} for all except Mac/LR/Ops, gPR, bPR and Ace/AR, which were 535 ± 25 nm, 9.4 mW mm^{-2}; see Supplementary Table 1 for details on the molecules screened; $n = 4$–16 neurons for each bar). Data are mean and s.e. Full species names from left to right: *Natronomonas pharaonis*, *Halobacterium salinarum*, *Halorubrum sodomense*, *Halorubrum species aus-1*, *Haloarcula vallismortis*, *Haloarcula marismortui*, *Haloquadratum walsbyi*, *Haloterrigena species Arg-4*, *Salinibacter ruber*, *Podospora anserina*, *Neurospora crassa*, *Leptosphaeria maculans*, *Halorubrum sodomense*, *Halobacterium salinarum*, *Haloarcula species Arg-1*, *uncultured gamma-proteobacterium BAC31A8*, *uncultured gamma-proteobacterium Hot75m4* and *Acetabularia acetabulum*[5]. **b**, Action spectrum of Arch measured in cultured neurons by scanning illumination light wavelength through the visible spectrum ($n = 7$ neurons). **c**, Confocal fluorescence image of a lentivirally infected cultured neuron expressing Arch–GFP. Scale bar, 20 μm. **d**, Raw current trace of a neuron lentivirally infected with Arch, illuminated by a 15-s light pulse (575 ± 25 nm, irradiance 7.8 mW mm^{-2}) followed by 1-s test pulses delivered at 15, 45, 75, 105 and 135 s after the end of the 15-s light pulse. **e**, Population data of averaged Arch photocurrents ($n = 11$ neurons) sampled at the times indicated by the vertical dotted lines that extend into **d**. **f**, Photocurrents of Arch versus Halo measured as a function of 575 ± 25 nm light irradiance (or effective light irradiance; see Methods for details), in patch-clamped cultured neurons ($n = 4$–16 neurons for each point), for low (left) and high (right) light powers. The line is a single Hill fit to the data.

350 ± 35 pA ($n = 16$ neurons) to 863 ± 62 pA ($n = 8$ neurons), respectively. Current-clamped neurons under these two irradiance conditions were hyperpolarized by −69.6 ± 7.3 mV ($n = 10$) and −76.2 ± 10.1 mV ($n = 8$), respectively. Notably, these voltage deflections, although both large, were not significantly different from one another ($P > 0.7$, t-test), suggesting the existence of a rapidly activated transporter or exchanger (perhaps the Na$^+$-dependent Cl$^-$/HCO$_3^-$ exchanger), or the opening of hyperpolarization-gated channels capable of shunting protons, which limit the effects of Arch on accumulated proton (or other charge carrier) gradients across neural membranes. This enabling of effective but not excessive silencing may make Arch safer than pumps that accumulate ions without self-regulation.

We next assessed the changes in intracellular pH (pH$_i$) driven by illumination of Arch-expressing cultured neurons, using the fluorescent pH indicator carboxy-SNARF-1. Within 1 s of illumination with strong green light (Fig. 2c), pH$_i$ rose from 7.309 ± 0.011 to 7.431 ± 0.020, plateauing rapidly. pH$_i$ increased slightly further after 15 s of illumination to 7.461 ± 0.024 (Fig. 1e). The fast stabilization of pH$_i$ may reflect the same self-limiting influence that limits proton-mediated voltage swings as described earlier, and may contribute to the safe operation of Arch in neurons by preventing large pH$_i$ swings.

The changes in pH$_i$ observed here are comparable in magnitude to those observed during illumination of ChR2-expressing cells[14] (owing to the proton currents carried by ChR2; refs 3, 15), and are also within the magnitudes of changes observed during normal neural activity[16–19]. Passive electrical properties of neurons were not affected by Arch expression (Fig. 2e–g; $P > 0.6$ for each measure, t-test), nor was cell death ($P > 0.6$, $\chi^2 = 0.26$; Fig. 2d).

We estimated the tissue volumes that could be silenced, using *in vitro* experiments and computational modelling. In cultured neurons expressing Arch or a trafficking-improved variant of Halo, eNpHR[8,9], we somatically injected brief current pulses at magnitudes chosen to mimic the current drives of neurons in the intact nervous system[20–23]. We exposed these neurons to periods of 575 nm yellow light (0.35, 1.28 or 6 mW mm^{-2}, simulating irradiance ~1.7, 1.2 or 0.6 mm away from the tip of a 200-μm fibre emitting 200 mW mm^{-2} irradiance, as modelled by Monte Carlo methods; see Supplementary Fig. 3), and measured the reduction in spike rate for each condition (Fig. 3a). In general, Arch-expressing neurons were significantly more inhibited than eNpHR-expressing cells. According to our model and the 350 pA data in Fig. 3a, the increase in brain tissue volume that would be 45–55% optically silenced would be ~10 times larger for Arch than for eNpHR.

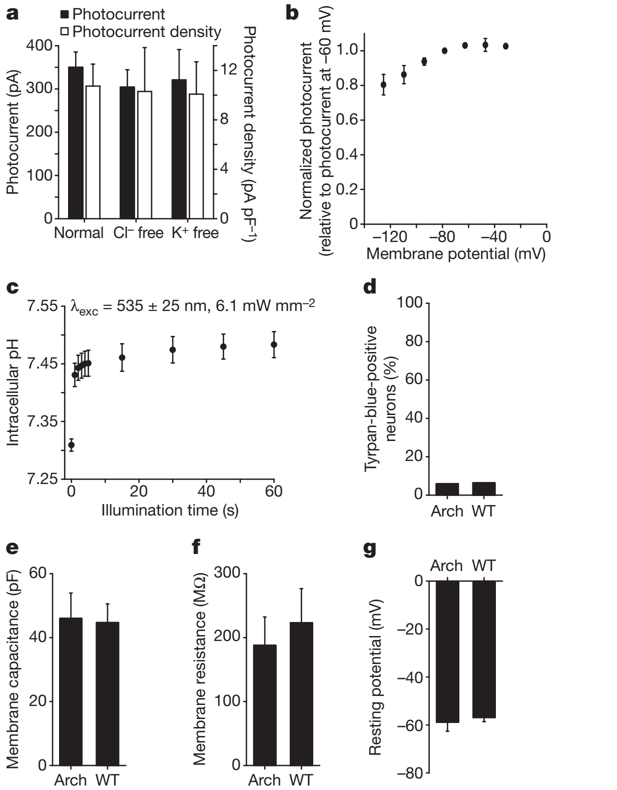


Figure 2 | Functional properties of the light-driven proton pump Arch in neurons. a, Photocurrent of Arch measured as a function of ionic composition (575 ± 25 nm light, 7.8 mW mm^{-2}), showing no significant dependence of photocurrent on the concentration of Cl^- or K^+ ions in bath and intracellular solutions ($n = 16$, 8 and 7 neurons, from left to right). **b**, Arch proton photocurrent versus holding potential ($n = 4$ neurons). **c**, Intracellular pH measurements over a 1-min period of continuous illumination and simultaneous imaging (535 ± 25 nm light, 6.1 mW mm^{-2}) using SNARF-1 pH-sensitive ratiometric dye ($n = 10$–20 cells per data point). **d**, Trypan-blue staining of neurons lentivirally infected with Arch versus wild-type (WT) neurons, measured at 18 days *in vitro* ($n = 669$ Arch-expressing, 512 wild-type neurons). **e**–**g**, Membrane capacitance (**e**), membrane resistance (**f**), and resting potential (**g**) in neurons lentivirally infected with Arch versus wild-type neurons, measured at 11 days *in vitro* ($n = 7$ cells each).

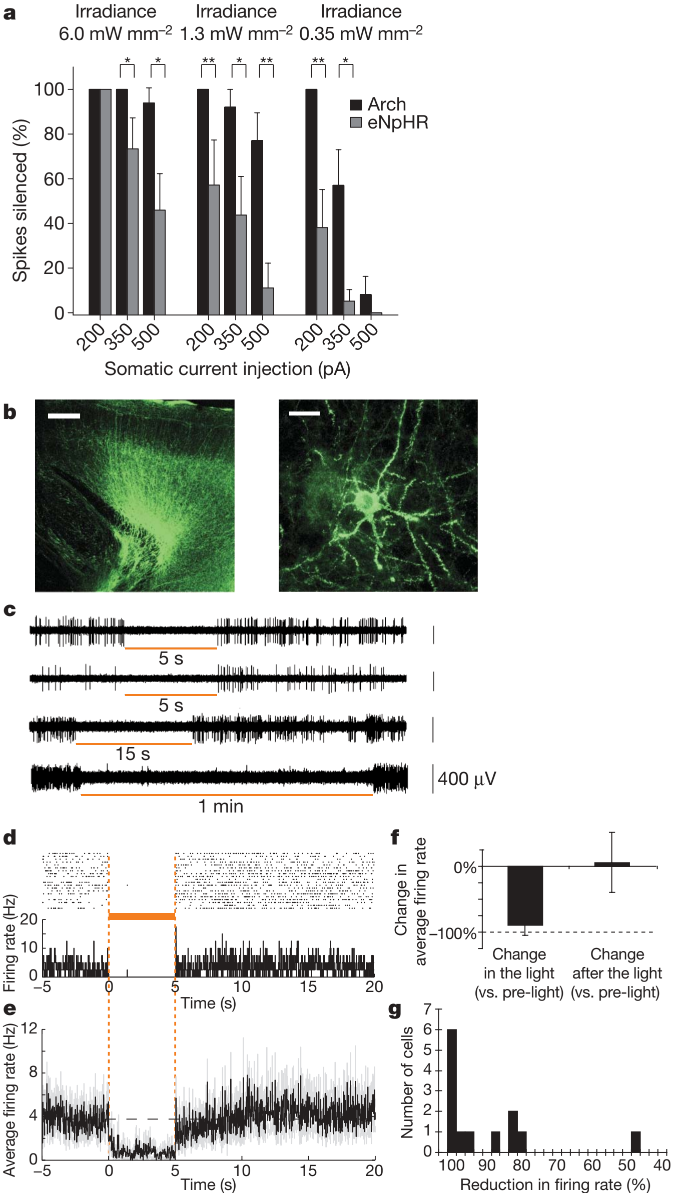


Figure 3 | High-performance Arch-mediated optical neural silencing of neocortical regions in awake mice. a, *In vitro* data showing, in cultured neurons expressing Arch or eNpHR and receiving trains of somatic current injections (15-ms pulse durations at 5 Hz), the per cent reduction of spiking under varying light powers (575 ± 25 nm light) as might be encountered *in vivo*. $^*P < 0.05$, $^{**}P < 0.01$, *t*-test. $n = 7$–8 cells for each condition. **b**, Fluorescence images showing Arch–GFP expression in mouse cortex ~1 month after lentiviral (FCK–Arch–GFP) injection. Scale bars, 200 μm (left) and 20 μm (right). **c**, Representative extracellular recordings showing neurons undergoing 5-s, 15-s and 1-min periods of light illumination (593 nm; ~150 mW mm^{-2} radiant flux out the fibre tip; and expected to be ~3 mW mm^{-2} at the electrode tip ~800 μm away[12,13,24]). **d**, Neural activity in a representative neuron before, during and after 5 s of yellow light illumination, shown as a spike raster plot (top), and as a histogram of instantaneous firing rate averaged across trials (bottom; bin size, 20 ms). **e**, Population average of instantaneous firing rate before, during and after yellow light illumination (black line, mean; grey lines, mean ± s.e.; $n = 13$ units). **f**, Average change in spike firing during 5-s of yellow light illumination (left) and during the 5 s immediately after light offset (right), for the data shown in **d**. **g**, Histogram of percentage reductions in spike rate, for each individual neuron, integrated across all 5-s silencing period.

To assess Arch *in vivo* directly, we injected lentivirus encoding for Arch into mouse cortex and recorded neural responses ~1 month later. Arch expressed well (Fig. 3b, left) and appeared well localized to the plasma membrane, labelling cell bodies, processes and dendritic spines (Fig. 3b, right). We recorded neurons in awake head-fixed mice, illuminating neurons by a 200-μm optical fibre coupled to a 593-nm laser (power at electrode tip estimated at ~3 mW mm^{-2}; refs 12, 13, 24). After light onset, firing rates of many units immediately and strongly declined, and remained low throughout the period of illumination, for both brief (Fig. 3c, top, d) and long (Fig. 3c, bottom) pulses. We recorded 13 single units that showed any decrease in firing during illumination, objectively identified as described in the Methods, and found spiking rates during exposure to 5 s yellow light (Fig. 3d) to drop by an average of 90 ± 15% (mean ± s.d.; Fig. 3e, f), restoring to levels indistinguishable from baseline after light cessation ($P > 0.2$, paired *t*-test; Fig. 3f). Six of the 13 units decreased spike rate by at least 99.5%, and the median decrease was 97.1% (Fig. 3g). One possibility is that Arch-expressing cells were almost completely silenced, whereas non-infected cells decreased activity owing to network activity reduction during illumination; note that only excitatory cells were genetically targeted here. Optical silencing was consistent across trials ($P > 0.1$, paired *t*-test comparing, for each neuron, responses to the first three versus the last three light exposures; ~20 trials per neuron). The kinetics of silencing were rapid: for the six neurons that underwent >99.5% silencing, spike firing reduced with near-0-ms latency, rarely firing spikes after light onset; averaged across all cells, firing-rate reductions plateaued within 229 ± 310 ms (mean ± s.d.) after light onset. After light cessation, firing rate restored quickly for the highly silenced neurons; averaged across all cells, firing rates took 355 ± 505 ms to recover after light offset. The level of post-light firing did not vary with repeated light exposure ($P > 0.7$, paired t-test comparing, for each neuron, after-light firing rates during the first three versus the last three trials). Thus, Arch could mediate reliable, near-digital silencing of neurons in the awake mammalian brain.

Proton pumps naturally exist that are activated by many colours of light (see Supplementary Table 1), in contrast to chloride pumps, which are primarily driven by yellow–orange light (even with significant mutagenesis of retinal-flanking residues; Supplementary Table 3). The light-driven proton pump Mac (Fig. 1a), in our screen, had an action spectrum strongly blueshifted relative to that of the light-driven chloride pump Halo (Fig. 4a). We found that Mac-expressing neurons could undergo 4.1-fold larger hyperpolarizations with blue light than with red light, and Halo-expressing neurons could undergo 3.3-fold larger hyperpolarizations with red light than with blue light, when illuminated with appropriate filters (Fig. 4b). Accordingly, we could demonstrate selective silencing of spike firing in Mac-expressing neurons in response to blue light, and selective silencing of spike firing in Halo-expressing neurons in response to red light (Fig. 4c). Thus, the spectral diversity of proton pumps points the way towards independent multicolour silencing of separate neural populations. This result opens up new kinds of experiment, in which, for example, two neuron classes, or two sets of neural projections from a single site, can be independently silenced during a behavioural task.

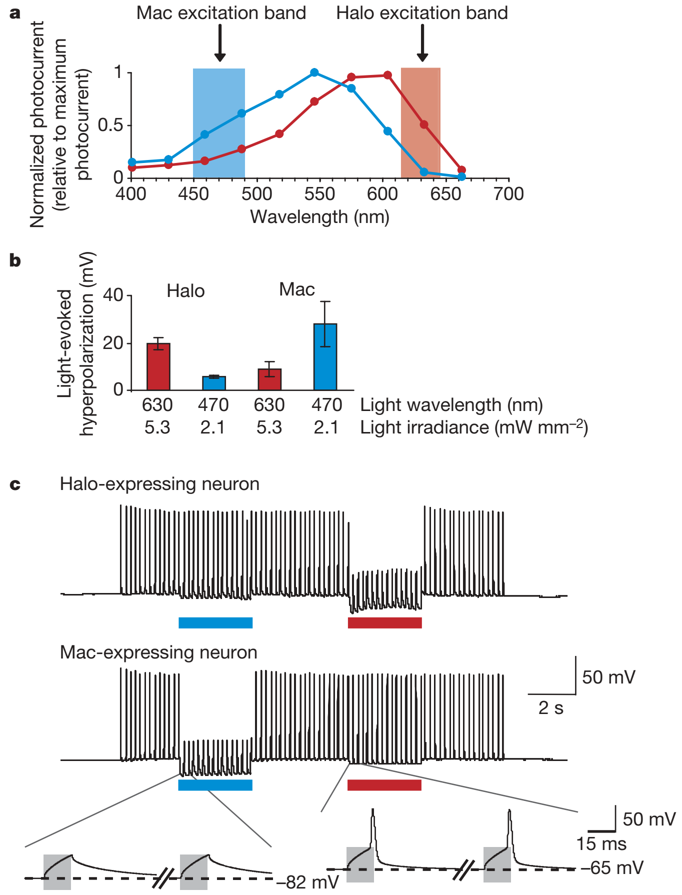


Figure 4 | Multicolour silencing of two neural populations, enabled by blue- and red-light drivable ion pumps of different classes. a, Action spectra of Mac versus Halo; rectangles indicate filter bandwidths used for multicolour silencing *in vitro*. Blue light is delivered by a 470 ± 20 nm filter at 5.3 mW mm^{-2}, and red light is delivered by a 630 ± 15 nm filter at 2.1 mW mm^{-2}. **b,** Membrane hyperpolarizations elicited by blue versus red light, in cells expressing Halo or Mac ($n = 5$ Mac-expressing and $n = 6$ Halo-expressing neurons). **c,** Action potentials evoked by current injection into patch-clamped cultured neurons transfected with Halo (top) were selectively silenced by the red light but not by the blue light, and vice-versa in neurons expressing Mac (middle). Grey boxes in the inset (bottom) indicate periods of patch-clamp current injection.

Arch and Mac represent members of a new, diverse and powerful class of optical neural silencing reagent, the light-driven proton pump, which operates without the need for exogenous chemical supplementation in mammalian cells. The efficacy of these proton pumps is surprising, given that protons occur, in mammalian tissue, at a million-fold lower concentration than the ions carried by other optical control molecules. This high efficacy may be due to the fast photocycle of Arch (see also refs 25, 26), but it may also be due to the ability of high-pK_a residues in proton pumps to mediate proton uptake[25,27]. We discovered several facts about this class of molecules that point the way for future neuroengineering innovation. First, proton pumping is a self-limiting process in neurons, providing for a safe and naturalistic form of neural silencing. Second, proton pumps recover spontaneously after optical activation, improving their relevance for behaviourally relevant silencing over the class of halorhodopsins. Finally, proton pumps exist with a wide diversity of action spectra, thus enabling multiple-colour silencing of distinct neural populations. Structure-guided mutagenesis of Arch and Mac may further facilitate development of neural silencers with altered spectrum or ion selectivity, given the significant amount of structure–function knowledge of the proton pump family (for example, refs 28–30).

Our study highlights the importance of ecological and genomic diversity in providing new molecular reagents for optical control of biological processes, as has previously benefited the fluorescent protein community. These opsins are likely to find uses across the spectrum of neuroscientific, biological and bioengineering fields. For example, expression of these opsins in neurons, muscle, immune cells and other excitable cells will allow control over their membrane potential, providing the opportunity to investigate the causal role of specific cells' activities in intact organisms, and, potentially, to understand the causal contribution of such cells to disease states in animal models. With the recent demonstration of the safe and efficacious use of the microbial opsin ChR2 to control neurons in non-human primates[24], it is in principle possible that in the future, these opsins may subserve new forms of neuromodulation technology that bear clinical benefit.

METHODS SUMMARY

Constructs with Arch, Mac and Halo are available at http://syntheticneurobiology.org/protocols. In brief, codon-optimized genes were synthesized by Genscript and fused to GFP in lentiviral and mammalian expression vectors as used previously[6,24] for transfection or viral infection of neurons. Primary hippocampal or cortical neurons were cultured and then transfected with plasmids or infected with viruses encoding for genes of interest, as described previously[6]. Images were taken using a Zeiss LSM 510 confocal microscope. Patch-clamp recordings were made using glass microelectrodes and a Multiclamp 700B/Digidata electrophysiology setup, using appropriate pipette and bath solutions for the experimental goal at hand. Neural pH imaging was done using carboxy-SNARF-1-AM ester (Invitrogen). Cell health was assayed using trypan blue staining (Gibco). HEK cells were cultured and patch-clamped using standard protocols. Mutagenesis was performed using the QuikChange kit (Stratagene). Computational modelling of light propagation was done with Monte Carlo simulation with MATLAB. *In vivo* recordings were made on head-fixed awake mice, which were surgically injected with lentivirus, and implanted with a headplate as described before[24]. Glass pipettes attached to laser-coupled optical fibres were inserted into the brain, to record neural activity during laser illumination in a photoelectrochemical artefact-free way. Data analysis was performed using Clampfit, Excel, Origin and MATLAB. Histology was performed using transcardial formaldehyde perfusion followed by sectioning and subsequent confocal imaging.

Received 27 August; accepted 9 November 2009.

1. Ihara, K. *et al.* Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *J. Mol. Biol.* 285, 163–174 (1999).
2. 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).

3. Nagel, G. *et al.* Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc. Natl Acad. Sci. USA* **100,** 13940–13945 (2003).
4. Waschuk, S. A., Bezerra, A. G., Shi, L. & Brown, L. S. Leptosphaeria rhodopsin: Bacteriorhodopsin-like proton pump from a eukaryote. *Proc. Natl Acad. Sci. USA* **102,** 6879–6883 (2005).
5. Klare, J. P., Chizhov, I. & Engelhard, M. Microbial rhodopsins: scaffolds for ion pumps, channels, and sensors. *Results Probl. Cell Differ.* **45,** 73–122 (2008).
6. Han, X. & Boyden, E. S. Multiple-color optical activation, silencing, and desynchronization of neural activity, with single-spike temporal resolution. *PLoS One* **2,** e299 (2007).
7. Zhang, F. *et al.* Multimodal fast optical interrogation of neural circuitry. *Nature* **446,** 633–639 (2007).
8. Zhao, S. *et al.* Improved expression of halorhodopsin for light-induced silencing of neuronal activity. *Brain Cell Biol.* **36,** 141–154 (2008).
9. Gradinaru, V., Thompson, K. R. & Deisseroth, K. eNpHR: a Natronomonas halorhodopsin enhanced for optogenetic applications. *Brain Cell Biol.* **36,** 129–139 (2008).
10. Tateno, M., Ihara, K. & Mukohata, Y. The novel ion pump rhodopsins from *Haloarcula* form a family independent from both the bacteriorhodopsin and archaerhodopsin families/tribes. *Arch. Biochem. Biophys.* **315,** 127–132 (1994).
11. Bamberg, E., Tittor, J. & Oesterhelt, D. Light-driven proton or chloride pumping by halorhodopsin. *Proc. Natl Acad. Sci. USA* **90,** 639–643 (1993).
12. Aravanis, A. M. *et al.* An optical neural interface: *in vivo* control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.* **4,** S143–S156 (2007).
13. Bernstein, J. G. *et al.* Prosthetic systems for therapeutic optical activation and silencing of genetically-targeted neurons. *Proc. Soc. Photo Opt. Instrum. Eng.* **6854,** 68540H (2008).
14. Lin, J. Y., Lin, M. Z., Steinbach, P. & Tsien, R. Y. Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophys. J.* **96,** 1803–1814 (2009).
15. Berthold, P. *et al.* Channelrhodopsin-1 initiates phototaxis and photophobic responses in *Chlamydomonas* by immediate light-induced depolarization. *Plant Cell* **20,** 1665–1677 (2008).
16. Bevensee, M. O., Cummins, T. R., Haddad, G. G., Boron, W. F. & Boyarsky, G. pH regulation in single CA1 neurons acutely isolated from the hippocampi of immature and mature rats. *J. Physiol. (Lond.)* **494,** 315–328 (1996).
17. Chesler, M. Regulation and modulation of pH in the brain. *Physiol. Rev.* **83,** 1183–1221 (2003).
18. Meyer, T. M., Munsch, T. & Pape, H. C. Activity-related changes in intracellular pH in rat thalamic relay neurons. *Neuroreport* **11,** 33–36 (2000).
19. Trapp, S., Luckermann, M., Brooks, P. A. & Ballanyi, K. Acidosis of rat dorsal vagal neurons in situ during spontaneous and evoked activity. *J. Physiol. (Lond.)* **496,** 695–710 (1996).
20. Leinekugel, X. *et al.* Correlated bursts of activity in the neonatal hippocampus *in vivo. Science* **296,** 2049–2052 (2002).
21. Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426,** 442–446 (2003).
22. Richter, D. W., Pierrefiche, O., Lalley, P. M. & Polder, H. R. Voltage-clamp analysis of neurons within deep layers of the brain. *J. Neurosci. Methods* **67,** 121–131 (1996).
23. Narikawa, K., Furue, H., Kumamoto, E. & Yoshimura, M. *In vivo* patch-clamp analysis of IPSCs evoked in rat substantia gelatinosa neurons by cutaneous mechanical stimulation. *J. Neurophysiol.* **84,** 2171–2174 (2000).
24. Han, X. *et al.* Millisecond-timescale optical control of neural dynamics in the nonhuman primate brain. *Neuron* **62,** 191–198 (2009).
25. Ming, M. *et al.* pH dependence of light-driven proton pumping by an archaerhodopsin from Tibet: comparison with bacteriorhodopsin. *Biophys. J.* **90,** 3322–3332 (2006).
26. Lukashev, E. P. *et al.* pH dependence of the absorption spectra and photochemical transformations of the archaerhodopsins. *Photochem. Photobiol.* **60,** 69–75 (1994).
27. Lanyi, J. K. Proton transfers in the bacteriorhodopsin photocycle. *Biochim. Biophys. Acta* **1757,** 1012–1018 (2006).
28. Enami, N. *et al.* Crystal structures of archaerhodopsin-1 and -2: common structural motif in archaeal light-driven proton pumps. *J. Mol. Biol.* **358,** 675–685 (2006).
29. Mogi, T., Marti, T. & Khorana, H. G. Structure-function studies on bacteriorhodopsin. IX. Substitutions of tryptophan residues affect protein-retinal interactions in bacteriorhodopsin. *J. Biol. Chem.* **264,** 14197–14201 (1989).
30. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291,** 899–911 (1999).

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

Acknowledgements E.S.B. acknowledges funding by the NIH Director's New Innovator Award (DP2 OD002002-01), as well as the NSF (0835878 and 0848804), the McGovern Institute Neurotechnology Award Program, the Department of Defense, NARSAD, the Alfred P. Sloan Foundation, Jerry and Marge Burnett, the SFN Research Award for Innovation in Neuroscience, the MIT Media Lab, the Benesse Foundation, and the Wallace H. Coulter Foundation. X.H. acknowledges the Helen Hay Whitney Foundation and NIH 1K99MH085944. The authors thank E. Klinman for help with transfections, R. Desimone for advice, J. Lin for technical aid on intracellular pH measurements, K. Ihara for discussions about archaerhodopsins, and M. Hemann and N. Gershenfeld and the Center for Bits and Atoms for use of their respective laboratory facilities.

Author Contributions B.Y.C., X.H. and E.S.B. designed experiments, analysed data and wrote the paper. B.Y.C. and X.H. carried out experiments. A.S.D. assisted with electrophysiological recording. X.Q., M.L. and A.S.C. assisted with molecular biology, virus making, and transfections. M.A.H. performed Monte Carlo modelling. P.E.M., G.M.B. and Y.L. created hippocampal and cortical neural cultures.

Author Information Sequences are available to download from GenBank (http://www.ncbi.nlm.nih.gov/) under accession numbers: GU045593 (mammalian codon-optimized Arch), GU045594 (mammalian codon-optimized Arch fused to GFP), GU045595 (mammalian codon-optimized Mac), GU045596 (mammalian codon-optimized Mac fused to GFP), GU045597 (ss-Prl-Arch), GU045598 (ss-Arch-GFP-ER2) and GU045599 (ss-Prl-Arch-GFP). 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 E.S.B. (esb@media.mit.edu).

Vol 463 | 7 January 2010 | doi:10.1038/nature08646

Deubiquitinase USP9X stabilizes MCL1 and promotes tumour cell survival

Martin Schwickart[1]*, XiaoDong Huang[1]*, Jennie R. Lill[2], Jinfeng Liu[3], Ronald Ferrando[4], Dorothy M. French[4], Heather Maecker[5], Karen O'Rourke[1], Fernando Bazan[6], Jeffrey Eastham-Anderson[4], Peng Yue[3], David Dornan[7], David C. S. Huang[8] & Vishva M. Dixit[1]

MCL1 is essential for the survival of stem and progenitor cells of multiple lineages[1,2], and is unique among pro-survival BCL2 family members in that it is rapidly turned over through the action of ubiquitin ligases[3–6]. B- and mantle-cell lymphomas, chronic myeloid leukaemia, and multiple myeloma[7–9], however, express abnormally high levels of MCL1, contributing to chemoresistance and disease relapse. The mechanism of MCL1 overexpression in cancer is not well understood. Here we show that the deubiquitinase USP9X stabilizes MCL1 and thereby promotes cell survival. USP9X binds MCL1 and removes the Lys 48-linked polyubiquitin chains that normally mark MCL1 for proteasomal degradation. Increased USP9X expression correlates with increased MCL1 protein in human follicular lymphomas and diffuse large B-cell lymphomas. Moreover, patients with multiple myeloma overexpressing USP9X have a poor prognosis. Knockdown of USP9X increases MCL1 polyubiquitination, which enhances MCL1 turnover and cell killing by the BH3 mimetic ABT-737. These results identify USP9X as a prognostic and therapeutic target, and they show that deubiquitinases may stabilize labile oncoproteins in human malignancies.

Failure to eliminate damaged cells by apoptosis is vital in tumour development[10]. Pro-survival BCL2 family members can be overexpressed owing to gene translocation or amplification, but in the case of MCL1, altered post-translational ubiquitination can increase MCL1 protein levels without the need for altered *MCL1* transcription[4,5,11]. HeLa and HT1080 cells treated with the proteasome inhibitor MG-132, for example, failed to degrade ubiquitinated MCL1 and contained more MCL1 protein (Supplementary Fig. 1). To identify deubiquitinases (DUBs) that remove ubiquitin from MCL1, we analysed proteins co-immunoprecipitating specifically with 3×Flag-tagged MCL1 from HEK293T cells. MULE, a ubiquitin ligase for MCL1 (ref. 4), interacted specifically with MCL1, as did the DUB USP9X[11] (Fig. 1a, b). Whereas the carboxy-terminal BCL2 homology domains of MCL1 bind the BH3 domain of MULE[4], the amino terminus of MCL1 was necessary for binding USP9X (Supplementary Fig. 2a, b). The interaction of MCL1 with USP9X (or MULE) occurred between endogenous (Fig. 1c) and *in vitro* translated proteins (Supplementary Fig. 2c), the latter suggesting direct binding. MCL1 interacted with USP9X and MULE less efficiently than with the pro-apoptotic protein NOXA (also known as PMAIP1)[12] (data not shown), as only a portion of endogenous USP9X and MULE in U2OS, Colo201 or Colo205 cells was in the membrane-enriched cellular fraction containing MCL1 (Supplementary Fig. 3a). Immunofluorescence microscopy of U2OS cells confirmed that most MCL1 but only a portion of USP9X co-localized with mitochondria (Supplementary Fig. 3b, c).

If the USP9X–MCL1 interaction enhances MCL1 stability, then human tumours overexpressing MCL1 protein might also overexpress USP9X. Immunoblotting identified five human follicular lymphomas with increased USP9X protein, and all contained more MCL1 than normal lymph nodes (Fig. 2a). Immunohistochemical staining of additional follicular lymphoma ($n = 25$) and normal lymphoid tissue samples ($n = 13$) (Fig. 2b–d) showed that MCL1 and USP9X levels were increased in the lymphomas (Mann–Whitney–Wilcoxon test; $P = 0.003$, MCL1; $P = 0.001$, USP9X), and their expression correlated significantly ($P = 0.0006$, Spearman's rank correlation). We do not think that differences in USP9X levels reflect proliferating tumour cells versus predominantly quiescent normal tissues, because IgM-stimulated normal human B cells did not express significantly more USP9X than freshly isolated, resting B cells (Supplementary Fig. 4). Increased MCL1 in the activated B cells was probably due to increased *MCL1* transcription (Supplementary Fig. 4)[13].

The correlation between USP9X and MCL1 was not restricted to follicular lymphomas: four out of five human ductal adenocarcinomas, four out of five human colon adenocarcinomas, and three out of three human small cell lung carcinomas had increased MCL1 and USP9X (Supplementary Fig. 5), and a correlation was observed in diffuse large B-cell lymphoma (DLBCL) (Supplementary Fig. 6a; Spearman's rank correlation coefficient = 0.51, $P = 0.00016$). USP9X protein and messenger RNA correlated in DLBCL cell lines (Supplementary Fig. 6b, c), suggesting that increased USP9X protein in cases of DLBCL reflects increased *USP9X* transcription. Interrogation of public expression databases[14,15] showed that increased *USP9X* mRNA in tumours was significantly associated with poor prognosis for patients with multiple myeloma; patients expressing *USP9X* mRNA highly had a 5.5-fold greater risk of death (Fig. 2e). Comparison of *USP9X* mRNA expression to other gene expression-based predictors of survival with multiple myeloma, such as the 70-gene high-risk predictor, indicated that USP9X provides an extra prognostic signal independent of the high-risk predictor (Fig. 2f; Cox regression, $P = 0.013$). USP9X also provides a further prognostic signal for multiple myeloma with a low PR (proliferation) signature[16] (Fig. 2g; Cox regression, $P = 0.005$). Our findings indicate that the interaction of USP9X and MCL1 is of prognostic relevance for several human malignancies.

Next we sought direct evidence that USP9X affects MCL1 stability. Knockdown of USP9X in HEK293T cells with short interfering RNAs (siRNAs) decreased MCL1 protein (Fig. 3a) but not mRNA (Supplementary Fig. 7a). MULE and β-tubulin were unaffected (Fig. 3a). Decreased MCL1 was a specific effect of USP9X knockdown because

[1]Department of Physiological Chemistry, [2]Department of Protein Chemistry, [3]Department of Bioinformatics, [4]Department of Pathology, [5]Department of Translational Oncology, [6]Department of Protein Engineering, [7]Department of Research Oncology Diagnostics, Genentech, Inc., 1 DNA Way, South San Francisco, California 94080, USA. [8]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia.
*These authors contributed equally to this work.

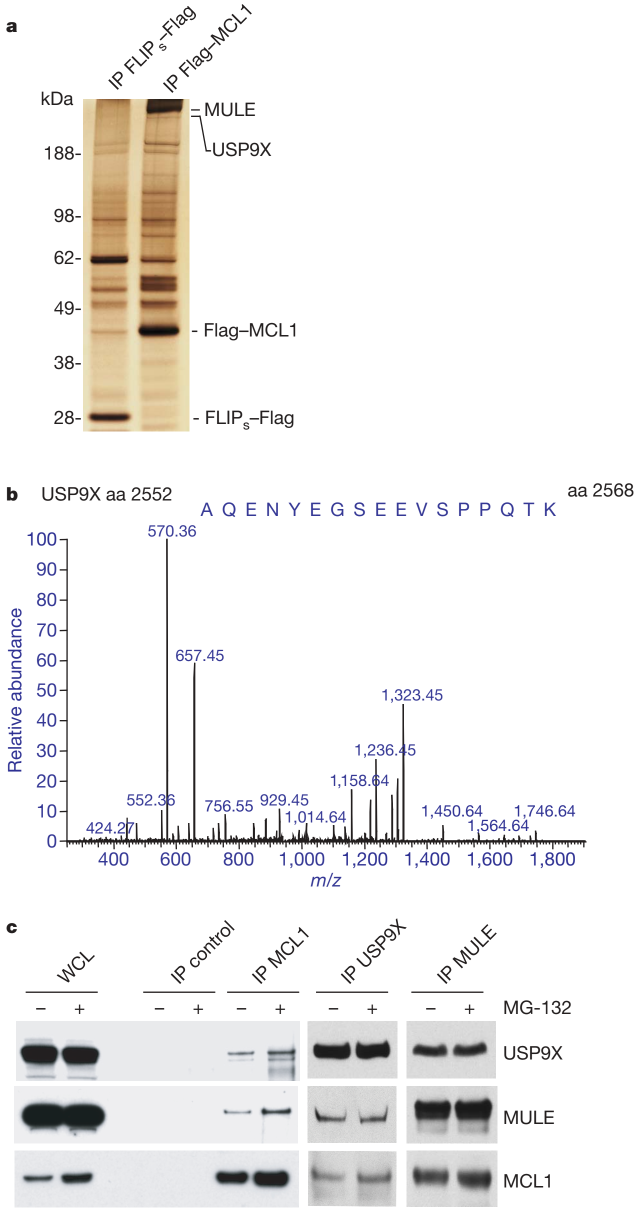


Figure 1 | USP9X binds MCL1. a, Silver-stained gel of proteins co-immunoprecipitated from HEK293T cells with 3×Flag–MCL1 or control protein FLIP$_S$ (also known as CFLAR)–Flag. IP, immunoprecipitation. **b**, Fragmentation spectrum of the only identified USP9X peptide. aa, amino acid. **c**, HEK293T cells treated with (+) or without (−) 10 μM MG-132 for 4 h were used for endogenous MCL1, USP9X and MULE immunoprecipitations. WCL, whole cell lysate.

siRNAs targeting the USP9X 3′ untranslated region (UTR) had no effect in cells expressing ectopic USP9X without its 3′ UTR. In contrast, ectopic USP9X with catalytic residue Cys 1566 mutated to serine (Cys1566Ser) did not counter the effects of the siRNAs (Supplementary Fig. 7b). Modulating USP9X altered the level of MCL1 protein in diverse cell types; expression of USP9X in the USP9X-deficient melanoma cell line LOX.IMVI increased MCL1 (Supplementary Fig. 7c), whereas knockdown of USP9X in HeLa, HEK293T and HCT116 cells decreased MCL1 protein but not mRNA (Fig. 3a, b and Supplementary Fig. 7a, d). Reportedly, USP9X deubiquitinates ITCH—a ubiquitin ligase for transcription factors p63 (also known as TP63) and p73 (TP73)[17]. USP9X knockdown decreased ITCH levels, but no significant changes in p63 and p73 were detected (Supplementary Fig. 7d). BCL2 protein levels increased after USP9X knockdown in HeLa cells (Fig. 3b), but this did not occur in HEK293T or HCT116 cells (Supplementary Fig. 7d). Increased BCL2 in HeLa cells was probably due to increased *BCL2* transcription (Supplementary Fig. 7e) by an unknown mechanism.

Next we determined the effect of USP9X on MCL1 turnover. USP9X knockdown in HeLa cells reduced the half-life of MCL1 from ~17 to ~9 min (Fig. 3b, c). Binding of USP9X was essential for MCL1 stabilization because a mutant MCL1 protein unable to bind USP9X (MCL1ΔN; Supplementary Fig. 2b) was less stable than wild-type MCL1 (Supplementary Fig. 7f). Consistent with USP9X removing ubiquitin from MCL1 to preserve its stability, cells contained more ubiquitinated MCL1 after USP9X knockdown. Conversely, knockdown of MULE reduced the amount of ubiquitinated MCL1 (Fig. 3d). To see whether USP9X deubiquitinated MCL1 *in vitro*, purified MCL1 and USP9X were mixed under conditions supporting MCL1 deubiquitination by the promiscuous catalytic subunit of rat Usp2 (Fig. 3e). USP9X, but not mutant USP9X(Cys1566Ser), was comparable to Usp2 at deubiquitinating MCL1 and generating free monoubiquitin. USP9X probably removed degradative Lys 48-linked polyubiquitin chains[18], because MCL1 purified from HEK293T cells was immunoblotted with a Lys 48 linkage-specific but not a Lys 63 linkage-specific polyubiquitin antibody (Supplementary Fig. 8). Furthermore, USP9X, but not mutant USP9X(Cys1566Ser), cleaved Lys 48-linked di-, tri- and tetraubiquitin, resulting in increased monoubiquitin (Fig. 3f), although USP9X cleaved synthetic polyubiquitin chains slower than polyubiquitin on MCL1 (data not shown). Together, these data suggest that USP9X stabilizes MCL1 by removing its degradative Lys 48-linked polyubiquitin chains.

Phosphorylation of MCL1 residues Thr 92 and Thr 163 by extracellular signal-regulated kinases (ERKs) is reported to stabilize MCL1 (refs 19 and 20). Neither ERK activation by epidermal growth factor nor ERK inhibition with PD0325901 affected the USP9X–MCL1 interaction in HEK293T cells (Supplementary Fig. 9a), and *USP9X* and *MCL1* mRNA levels were unchanged (Supplementary Fig. 9b). Mutation of both ERK phosphorylation sites in MCL1 to either alanine or aspartic acid also had no effect on USP9X binding to MCL1 (Supplementary Fig. 9c). These data indicate that the regulation of MCL1 by USP9X is independent of ERK activity. Phosphorylation of MCL1 residue Ser 159 by GSK3β was reported to increase MCL1 turnover and thereby promote apoptosis[21]. Mutation of MCL1 residues Ser 155, Ser 159 and Thr 163 to alanine enhanced the interaction of MCL1 with USP9X, whereas mutation of these three serine residues to aspartic acid as a phospho-mimic decreased the interaction (Supplementary Fig. 9c). These results indicate that MCL1 phosphorylation at these sites, perhaps by GSK3β, disrupts the USP9X–MCL1 interaction. The phosphatidylinositol 3-kinase (PI3K) inhibitor LY294002 was shown to reduce MCL1 levels rapidly, probably as a result of GSK3β activation[22], and it, too, decreased the binding of USP9X to MCL1 (Supplementary Fig. 9d). These data indicate that the USP9X–MCL1 interaction is regulated and context-dependent.

ABT-737 is a small molecule antagonist of the pro-survival proteins BCL2, BCL-x$_L$ (also known as BCL2L1) and BCL-w (BCL2L2), but it does not inhibit MCL1 or A1 (BCL2A1)[6]. It kills tumours with low MCL1, whereas tumours and cell-line derivatives overexpressing MCL1 are resistant to single agent ABT-737 (ref. 23). For example, colon carcinoma cell lines HCT116 and DLD-1 and the leukaemia cell line D1.1 are dependent on MCL1 for their survival and are relatively resistant to ABT-737 (refs 6 and 23). MCL1 knockdown, however, rendered these cells sensitive to ABT-737 (refs 6 and 24). USP9X knockdown reduced MCL1 in HCT116, DLD-1 and D1.1 cells (Supplementary Fig. 10) and increased sensitivity to ABT-737 by up to fivefold (Fig. 4a–c). The more USP9X was knocked down, the greater the reduction in MCL1, and the increase in sensitivity to ABT-737 (Supplementary Fig. 11). The cells still contained residual MCL1, so they were not as sensitive to ABT-737 as cells subjected to MCL1 knockdown (up to 14-fold sensitization) (Fig. 4a–c). The effect of USP9X on tumour cell sensitivity to ABT-737 *in vivo* was investigated with a xenograft tumour model. Pancreatic BxPC-3 cells expressing a

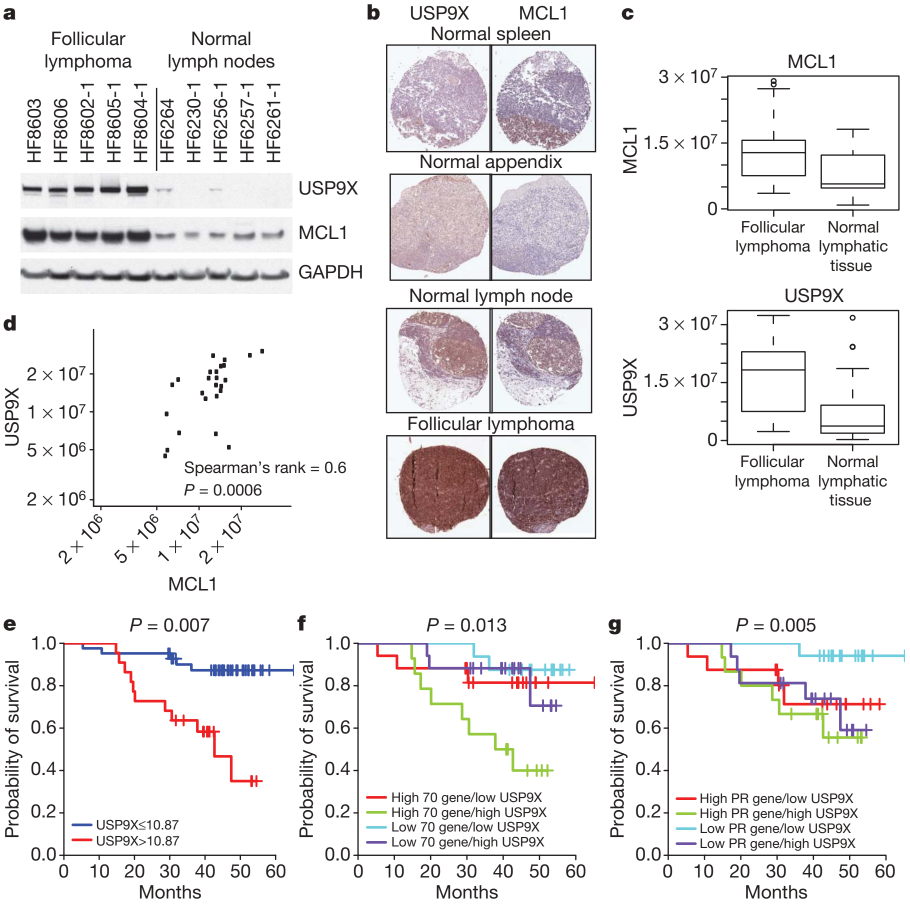


Figure 2 | USP9X overexpression in tumours correlates with increased MCL1 protein expression and poor prognosis. a, Western blots of USP9X and MCL1 in human follicular lymphomas and normal lymph nodes. **b,** Representative immunohistochemical staining for USP9X and MCL1 in human follicular lymphomas ($n = 25$) and normal lymphoid tissues ($n = 13$). Original magnification, ×100. **c,** Box-and-whisker plots of the staining in **b** quantified by morphometry. Boxes represent the upper and lower quartiles and median; whiskers show the lowest data point within 1.5 interquartile ranges (IQR) from the lower quartile, and the highest data point within 1.5 IQR from the upper quartile. **d,** Paired MCL1 and USP9X staining intensities of follicular lymphoma samples in **b**. Spearman's rank correlation coefficient = 0.6; $P = 0.0006$. The quantification data in **c** and **d** are expressed as the integrated DAB (3,3′-diaminobenzidine) intensity—that is, the sum of the light intensity of all pixels identified as 'brown' using an RGB threshold specific to DAB staining. **e–g,** Multiple myeloma patient survival on the basis of *USP9X* mRNA expression alone (**e**), or in combination with either the 70-gene high-risk predictor (**f**) or the PR (proliferation) signature (**g**). Using the optimal cutoff for USP9X of 10.87, the hazard ratio in **e** is 5.53 ($P = 0.0017$). When Cox regression model fitting the *USP9X* mRNA expression as a continuous variable, $P = 0.007$. In **f** and **g**, median gene signatures and USP9X levels were used to subdivide the population.

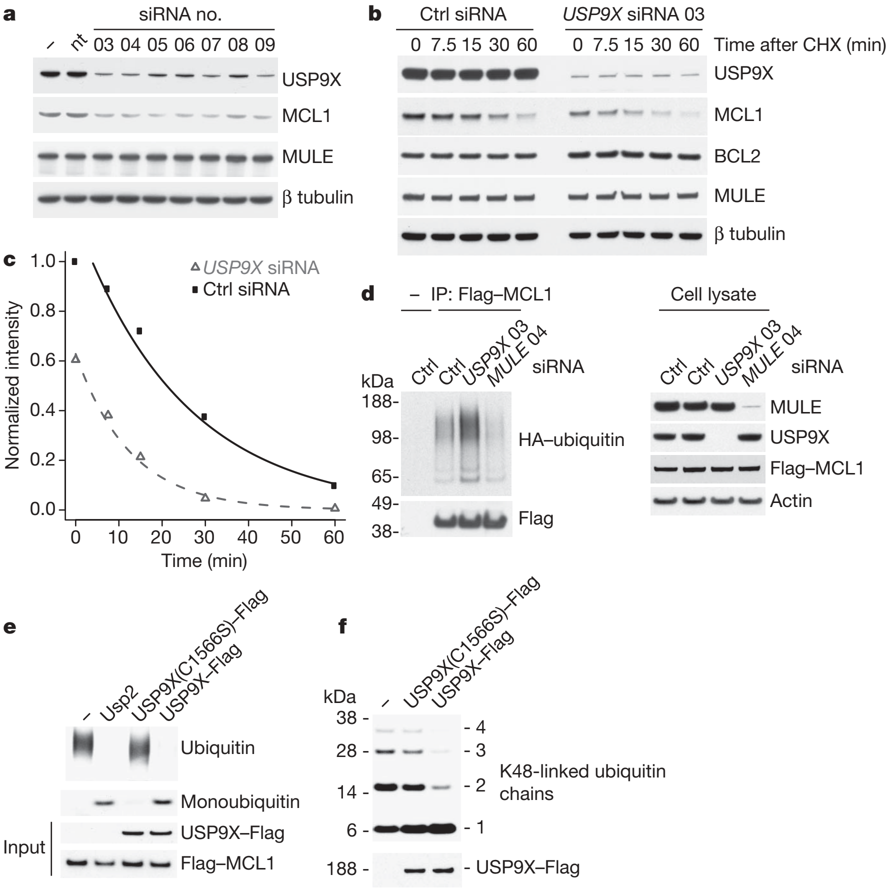


Figure 3 | USP9X deubiquitinates MCL1 and regulates its degradation. a, Knockdown of USP9X in HEK293T cells with seven different siRNAs (–, no siRNA; nt, non-targeting siRNA). **b,** HeLa cells were cultured in 10 μg ml^{-1} cycloheximide (CHX) at 72 h after siRNA knockdown of USP9X. Ctrl, control. **c,** Quantification of MCL1 in **b** by densitometry. **d,** HEK293T cells stably expressing 3×Flag–MCL1 were transfected with the siRNA indicated and haemagglutinin (HA)–ubiquitin. MCL1 was immunoprecipitated and blotted for HA–ubiquitin. **e,** Deubiquitination of MCL1 by USP9X (or the rat Usp2 catalytic subunit) *in vitro*. USP9X–Flag, USP9X(Cys1566Ser)–Flag and 3×Flag–MCL1 were purified individually from HEK293T cells. **f,** *In vitro* cleavage of Lys 48-linked polyubiquitin chains by purified USP9X or USP9X(Cys1566Ser).

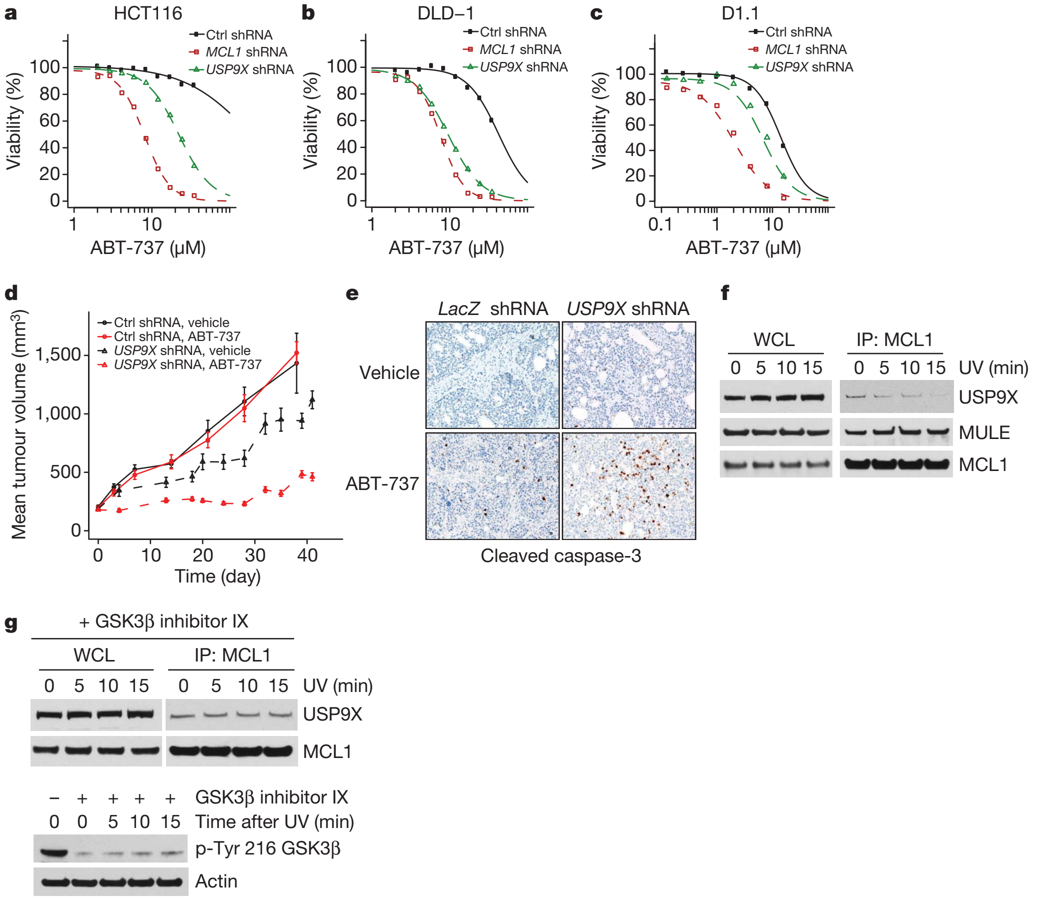


Figure 4 | DNA damage negates USP9X inhibition of MCL1-regulated cell death. a–c, HCT116 (**a**), DLD-1 (**b**) and D1.1 (**c**) clonal cell populations expressing shRNAs against human *USP9X* or *MCL1* were treated with ABT-737 for 48 h. **d**, Growth of BxPC-3 xenograft tumours expressing *LacZ* or *USP9X* shRNAs and treated with or without daily ABT-737 for 21 days. Error bars denote s.e.m. ($n = 10$). **e**, Immunohistochemical staining for cleaved caspase-3 in the BxPC-3 xenograft tumours. Original magnification, ×40. **f**, USP9X–MCL1 dissociation in HEK293T cells after ultraviolet (UV) irradiation (200 J m^{-2}). **g**, HEK293T cells pretreated with or without 100 nM GSK3β inhibitor IX for 1 h were ultraviolet irradiated (200 J m^{-2}).

doxycycline-inducible *USP9X* or *LacZ* short hairpin RNA (shRNA) were introduced into severe combined immunodeficient (SCID) beige mice. Tumours expressing the *USP9X* shRNA contained markedly less USP9X and MCL1 than tumours expressing the *LacZ* shRNA after 5 days of doxycycline treatment (Supplementary Fig. 12). Although USP9X knockdown alone caused a modest decrease in tumour growth, it delayed tumour growth significantly in combination with ABT-737 (Fig. 4d). USP9X knockdown did not affect cell proliferation *in vitro* (data not shown), so the reduction in tumour growth, amplified by exposure to ABT-737, was probably due to increased apoptosis as measured by caspase-3 cleavage (Fig. 4e, $P = 0.04$). These findings are consistent with USP9X promoting cell survival, at least in part, through the stabilization of MCL1. USP9X knockdown also sensitized HeLa cells to killing by docetaxel and etoposide, but not by TNF-α or the agonistic Fas antibody CH11 (Supplementary Fig. 13). These data demonstrate that USP9X does not affect all apoptosis signalling pathways, but specifically modulates those regulated by MCL1.

Cellular MCL1 levels decrease after DNA damage[5,25] or growth factor deprivation[5,26,27]. Ultraviolet-irradiated HEK293T cells contained significantly less MCL1 within 2 and 3 h (Supplementary Fig. 14a). We proposed that MCL1 was freed from USP9X, then ubiquitinated and degraded. Indeed, no interaction between endogenous USP9X and MCL1 was detected 15 min after ultraviolet exposure (Fig. 4f and Supplementary Fig. 14b), even though USP9X and MCL1 levels were not altered at this time, and MCL1 bearing Lys 48-linked polyubiquitin was increased (Supplementary Fig. 14c, d). Irradiation did not alter binding of MCL1 to pro-apoptotic BAK (Supplementary Fig. 14b). Pretreatment of cells with GSK3β inhibitor IX blocked ultraviolet-induced dissociation of USP9X from MCL1 (Fig. 4g), suggesting that phosphorylation of MCL1 by GSK3β after ultraviolet exposure is critical for negating the stabilizing effect of USP9X on MCL1.

MCL1 overexpression is associated with poor prognostic outcome in multiple myeloma[9], breast cancer[22], relapsing acute myeloid leukaemia and acute lymphocytic leukaemia[28], so the mechanisms that increase MCL1 levels are of paramount importance. We show that USP9X is a DUB for MCL1 that promotes cell survival by removing Lys 48-linked polyubiquitin chains that would otherwise target MCL1 for proteasomal degradation. Hence, tumours overexpressing USP9X are expected to have a survival advantage because MCL1 is stabilized. Inhibition of USP9X to decrease MCL1 potentiates ABT-737, and probably a compound in the same class undergoing evaluation in clinical trials, ABT-263 (http://clinicaltrials.gov/). Conversely, increasing and maintaining MCL1 levels by promoting USP9X-mediated MCL1 stabilization may have promise in the treatment of degenerative conditions in which there is excessive apoptosis.

METHODS SUMMARY

Mass spectrometry. 3×Flag–MCL1 stably expressed in HEK293T cells was immunoprecipitated with anti-Flag M2 affinity gel and eluted with 500 μg ml^{-1} 3×Flag peptide (Sigma). Eluted proteins were identified with a gel-based liquid chromatography–tandem mass spectrometry (Gel-LC–MS/MS) approach and ion trap mass spectrometry (LTQ; Thermo Electron). A Mascot database search and the Scaffold program (Proteome Software) were used to visualize and validate results.

***In vitro* deubiquitination.** 3×Flag–MCL1 or USP9X–Flag in transfected HEK293T cells was purified with anti-Flag M2 affinity gel in RIPA buffer. High salt (20 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 420 mM NaCl, 0.2 mM EDTA, 25% glycerol) and low salt (20 mM Tris, pH 7.4, 300 mM NaCl, 0.2 mM EDTA, 20% glycerol, 0.1% NP-40) washes preceded elution with 500 μg ml^{-1} 3×Flag peptide. MCL1 and USP9X were combined in 50 mM HEPES, pH 7.5, 10 mM 2-mercaptoethanol and 0.5 mM EDTA at 30 °C for 30 min. USP9X was incubated with 40 ng Lys 48- or Lys 63-linked polyubiquitin (Boston Biochem).

ABT-737 sensitivity. Viability of clonal cell populations treated for 48 h with ABT-737 (Abbott Laboratories) was determined with CellTiter-Glo (Promega). Dose response curves were fitted with a three-parameter Hill model using the R package 'drc' (http://www.jstatsoft.org/v12/i05).

Xenografts. Eight-week-old female C.B-17 SCID.bg mice (Charles Rivers Laboratories) were injected subcutaneously in the right hind flank with 1×10^7 BxPC-3 cells in 100 μl of HBSS–matrigel. Seven days later, mice were given drinking water containing 1 mg ml^{-1} doxycycline and 5% sucrose. When tumours achieved a mean volume of 200 mm^3, mice were randomly assigned for intraperitoneal injections of either ABT-737 (100 mg kg^{-1}; formulated in 30% polypropylene glycol, 5% Tween 80, and 65% D5W (5% dextrose in water, pH 4–5)) or vehicle control, daily for 21 days. Mice were weighed and tumours were measured for 40 days after treatment initiation.

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

Received 17 December 2008; accepted 5 November 2009.
Published online 20 December 2009.

1. Opferman, J. T. *et al.* Development and maintenance of B and T lymphocytes requires antiapoptotic MCL-1. *Nature* **426,** 671–676 (2003).
2. Opferman, J. T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Science* **307,** 1101–1104 (2005).
3. Zhao, Y. *et al.* Glycogen synthase kinase 3α and 3β mediate a glucose-sensitive antiapoptotic signaling pathway to stabilize Mcl-1. *Mol. Cell. Biol.* **27,** 4328–4339 (2007).
4. Zhong, Q. *et al.* Mule/ARF-BP1, a BH3-only E3 ubiquitin ligase, catalyzes the polyubiquitination of Mcl-1 and regulates apoptosis. *Cell* **121,** 1085–1095 (2005).
5. Nijhawan, D. *et al.* Elimination of Mcl-1 is required for the initiation of apoptosis following ultraviolet irradiation. *Genes Dev.* **17,** 1475–1486 (2003).
6. van Delft, M. F. *et al.* The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. *Cancer Cell* **10,** 389–399 (2006).
7. Kitada, S. *et al.* Expression of apoptosis-regulating proteins in chronic lymphocytic leukemia: correlations with *in vitro* and *in vivo* chemoresponses. *Blood* **91,** 3379–3389 (1998).
8. Warr, M. R. & Shore, G. C. Unique biology of Mcl-1: therapeutic opportunities in cancer. *Curr. Mol. Med.* **8,** 138–147 (2008).
9. Wuillème-Toumi, S. *et al.* Mcl-1 is overexpressed in multiple myeloma and associated with relapse and shorter survival. *Leukemia* **19,** 1248–1252 (2005).
10. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
11. Wood, S. A. *et al.* Cloning and expression analysis of a novel mouse gene with sequence similarity to the *Drosophila* fat facets gene. *Mech. Dev.* **63,** 29–38 (1997).
12. Oda, E. *et al.* Noxa, a BH3-only member of the Bcl-2 family and candidate mediator of p53-induced apoptosis. *Science* **288,** 1053–1058 (2000).
13. Wang, J. M., Lai, M. Z. & Yang-Yen, H. F. Interleukin-3 stimulation of *mcl-1* gene transcription involves activation of the PU.1 transcription factor through a p38 mitogen-activated protein kinase-dependent pathway. *Mol. Cell. Biol.* **23,** 1896–1909 (2003).
14. Hummel, M. *et al.* A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* **354,** 2419–2430 (2006).
15. Carrasco, D. R. *et al.* High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell* **9,** 313–325 (2006).
16. Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108,** 2020–2028 (2006).
17. Mouchantaf, R. *et al.* The ubiquitin ligase itch is auto-ubiquitylated *in vivo* and *in vitro* but is protected from degradation by interacting with the deubiquitylating enzyme FAM/USP9X. *J. Biol. Chem.* **281,** 38738–38747 (2006).
18. Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67,** 425–479 (1998).
19. Ding, Q. *et al.* Down-regulation of myeloid cell leukemia-1 through inhibiting Erk/Pin 1 pathway by sorafenib facilitates chemosensitization in breast cancer. *Cancer Res.* **68,** 6109–6117 (2008).
20. Domina, A. M. *et al.* MCL1 is phosphorylated in the PEST region and stabilized upon ERK activation in viable cells, and at additional sites with cytotoxic okadaic acid or taxol. *Oncogene* **23,** 5301–5315 (2004).
21. Maurer, U. *et al.* Glycogen synthase kinase-3 regulates mitochondrial outer membrane permeabilization and apoptosis by destabilization of MCL-1. *Mol. Cell* **21,** 749–760 (2006).
22. Ding, Q. *et al.* Myeloid cell leukemia-1 inversely correlates with glycogen synthase kinase-3β activity and associates with poor prognosis in human breast cancer. *Cancer Res.* **67,** 4564–4571 (2007).
23. Oltersdorf, T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **435,** 677–681 (2005).
24. Tahir, S. K. *et al.* Influence of Bcl-2 family members on the cellular response of small-cell lung cancer cell lines to ABT-737. *Cancer Res.* **67,** 1176–1183 (2007).
25. Cuconati, A. *et al.* DNA damage response and MCL-1 destruction initiate apoptosis in adenovirus-infected cells. *Genes Dev.* **17,** 2922–2932 (2003).
26. Jourdan, M. *et al.* A major role for Mcl-1 antiapoptotic protein in the IL-6-induced survival of human myeloma cells. *Oncogene* **22,** 2950–2959 (2003).
27. Austin, M. & Cook, S. J. Increased expression of Mcl-1 is required for protection against serum starvation in phosphatase and tensin homologue on chromosome 10 null mouse embryonic fibroblasts, but repression of Bim is favored in human glioblastomas. *J. Biol. Chem.* **280,** 33280–33288 (2005).
28. Kaufmann, S. H. *et al.* Elevated expression of the apoptotic regulator Mcl-1 at the time of leukemic relapse. *Blood* **91,** 991–1000 (1998).

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

Acknowledgements We thank J. Cupp, W. Tombo, B. Yang, L. Gilmour, J.-A. Hongo, R. Vij, C. Quan, M. Vasser, P. Ng, W. Sandoval and T. Huang for technical assistance; B. Bolon for immunohistochemical scoring; S. Johnson for patient data acquisition; and K. Newton for manuscript preparation.

Author Contributions V.M.D. directed the study; M.S. and X.H., with assistance from K.O. and F.B., conducted all biochemical experiments; J.R.L. ran the mass spectrometry; J.L., P.Y. and D.D. performed bioinformatics analyses; R.F., J.E.-A. and D.M.F. generated the immunohistochemical data; H.M. ran the xenograft study; D.C.S.H. contributed to experiment design.

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. Correspondence and requests for materials should be addressed to V.M.D. (dixit@gene.com).

METHODS

Plasmids, siRNAs and shRNAs. Human *MCL1*, *MCL1-ΔN* (encoding residues 169–350), *MCL1 2A* (Thr92Ala/Thr1633Ala), *MCL1 2D* (Thr92Asp/Thr163Asp) and *MCL1 3D* (Ser155Asp/Ser159Asp/Thr163Asp) were cloned into p3×Flag-CMV-7.1 (Sigma). *MCL1 3A* (Ser155Ala/Ser159Ala/Thr163Ala) was cloned into pEF FlagA PGKpuro[29]. *USP9X* with a C-terminal Flag-tag was cloned into pRK. p3×Flag-CMV-7-BAP was from Sigma. *FLIPS* was cloned into pFlag-CMV-5a (Sigma). GIPZ lentiviral shRNAmir-GFP constructs were from Open Biosystems (*MCL1* shRNA 72721, targeting sequence: 5′-GAAATTCTTTCACTTCATT-3′; *USP9X* shRNA 41519, targeting sequence: 5′-CAATCAAGTTCAATGATTA-3′; control shRNA 33305, targeting sequence: 5′-CTGTGCCATCAATATCTTA-3′). Plasmids were transfected with Lipofectamine 2000 (Invitrogen), FuGENE 6 (Roche), or Geneporter 2 (Genlantis). All siRNAs were used at 25 nM and transfected with Lipofectamine 2000. siRNAs were obtained from Dharmacon (siGENOME and siON-TARGETplus), Qiagen, or synthesized at Genentech. See Supplementary Table 1 for all sequences.

Cell culture. HEK293T cells and HeLa cells were grown in DMEM plus 10% serum. HCT116, DLD-1 and D1.1 cells were grown in RPMI-1640 plus 10% serum. U2OS cells were maintained in McCoy's 5a plus 10% serum. Primary human B cells were obtained from donor buffy coats. Peripheral blood mononuclear cells were enriched with Ficoll-Plaque PLUS, and the B cells were purified with a human B-cell isolation kit (Miltenyi) and XS columns (Miltenyi). B cells were cultured at 1.3 × 10^6 cells per ml in RPMI-1640 medium, 10% FCS, 2 mM L-glutamine, 10 mM HEPES, pH 7.2 and 10 μg ml^{-1} anti-IgM antibody (Jackson ImmunoResearch, 109-006-129).

Lentivirus infections. HEK293T cells in a 10-cm dish were tranfected with 5 μg of shRNAmir-GFP construct, 5 μg of plasmid Δ8.9, and 3 μg of plasmid VSVG. Cells were incubated at 34 °C and the medium was replaced after 12 h. Virus-containing medium was collected 36 h after transfection, and supplemented with 8 μg ml^{-1} polybrene to spin-infect target cells in 6-well dishes. After centrifugation at 630*g* for 45 min, the cells were cultured at 37 °C. The medium was replaced after 3 h. GFP$^+$ cells were sorted by flow cytometry and single cell colonies were grown.

Antibodies and reagents. CH11 anti-Fas agonistic antibody was from Millipore. Flag-tagged proteins were detected with rabbit anti-Flag polyclonal antibody (Sigma). MCL1 was immunoprecipitated with rabbit anti-MCL1 antibody (sc-819, Santa Cruz). Western blots were performed with antibodies recognizing MCL1 (sc-819, Santa Cruz or AAM-241, Stressgen), MULE (BP300-486a, Bethyl), BCL2 (OP60, Calbiochem), BCL-x$_L$ (clone 2H12, BD Biosciences), BAK (G317-2, Pharmingen), BAX (clone 3, Pharmingen), ITCH (clone 32, BD Biosciences), p63 (clone 4A4, BD Biosciences), p73 (clone GC-15, BD Biosciences), NOXA (clone 114C307.1, Alexis), GSK3β (pY216) (612313, BD Biosciences), β-tubulin (T6199, Sigma), cytochrome C (556433, BD Biosciences), GAPDH (CSA-335, Stressgen), phosphorylated (Ser 15) p53 (9284, Cell Signaling), lamin C (ab3702, Chemicon), MKI67 (610968, BD Biosciences), ubiquitin (U5379, Sigma for Fig. 3e, f, and clone P4D, Santa Cruz for Supplementary Fig. 14c), Lys 48-linked or Lys 63-linked polyubiquitin (Genentech)[30], and USP9X (clone 4B3.1.1, Genentech or h00008239-m01, Novus). Rat monoclonal 4B3.1.1 was raised against human USP9X residues 2373–2570. It detected a single band of the appropriate molecular mass (292 kDa) in HEK293T cells by western blotting (Supplementary Fig. 3d). This band was reduced by *USP9X* siRNAs (Fig. 3a). Western blot films were quantified with ImageJ software (http://rsb.info.nih.gov/ij).

Other reagents: ABT-737, (R)-4-(dimethylamino-1-phenylsulphanylmethyl-propylamino)-N-{4-[4-(4′-chloro-biphenyl-2-ylmethyl)-piperazin-1-yl]-benzoyl}-3 nitro-benzenesulfonamide) (Abbott Laboratories); GSK3 inhibitor IX (EMD Biosciences); ERK inhibitor PD0325901 (Stemgent small molecules); PI3K inhibitor LY294002 (Cell Signaling Technology); EGF and TNF (R&D Systems); MG-132 and cycloheximide (Calbiochem).

Immunoprecipitations. Cell lysates were cleared by centrifugation at 90,000*g* for 30 min. Protein A or anti-Flag M2 beads were blocked with 100 mg ml^{-1} BSA and then incubated for 1 h with soluble lysate. Beads were washed eight times with lysis buffer containing 0.1% digitonin, and eluted in SDS–PAGE sample buffer with boiling. To assess ubiquitination of MCL1, HEK293T cells stably expressing 3×Flag–MCL1 were transfected with pRK5-HA-ubiquitin. Cells were lysed in 1% SDS, 25 mM Tris-HCl, pH 7.5, boiled for 30 min, and then diluted with 10 volumes of immunoprecipitation buffer (lysis buffer with 2% Triton X-100). The soluble lysate was precleared with protein G beads and then 3×Flag–MCL1 was immunoprecipitated with anti-Flag M2 beads. The beads were washed five times with immunoprecipitation buffer, followed by five washes with immunoprecipitation buffer and 0.5 M NaCl. 3×Flag–MCL1 was eluted with 500 μg ml^{-1} 3×Flag peptide. Immunoprecipitations with human anti-Lys 48-linked polyubiquitin antibody were performed on denatured whole cell lysates and protein G beads.

***In vitro* binding assays.** Recombinant Myc–MCL1 and USP9X–Flag were generated with a TNT T7/SP6-coupled wheat-germ extract system (Promega). The recombinant proteins were mixed in NP40 buffer (1% NP-40, 120 mM NaCl, 50 mM Tris-HCl, pH 7.4, 1 mM EDTA, pH 7.4) and immunoprecipitated with either anti-Flag M2 or anti-Myc beads (Sigma).

Cellular fractionation. U2OS, Colo201 and Colo205 cells were fractionated with the Qproteome Cell Compartment Kit (Qiagen). The fractionated lysates were quantified using a BCA Protein Assay Kit (Thermo Scientific), and equivalent amounts of fractionated cell lysates were analysed by immunoblotting.

Immunofluorescence microscopy. U2OS cells were incubated for 30 min at 37 °C with Mito Tracker Deep Red 633 (Invitrogen) to stain mitochondria, washed with PBS, and fixed for 20 min in 4% paraformaldehyde. After two PBS washes, quenching was performed for 10 min with 50 mM NH$_4$Cl in PBS. Cells were then permeabilized with 0.1% Triton X-100 for 5 min and stained with 10 μg ml^{-1} rat anti-USP9X antibody (clone 4B3.1.1) or a 1:20 dilution of anti-MCL1 antibody (BD Biosciences) in 10% goat serum for 30 min. After five PBS washes, the cells were incubated with Alexa Fluor 488-conjugated secondary antibodies (Invitrogen) for 30 min. Cells were washed five times with PBS and then mounted in Pro-long Gold with 4′,6-diamidino-2-phenylindole (DAPI, Invitrogen).

Immunohistochemistry. Formalin-fixed paraffin-embedded 3-μm tissue sections were deparaffinized in xylenes and rehydrated through a graded series of alcohols. Antigen retrieval was in Trilogy (Cell Marque) for mouse anti-USP9X antibody (clone 1C4, Novus Biologicals) or Target Retrieval (DAKO) for rabbit anti-MCL1 antibody (Cell Signaling). Blocking was performed with KPL blocking solution (Kirkegaard and Perry Laboratories), an avidin/biotin blocking kit (Vector Labs), and then either 10% horse serum/3%BSA/PBS or 10% goat serum/3% BSA/PBS. Sections were stained with 10 μg ml^{-1} USP9X or 2 μg ml^{-1} MCL1 antibody for 60 min at room temperature, followed by a biotinylated horse-anti-mouse or goat-anti-rabbit secondary antibody, an ABC-HRP Elite reagent (Vector Labs), and a metal enhanced DAB colourimetric peroxidase substrate (Pierce Laboratories). Sections were counterstained with Myer's Hematoxylin (Rowley Biochemical Institute).

For quantification, stained sections were independently scored by two pathologists using a 0–4+ scale. Furthermore, images were acquired by the Ariol SL-50 automated slide-scanning platform (Genetix Ltd) at ×100 final magnification. A range of pixel colours that corresponded to regions of positive staining was defined with the Metamorph software package (MDS Analytical Technologies). The brown DAB-specific staining was isolated from the haematoxylin counterstain using a blue-normalization algorithm as described previously[31]. The average pixel intensity inside these regions was multiplied by the area to generate a staining index for each core, which is indicative of how much total staining is present. For cores of less uniform size, the average pixel intensity was used as a parameter to determine protein levels. Differences in protein expression between human follicular lymphoma and normal lymphatic tissues were assessed by Mann–Whitney–Wilcoxon test. The correlation of protein expression between MCL1 and USP9X was assessed by Spearman's rank correlation.

mRNA and protein correlation with patient outcome. Two publicly available mRNA expression data sets (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) were used to investigate whether *USP9X* mRNA levels are a marker for patient prognosis. The data set GSE4452 was used for analysis in multiple myeloma (MM). Log2 was taken from the pre-processed MAS5 values in GSE4452. The association of *USP9X* mRNA expression with prognostic outcome was tested in two different ways: by treating the expression value (1) as a continuous variable, or (2) as a categorical variable. For the latter, the expression values of USP9X were first dichotomized to find the optimal cutoffs using the log rank test. In GSE4452, the cutoff between the two groups was defined at 10.87 (65.8% quantile). The association of USP9X protein level and prognostic outcome was tested by dividing the patient samples into two groups according to their score ($\leq$2, >2). Survival analysis was then done using Kaplan–Meier curves and the Cox proportional hazards model. Samples were cultured before treatment, and the small hatch marks on the Kaplan–Meier graph represent patients that left the trial.

Patient samples. Patient samples for immunohistochemistry and western analysis are part of a Genentech patient sample collection. The DLBCL tissues evaluated by immunohistochemistry are from patients that were not treated uniformly. The lymphoma TMA was constructed externally (Cybrdi, Inc.).

29. Huang, D. C., Cory, S. & Strasser, A. Bcl-2, Bcl-XL and adenovirus protein E1B19kD are functionally equivalent in their ability to inhibit cell death. *Oncogene* **14**, 405–414 (1997).
30. Newton, K. *et al.* Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. *Cell* **134**, 668–678 (2008).
31. Brey, E. M. *et al.* Automated selection of DAB-labeled tissue for immunohistochemical quantification. *J. Histochem. Cytochem.* **51**, 575–584 (2003).

LETTERS

Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor

Michael P. Bokoch[1], Yaozhong Zou[1], Søren G. F. Rasmussen[1], Corey W. Liu[2], Rie Nygaard[1], Daniel M. Rosenbaum[1], Juan José Fung[1], Hee-Jung Choi[1,3], Foon Sun Thian[1], Tong Sun Kobilka[1], Joseph D. Puglisi[2,3], William I. Weis[1,3], Leonardo Pardo[4], R. Scott Prosser[5], Luciano Mueller[6] & Brian K. Kobilka[1]

G-protein-coupled receptors (GPCRs) are seven-transmembrane proteins that mediate most cellular responses to hormones and neurotransmitters. They are the largest group of therapeutic targets for a broad spectrum of diseases. Recent crystal structures of GPCRs[1–5] have revealed structural conservation extending from the orthosteric ligand-binding site in the transmembrane core to the cytoplasmic G-protein-coupling domains. In contrast, the extracellular surface (ECS) of GPCRs is remarkably diverse and is therefore an ideal target for the discovery of subtype-selective drugs. However, little is known about the functional role of the ECS in receptor activation, or about conformational coupling of this surface to the native ligand-binding pocket. Here we use NMR spectroscopy to investigate ligand-specific conformational changes around a central structural feature in the ECS of the β_2 adrenergic receptor: a salt bridge linking extracellular loops 2 and 3. Small-molecule drugs that bind within the transmembrane core and exhibit different efficacies towards G-protein activation (agonist, neutral antagonist and inverse agonist) also stabilize distinct conformations of the ECS. We thereby demonstrate conformational coupling between the ECS and the orthosteric binding site, showing that drugs targeting this diverse surface could function as allosteric modulators with high subtype selectivity. Moreover, these studies provide a new insight into the dynamic behaviour of GPCRs not addressable by static, inactive-state crystal structures.

In the ligand-free basal state, GPCRs exist in an equilibrium of conformations[6]. Ligand binding modulates receptor function by stabilizing different intramolecular interactions and establishing a new conformational equilibrium. Activating ligands (agonists) stabilize receptor conformations that increase signalling through G proteins; inhibiting ligands (inverse agonists) stabilize other conformations that decrease the basal, agonist-independent level of signalling (Supplementary Fig. 1). When a GPCR is activated, structural changes occur in the cytoplasmic G-protein-coupling domains. These changes have been characterized for several receptors, including rhodopsin[7–10] and the β_2 adrenergic receptor (β_2AR)[11–13]. Recent solid-state NMR data show that light activation of rhodopsin also induces conformational changes in extracellular loop (ECL) 2 (ref. 14). In rhodopsin, ECL2 forms a structured cap over the covalently bound ligand retinal and interacts with transmembrane (TM) segments involved in activation. However, little is known about the effects of diffusible ligand binding on the extracellular domains of other family A GPCRs, in which ECL2 is displaced away from the ligand-binding pocket. Here we show that ligands known to affect cytoplasmic domain conformation differentially also stabilize distinct ECS conformations (Supplementary Fig. 1).

Understanding conformational changes in the ECS of GPCRs may provide new avenues for drug design. Comparison of the crystallographically identified orthosteric binding pockets of β_2AR and β_1AR reveals that 15 of 16 amino acids (94%) are identical[1,5]. This observation underscores the challenge of identifying subtype-selective drugs for families containing several closely related receptors (for example adrenergic, serotonin or dopamine receptors)[15]. In contrast, although the backbone structure of the β_2AR and β_1AR extracellular domains are similar, 22 of 39 residues (56%) in ECLs 2 and 3 differ. The ECS therefore provides a diverse site for the development of subtype-selective drugs.

Most of the β_2AR ECS consists of ECL2, connecting TMs 4 and 5, and ECL3, connecting TMs 6 and 7 (Fig. 1a)[1,2]. ECL2 forms a two-turn α-helix that is displaced away from the ligand-binding site entrance (Fig. 1b). Two disulphide bonds stabilize ECL2, one within the loop and one to the end of TM3. A salt bridge formed by Lys 305$^{7.32}$ and Asp 192$^{\mathrm{ECL2}}$ connects ECL3–TM7 to ECL2 (superscripts in this form indicate Ballesteros–Weinstein numbering for conserved GPCR residues)[16]. Carazolol is an inverse agonist that binds in the orthosteric pocket formed by TMs 3, 5, 6 and 7. The only direct interaction between the ECS and carazolol is through an aromatic interaction with Phe 193$^{\mathrm{ECL2}}$. Given these specific associations between ECLs, the orthosteric ligand-binding site and TMs involved in activation[17], we speculated that β_2AR extracellular domains and the associated salt bridge rearrange on activation.

To monitor the environment around the Lys 305–Asp 192 salt bridge by NMR, we selectively labelled lysine side chains in a modified β_2AR (β_2AR365) with carbon-13 by reductive methylation[18] (Methods, Supplementary Figs 2–4 and Supplementary Table 1):

$$\mathrm{R\text{-}\ddot{N}H_2} \rightleftharpoons \mathrm{R\text{-}\overset{+}{N}H_3} \xrightarrow[\mathrm{NaBH_3CN}]{^{13}\mathrm{CH_2O}} \mathrm{R\text{-}\ddot{N}(^{13}CH_3)_2} \rightleftharpoons \mathrm{R\text{-}\overset{+}{N}H(^{13}CH_3)_2} \qquad (1)$$

This approach exploits the sensitivity of methyl groups as NMR probes for the analysis of large protein structure and dynamics[19]. Reductive methylation adds two [^{13}C]methyl groups to the ε-NH$_2$ of lysine side chains and the α-NH$_2$ at the receptor amino terminus. The [^{13}C]dimethyllysines serve as conformational probes in two-dimensional ^{1}H–^{13}C correlation NMR experiments. Dimethylation does not alter the positive charge on the lysine residue (equation (1)) and causes little structural perturbation[20]. We observed no significant changes between the crystal structure of a methylated β_2AR–Fab complex bound to carazolol and that of the non-methylated receptor (Supplementary Fig. 5 and Supplementary Table 2). Reductively [^{13}C]methylated β_2AR ([^{13}C]methyl-β_2AR) has ligand-binding properties identical to those of unlabelled β_2AR, and G-protein coupling is unimpaired (Supplementary Fig. 6).

[1]Department of Molecular and Cellular Physiology, [2]Stanford Magnetic Resonance Laboratory, and [3]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [4]Laboratori de Medicina Computacional, Unitat de Bioestadística, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain. [5]Department of Chemistry, University of Toronto, UTM, Mississauga, Ontario, Canada L5L 1C6. [6]Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, New Jersey 08543, USA.

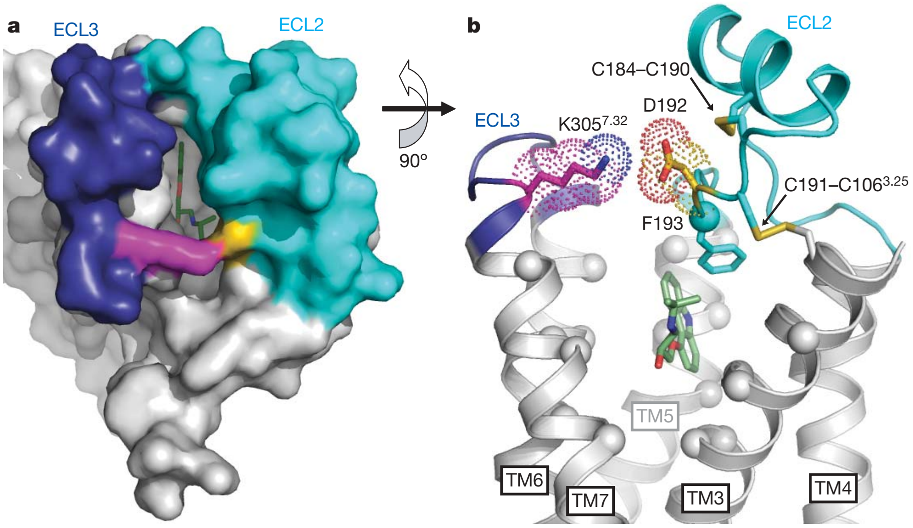


Figure 1 | Extracellular domains of carazolol-bound β_2AR. a, The extracellular surface (ECS) of β_2AR showing ECL2 (cyan, Met 171–Ala 198), ECL3 (dark blue, His 296–Glu 306), Lys 305$^{7.32}$ (magenta), Asp 192 (yellow) and the inverse agonist carazolol (green). ECL1 (Met 96–Phe 108) is part of the ECS but is not coloured. **b**, Intramolecular and ligand-binding interactions. Spheres indicate the α carbons of residues in direct contact with carazolol (at least one atom within 4 Å distance). Disulphide bonds are shown as yellow sticks. Other colours are the same as in **a**. Transmembrane helices 1 and 2 have been removed for clarity. Asp 192 and Lys 305 form the only lysine salt bridge observed in the crystal structure. The relative accessibilities of Asp 192 and Lys 305 are 35% and 75%, respectively, compared with the accessibility of that residue type in an extended Ala-x-Ala tripeptide[30].

Intense peaks from dimethylamines (dimethyllysines and the dimethyl-amino terminus) are observed in the ^{1}H–^{13}C NMR spectrum of [^{13}C]methyl-β_2AR bound to the inverse agonist carazolol in detergent buffer (Fig. 2, dimethylamine region; Supplementary Fig. 7, full spectral width). We used both heteronuclear single-quantum coherence (HSQC) and saturation transfer differencing (STD)-filtered ^{1}H-detected heteronuclear multiple-quantum coherence (HMQC) pulse sequences throughout this work; STD-filtered HMQC improved the spectral quality at the expense of longer acquisition times (Supplementary Fig. 8).

Several features of the [^{13}C]methyl-β_2AR NMR spectrum are notable (Fig. 2a). The sharpest peak is assigned to the dimethyl-amino terminus on the basis of protease digestion (Fig. 2b and Supplementary Fig. 9). The remaining peaks are assigned to dimethyllysines. The region is dominated by a cluster of overlapping peaks (centred at ^{1}H chemical shift 2.8 p.p.m.) attributed to solvent-exposed, highly mobile lysines. The intensity of this cluster is decreased by about 50% after the mutation of seven cytoplasmic lysines to arginine (Fig. 2b and Supplementary Fig. 10). Two broad dimethyllysine peaks are shifted upfield in the ^{1}H dimension (Fig. 2a, peaks 1 and 2). These peaks represent the two [^{13}C]methyl groups on Lys 305, as determined by mutation of Lys 305 to Arg (Fig. 2c and Supplementary Fig. 10). The fine structure features in peaks 1 and 2 might suggest conformational heterogeneity; however, they are most probably due to relatively low signal-to-noise ratios and subtle baseline distortions.

Lys 305 is the only lysine that forms a salt bridge in the β_2AR crystal structure, which is consistent with the unique chemical shifts of peaks 1 and 2. The presence of two peaks implies that the two methyl groups on Lys 305 exist in non-equivalent chemical environments[21–23]. The two peaks merge under conditions of increased temperature and ionic strength, presumably as a result of weakening of the salt bridge (Supplementary Fig. 11). Reduction of the disulphide bonds stabilizing ECL2 in the β_2AR abolishes the Lys 305 peaks (Supplementary Fig. 12), demonstrating that the salt bridge is sensitive to conformational changes in the ECS. Taken as a whole, these data show that the Lys 305–Asp 192 salt bridge is formed in solution as well as in crystal lattices for carazolol-bound β_2AR[2,24]. This conclusion is further supported by measurements of ^{13}C transverse relaxation (T_2) times indicating restricted motion of Lys 305 compared with other β_2AR lysines (Supplementary Fig. 13 and Supplementary Table 3).

The Lys 305 peaks are a probe for conformational changes in the receptor ECS. Both peaks are present in the NMR spectrum of unliganded β_2AR, indicating that the Lys 305–Asp 192 salt bridge also forms in the basal state (Fig. 3a). Two differences are seen relative to the carazolol-bound state (Fig. 3b). First, peaks 1 and 2 have a larger upfield shift in the ^{1}H dimension. Second, the ^{13}C chemical-shift separation between peaks 1 and 2 is increased. These data show that carazolol binding alters the chemical environment around Lys 305, demonstrating a change in the ECS. In contrast with inverse agonists, neutral antagonists (for example alprenolol) do not alter basal receptor activity. As might be predicted on the basis of ligand efficacy, NMR detects no difference in Lys 305 chemical shifts between unliganded and alprenolol-bound β_2AR (Fig. 3c).

The inverse agonist-induced conformational change (Fig. 3) probably involves Phe 193 in ECL2 (Fig. 1b), which forms a favourable edge-to-face interaction with the tricyclic aromatic ring of carazolol in the β_2AR crystal structure[1]. Nearly identical interactions are observed with other inverse agonists: Phe 193 and timolol in β_2AR[24], and the homologous Phe 201 and cyanopindolol in β_1AR[5]. By contrast, alprenolol has a single aromatic ring that cannot interact strongly with Phe 193 when docked at the carazolol position. Molecular dynamics simulations show that Phe 193 adopts the *trans* conformation pointing towards TM5 in the presence of carazolol, but it has increased mobility and is able to assume multiple conformations in the alprenolol-bound state (Supplementary Fig. 14). Phe 193 can form close encounters (less than 5 Å) with the Lys 305 amine when alprenolol is bound. The observed upfield chemical shift change of peaks 1 and 2 in alprenolol-bound receptor relative to carazolol-bound receptor is therefore most probably due to aromatic ring current effects, although we cannot exclude the possibility of other changes in ECS conformation.

Agonists induce ECS conformational changes that differ from those induced by inverse agonists. Adding formoterol, a β_2AR agonist with nanomolar affinity, attenuates the Lys 305 resonances (Fig. 4a, b). The effect is titratable (Supplementary Fig. 15) and reverses when formoterol is replaced by carazolol by dialysis (Fig. 4c). Attenuation of Lys 305 resonances was also observed with another, structurally distinct high-affinity agonist (Supplementary Fig. 16). These NMR data suggest that the Lys 305–Asp 192 salt bridge is weakened in the β_2AR active state. A loss of interaction with Asp 192 would abolish the unique chemical environment of Lys 305. On the basis of the β_2AR

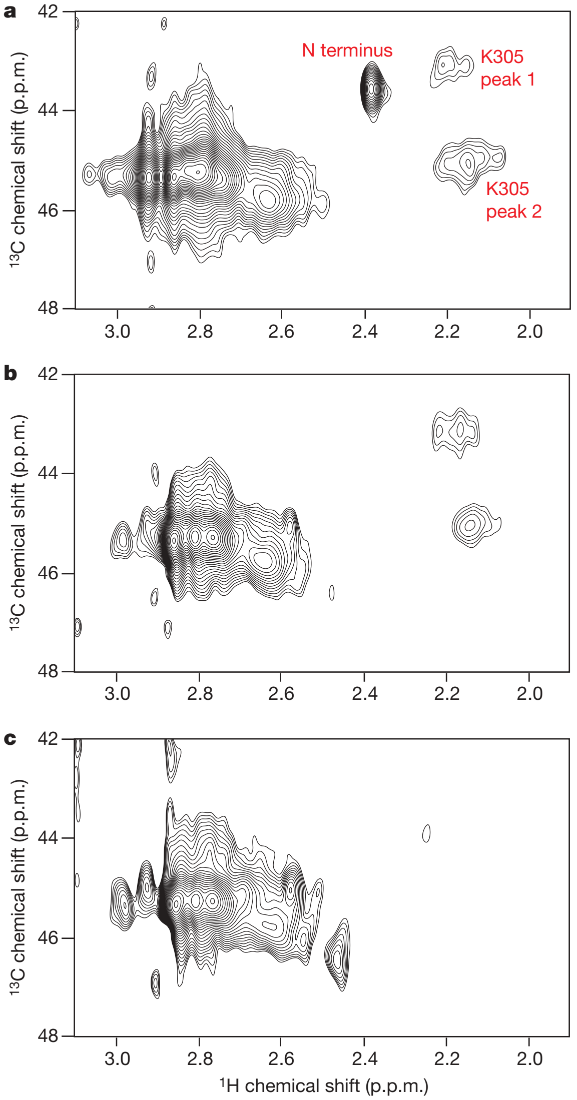


Figure 2 | Dimethyllysine NMR spectroscopy of [^{13}C]methyl-β_2AR and assignment of Lys 305. **a**, Dimethylamine region of STD-filtered ^{1}H–^{13}C HMQC spectrum of carazolol-bound β_2AR365, containing 14 Lys residues and an N-terminal Flag–TEV sequence (Supplementary Fig. 2). **b**, Spectrum of β_2AR365 with seven cytoplasmic Lys→Arg mutations (β_2AR365 Δ7Lys) and the N terminus removed by TEV proteolysis. **c**, Spectrum of β_2AR365 Δ7Lys plus the Lys 305→Arg mutation and the N terminus removed by TEV proteolysis.

crystal structure, we estimate that a distance increase of 2.9 Å between the Cα carbons of Asp 192 and Lys 305 is the minimum needed to disrupt the geometrical criteria for a salt bridge established previously[25]. At equilibrium, the fraction of Lys 305 liberated from the salt bridge would be indistinguishable from solvent-exposed lysine residues (at 2.8 p.p.m. ^{1}H chemical shift), explaining the absence of any new peaks. Alternatively, if agonists were to induce conformational fluctuations on the millisecond timescale, line broadening would attenuate the signal for Lys 305. In either case, we interpret the formoterol-induced conformational change as a relative motion between ECL3–TM7 and ECL2. Conformational changes involving ECL2 are compatible with circular dichroism experiments demonstrating agonist-induced changes in the extracellular disulphide bond linking ECL2 and TM3 in the 5-hydroxytryptamine$_{4(a)}$ receptor[26].

On the basis of our NMR results and computational modelling, we propose that the extracellular ends of TMs 6 and 7 move on activation (Fig. 4d). In brief, formoterol-activated β_2AR was modelled on the basis of the crystal structure of ligand-free opsin[9] and a relaxed conformation of the highly distorted Pro 288$^{6.50}$-induced kink (Supplementary Fig. 17).

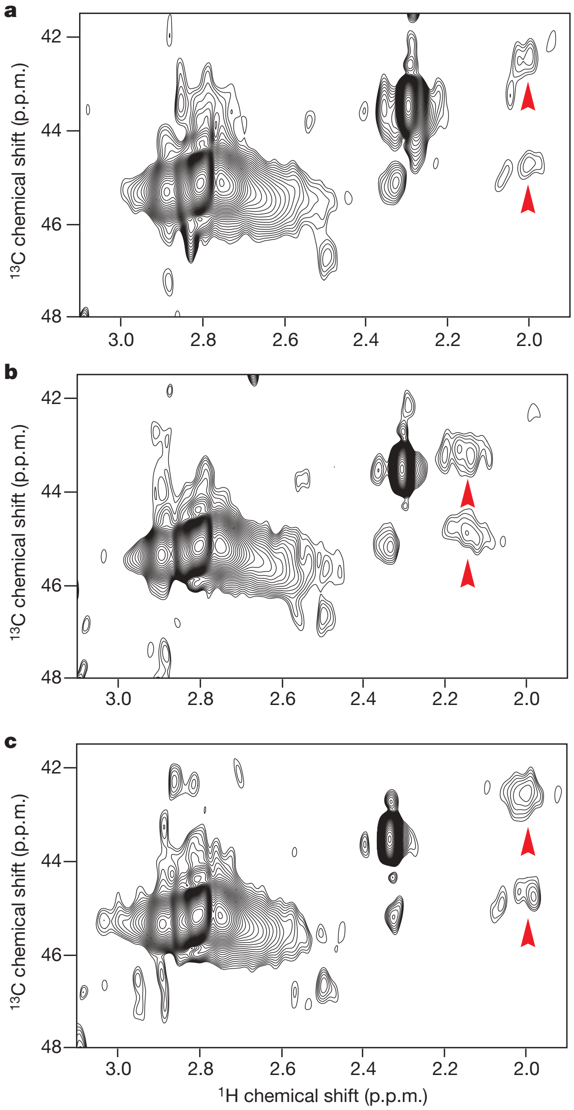


Figure 3 | Effect of inverse agonist and antagonist on the [^{13}C]dimethyl-Lys 305 NMR resonances. **a**–**c**, HSQC spectra of unliganded β_2AR (about 60 μM) (**a**), β_2AR bound to inverse agonist carazolol (**b**) and β_2AR bound to the neutral antagonist alprenolol (**c**).

In this active state model, an inward movement at the extracellular end of TM6 permits the known interaction between Asn 293$^{6.55}$ and the chiral β-hydroxyl of the agonist[27]. This motion, simultaneous with outward motion at the intracellular end of TM6 towards TM5 (ref. 9), agrees with the activation model derived from engineering GPCRs with metal-ion-binding sites[28,29]. The TM6 motion necessitates a lateral displacement of TM7 that reorients the Lys 305$^{7.32}$ salt bridge in agreement with NMR spectroscopy (Fig. 4). Inverse agonists may function in part by stabilizing bulky hydrophobic interactions with Phe 193$^{\mathrm{ECL2}}$ that block the motion of TM6.

Thus, NMR spectroscopy can be used to investigate structural changes in GPCRs, although the isotopic labelling methods employed here are limited to monitoring changes in the environment and dynamics of accessible lysine side chains. We provide direct biophysical evidence for three distinct conformations of the β_2AR extracellular surface: one for an unliganded receptor or a neutral antagonist, one for an inverse agonist, and one for an agonist (Fig. 4e). These conformations correspond to distinct functional behaviour. Unliganded and alprenolol-bound β_2AR are both able to couple to G_s, which is consistent with the basal activity of the receptor and the efficacy of alprenolol (neutral antagonist). In contrast, the inverse agonist carazolol prevents receptor–G_s coupling. Finally, agonists promote the strongest coupling. Ligands binding to the extracellular surface could therefore

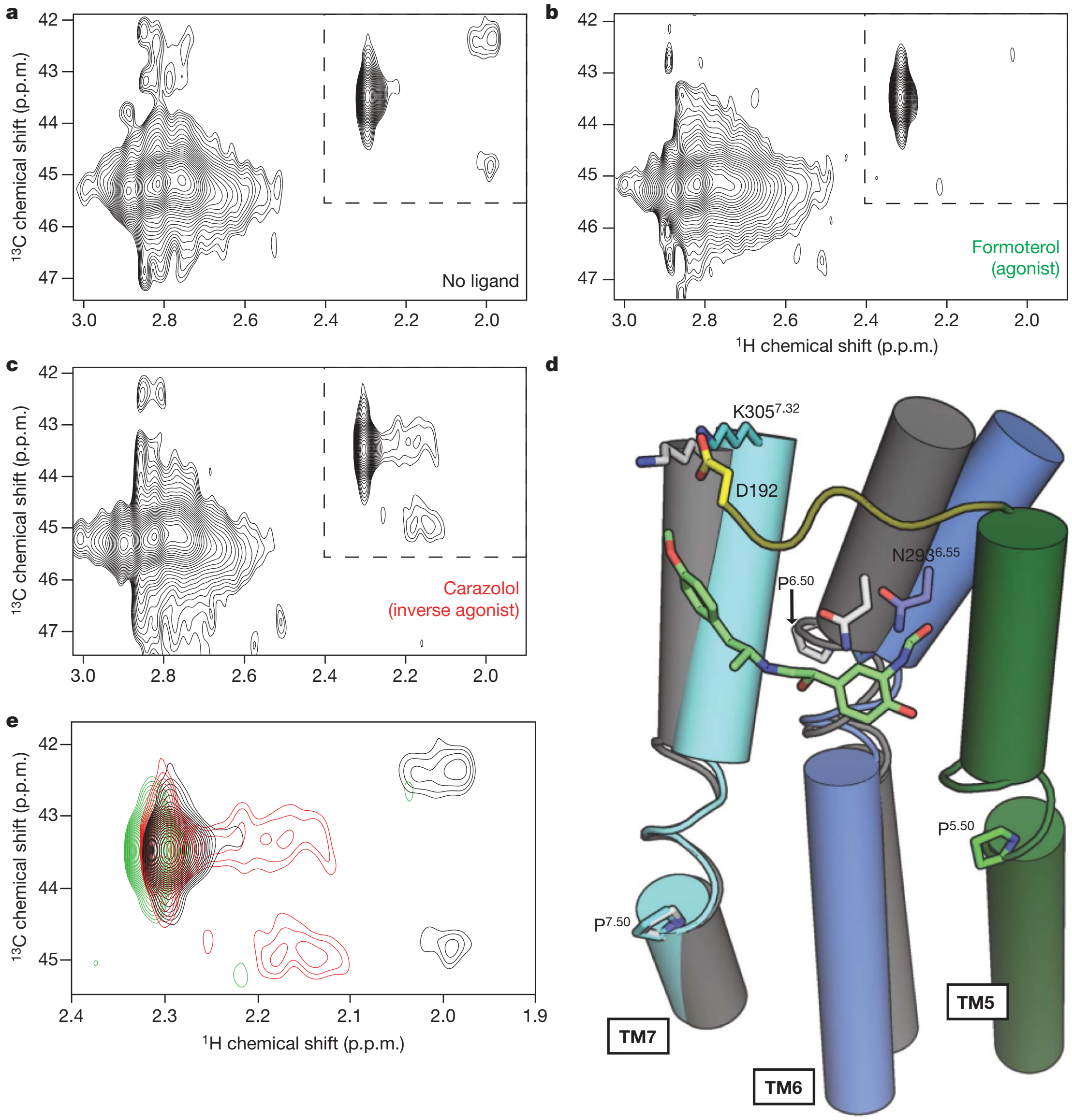


Figure 4 | Activation of β_2AR by formoterol. a–c, STD-filtered HMQC spectra of unliganded β_2AR (about 60 μM) (**a**), the same sample bound to a saturating concentration (320 μM) of agonist (*R*,*R*)-formoterol (**b**) and the same sample after replacing formoterol with the inverse agonist carazolol by dialysis (**c**). **d**, Model of β_2AR activation by formoterol (see Supplementary Fig. 16). Coloured helices, loops and side chains represent the carazolol-bound crystal structure. Grey helices and white side chains indicate the active-state model. Green sticks indicate (*R*,*R*)-formoterol and yellow indicates ECL2. **e**, Overlay of spectra corresponding to dashed regions shown in **a**–**c**. The spectrum of unliganded β_2AR from **a** is shown in black, agonist-bound β_2AR from **b** in green, and inverse agonist-bound β_2AR from **c** in red.

modulate receptor function, either by influencing the binding of orthosteric ligands or by direct allosteric modulation of cytoplasmic domain conformation (Supplementary Fig. 1). Although the specific salt bridge used to monitor these conformations may not be present in other GPCRs, it is ideally positioned to monitor ECS conformations in the β_2AR, and it is likely that our findings about ligand-induced changes in the ECS are relevant for other family A GPCRs.

METHODS SUMMARY

NMR spectroscopy of [^{13}C]methyl-β_2AR. Human β_2AR, tagged with an N-terminal Flag–tobacco etch virus (TEV) protease sequence, and truncated after residue Gly 365 (β_2AR365; Supplementary Fig. 2) was expressed in *Spodoptera frugiperda* (Sf9) insect cells with recombinant baculovirus. Sf9 cell membranes were solubilized with dodecylmaltoside and purified by sequential antibody affinity and alprenolol affinity chromatography, as described previously[3]. ^{13}C-Methyl labelling was performed by sequentially adding excess sodium cyanoborohydride followed by [^{13}C]formaldehyde to purified β_2AR. Methylation reagents were removed by extensive dialysis (unliganded β_2AR) or by anion-exchange chromatography (carazolol-bound β_2AR). For NMR spectroscopy, [^{13}C]methyl-β_2AR was dialysed against buffer containing 20 mM HEPES, pH 7.4, 100 mM NaCl and 0.1% dodecylmaltoside prepared in 98% $^{2}H_2O$ and concentrated to a final concentration of 50–200 μM. Two-dimensional ^{1}H–^{13}C correlation spectra of [^{13}C]methyl-β_2AR were recorded at 800 MHz for about 8 h (HSQC) or 24 h (STD-filtered HMQC) at 25 °C. Both pulse sequences used WATERGATE water suppression. See Supplementary Fig. 8 for all parameters and full details of NMR spectroscopy.

Crystal structure of [^{13}C]methyl-β_2AR–Fab5 complex. [^{13}C]Methyl-β_2AR–Fab5 complex was prepared and crystallized as described previously[3]. Diffraction images were obtained on a microfocus beam line, and the structure was solved by molecular replacement using β_2AR–Fab5 (PDB accession code 2R4R) as a search model.

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

Received 24 June; accepted 6 November 2009.

1. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).

2. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
3. Rasmussen, S. G. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).
4. Jaakola, V. P. *et al.* The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).
5. Warne, T. *et al.* Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–491 (2008).
6. Kobilka, B. K. & Deupi, X. Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **28,** 397–406 (2007).
7. Farrens, D. L., Altenbach, C., Yang, K., Hubbell, W. L. & Khorana, H. G. Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* **274,** 768–770 (1996).
8. Altenbach, C., Kusnetzow, A. K., Ernst, O. P., Hofmann, K. P. & Hubbell, W. L. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc. Natl Acad. Sci. USA* **105,** 7439–7444 (2008).
9. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–187 (2008).
10. Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455,** 497–502 (2008).
11. Ghanouni, P. *et al.* Functionally different agonists induce distinct conformations in the G protein coupling domain of the β_2 adrenergic receptor. *J. Biol. Chem.* **276,** 24433–24436 (2001).
12. Swaminath, G. *et al.* Probing the β_2 adrenoceptor binding site with catechol reveals differences in binding and activation by agonists and partial agonists. *J. Biol. Chem.* **280,** 22165–22171 (2005).
13. Yao, X. *et al.* Coupling ligand structure to specific conformational switches in the β_2-adrenoceptor. *Nature Chem. Biol.* **2,** 417–422 (2006).
14. Ahuja, S. *et al.* Helix movement is coupled to displacement of the second extracellular loop in rhodopsin activation. *Nature Struct. Mol. Biol.* **16,** 168–175 (2009).
15. Conn, P. J., Christopoulos, A. & Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature Rev. Drug Discov.* **8,** 41–54 (2009).
16. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G-protein coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).
17. Nygaard, R., Frimurer, T. M., Holst, B., Rosenkilde, M. M. & Schwartz, T. W. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol. Sci.* **30,** 249–259 (2009).
18. Zhang, M. & Vogel, H. J. Determination of the side chain pK_a values of the lysine residues in calmodulin. *J. Biol. Chem.* **268,** 22420–22428 (1993).
19. Tugarinov, V., Hwang, P. M., Ollerenshaw, J. E. & Kay, L. E. Cross-correlated relaxation enhanced ^{1}H–^{13}C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *J. Am. Chem. Soc.* **125,** 10420–10428 (2003).
20. Jentoft, J. E., Jentoft, N., Gerken, T. A. & Dearborn, D. G. ^{13}C NMR studies of ribonuclease A methylated with [^{13}C]formaldehyde. *J. Biol. Chem.* **254,** 4366–4370 (1979).
21. Gerken, T. A., Jentoft, J. E., Jentoft, N. & Dearborn, D. G. Intramolecular interactions of amino groups in ^{13}C reductively methylated hen egg-white lysozyme. *J. Biol. Chem.* **257,** 2894–2900 (1982).
22. Sherry, A. D. & Teherani, J. Physical studies of ^{13}C-methylated concanavalin A. pH- and Co^{2+}-induced nuclear magnetic resonance shifts. *J. Biol. Chem.* **258,** 8663–8669 (1983).
23. Abraham, S. J., Hoheisel, S. & Gaponenko, V. Detection of protein-ligand interactions by NMR using reductive methylation of lysine residues. *J. Biomol. NMR* **42,** 143–148 (2008).
24. Hanson, M. A. *et al.* A specific cholesterol binding site is established by the 2.8 Å structure of the human β_2-adrenergic receptor. *Structure* **16,** 897–905 (2008).
25. Kumar, S. & Nussinov, R. Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophys. J.* **83,** 1595–1612 (2002).
26. Baneres, J. L. *et al.* Molecular characterization of a purified 5-HT4 receptor: a structural basis for drug efficacy. *J. Biol. Chem.* **280,** 20253–20260 (2005).
27. Wieland, K., Zuurmond, H. M., Krasel, C., Ijzerman, A. P. & Lohse, M. J. Involvement of Asn-293 in stereospecific agonist recognition and in activation of the β_2-adrenergic receptor. *Proc. Natl Acad. Sci. USA* **93,** 9276–9281 (1996).
28. Elling, C. E. *et al.* Metal ion site engineering indicates a global toggle switch model for seven-transmembrane receptor activation. *J. Biol. Chem.* **281,** 17337–17346 (2006).
29. Schwartz, T. W., Frimurer, T. M., Holst, B., Rosenkilde, M. M. & Elling, C. E. Molecular mechanism of 7TM receptor activation—a global toggle switch model. *Annu. Rev. Pharmacol. Toxicol.* **46,** 481–519 (2006).
30. Hubbard, S. J., Campbell, S. F. & Thornton, J. M. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **220,** 507–530 (1991).

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

Acknowledgements We acknowledge support from National Institutes of Health, grants NS028471 (B.K.K.) and GM56169 (W.I.W.), the Stanford Medical Scientist Training Program (M.P.B.), the Lundbeck Foundation (S.G.F.R.), the University of Copenhagen and 7TM Pharma (R.N.), and the Instituto de Salud Carlos III (L.P.).

Author Contributions M.P.B. designed experiments, purified, labelled and functionally characterized β_2AR, collected and analysed NMR data and wrote the paper. Y.Z. made, expressed and purified β_2AR lysine mutants and collected NMR data. S.G.F.R. expressed and purified β_2AR for NMR and crystallized the ^{13}C-methylated β_2AR–Fab complex. C.W.L. designed, optimized and supervised NMR experiments and collected NMR data. D.M.R., H.-J.C. and W.I.W. collected diffraction data and refined the structure of the ^{13}C-methylated β_2AR–Fab complex. R.N. collected and analysed NMR data and optimized data processing. J.J.F. performed G-protein-coupling assays on labelled β_2AR. F.S.T. prepared insect cell cultures and purified β_2AR. T.S.K. purified β_2AR. J.D.P. advised on NMR spectroscopy experiments. L.P. performed molecular modelling and molecular dynamics simulations. R.S.P. designed and optimized NMR experiments, wrote NMR pulse sequences and collected data. L.M. conceived of lysine methylation of the β_2AR, wrote NMR pulse sequences and designed NMR experiments. B.K.K. supervised the overall project, designed experiments, collected diffraction data and wrote the paper.

Author Information Coordinates and structure factors for [^{13}C]methyl-β_2AR–Fab5 have been deposited in the Protein Data Bank under accession code 3KJ6. 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 B.K.K. (kobilka@stanford.edu).

METHODS

Buffers. Buffer A is 20 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% dodecylmaltoside. Buffer B is 20 mM HEPES, pH 7.4, 80 mM NaCl, 0.1% dodecylmaltoside. Buffer C is 20 mM HEPES, pH 7.4, 60 mM NaCl, 0.1% dodecylmaltoside. Buffer D is 20 mM HEPES, pH 7.4, 350 mM NaCl, 0.1% dodecylmaltoside. Buffer E is buffer A plus 2 mM $CaCl_2$ and 0.01% cholesterol hemisuccinate. Buffer F is buffer A plus 5 mM EDTA and 0.01% cholesterol hemisuccinate.

Preparation of modified β_2-adrenergic receptors for NMR. The coding sequence of wild-type human β_2AR was cloned into the pFastBac1 Sf9 insect cell expression vector (Invitrogen) and modified as described previously[3]. For small-scale expression trials (1 litre or less), recombinant baculovirus was made with the Bac-to-Bac system (Invitrogen). For large NMR and crystallography-scale expressions, the β_2AR cDNA was subcloned into the pVL1392 transfer vector and recombinant baculovirus was made with the BestBac system (Expression Systems). All cells were cultured in ESF 921 insect cell medium (Expression Systems). β_2AR365 (Supplementary Fig. 2) was expressed in Sf9 insect cells infected with baculovirus and solubilized in 1% *n*-dodecyl-β-D-maltopyranoside (dodecylmaltoside; Anatrace) in accordance with methods described previously[31]. M1 Flag affinity chromatography (Sigma) was used as the initial purification step. Flag-purified receptor was treated with 100 μM tris(2-carboxyethyl)phosphine followed by two additions of 2 mM iodoacetamide (twice, each for 1 h on ice) to alkylate reactive cysteine residues that can cause disulphide aggregation (these and all other reagents were from Sigma unless otherwise noted). Alternatively, β_2AR365 was labelled with the cysteine-reactive fluorophore monobromobimane before alkylation, to assess ligand-induced conformational changes[32]. Alkylation was quenched by the addition of 5 mM L-cysteine. Functional β_2AR365 was then selectively purified by alprenolol-Sepharose chromatography[31]. The next preparation steps varied depending on the particular NMR sample being prepared.

Preparation of carazolol-bound [^{13}C]methyl-β_2AR365. β_2AR365 (purified by alprenolol-Sepharose) was reductively methylated by sequentially adding 10 mM freshly prepared sodium cyanoborohydride, briefly vortex-mixing, and then adding 10 mM ^{13}C-enriched (99%) formaldehyde (CLM-806-1; Cambridge Isotope Labs). We feel that the order of reagent addition is important. We preferred to first set a reducing environment in solution to avoid protein cross-linking by formaldehyde. The reductive methylation reaction was then allowed to proceed overnight (minimum 8 h) at 4 °C with nutation. A second addition of sodium cyanoborohydride and [^{13}C]formaldehyde was made exactly as before, followed by another 4 h incubation at 4 °C. [^{13}C]Methyl-β_2AR365 was then dialysed extensively against buffer A plus 1 μM carazolol to remove unreacted methylation reagents and replace alprenolol with carazolol.

Reductive methylation destroys the antigenicity of the M1 Flag epitope, so we could not use a second M1 Flag affinity step to concentrate [^{13}C]methyl-β_2AR365 to NMR concentrations. The receptor was instead loaded onto Q Sepharose anion-exchange resin equilibrated in buffer B. We estimate the capacity of Q Sepharose to be about 2.5 mg of [^{13}C]methyl-β_2AR365 per ml of resin. Binding is largely mediated through the acidic Flag epitope (amino-acid sequence DYKDDDDA), because TEV-cleaved β_2AR365 does not bind tightly to Q Sepharose resin. The column was then washed with three column volumes of buffer C plus 1 μM carazolol, and eluted with buffer D plus 1 μM carazolol. A one-tenth volume of a saturated cholesterol hemisuccinate solution in buffer A was added to the eluate to enhance receptor stability. [^{13}C]methyl-β_2AR365 was typically eluted at a concentration of about 100 μM as determined by measurement of carazolol fluorescence[33]. N-linked glycosylations were removed by treatment with PNGase F (New England BioLabs; 750 units per mg of β_2AR365) for 1 h at 22–25 °C. The sample was dialysed extensively against buffer A plus 100 nM carazolol, followed by two dialysis steps against a small volume (about 25 ml) of buffer A prepared in 98% ^2H_2O (Cambridge Isotope Labs) plus 100 nM carazolol. Receptor was then concentrated to ≈200 μM with a 100-kDa cutoff Vivaspin concentrator (Vivascience). It is important to wash the concentrator membrane extensively with water, followed by ^2H_2O, to remove as much glycerol as possible because it can interfere with NMR spectroscopy.

Final NMR samples (about 270 μl) were loaded into Shigemi microtubes susceptibility-matched to ^2H_2O (Shigemi Inc.) and sealed. Carazolol-bound [^{13}C]methyl-β_2AR samples remained stable for more than four months with no visible precipitation, degradation of NMR spectral quality, or decrease in bound carazolol fluorescence. This method was used to prepare the samples shown in Fig. 2a and Supplementary Figs 7–9 and 13.

Preparation of unliganded [^{13}C]methyl-β_2AR365. β_2AR365 (purified by alprenolol-Sepharose, 2 mM $CaCl_2$ added) was loaded directly onto M1 resin for a second Flag affinity step. After loading, the column was washed at 10 ml h^{-1} with six column volumes of buffer E plus 30 μM atenolol. Atenolol is an antagonist with relatively low affinity for the β_2AR ($K_d \approx 1$ μM)[34]. This wash step was included to displace the higher-affinity antagonist alprenolol ($K_d \approx 1$ nM) from β_2AR365 by competition. The Flag M1 column was then washed extensively with buffer E to remove atenolol and guarantee that all bound ligand was removed. Unliganded β_2AR365 was then eluted with buffer F plus Flag peptide (100 μg ml^{-1}). Glycosylations were removed by treatment with PNGase F as described earlier. Receptor was dialysed against buffer A with a 20-kDa cutoff dialysis cassette (Slide-A-Lyzer; Pierce) to ensure the removal of Flag peptide before [^{13}C]methylation. Flag peptide (amino-acid sequence DYKDDDDK) has three primary amines that can be [^{13}C]methylated and cause undesirable NMR background.

Unliganded β_2AR365 was reductively methylated (as described earlier), dialysed against buffer A to remove labelling reagents, dialysed against buffer A prepared in 98% ^2H_2O, and concentrated for NMR as described above. This method was used to prepare the samples shown in Figs 3 and 4 and Supplementary Figs 12 and 15. Ligand additions to unliganded [^{13}C]methyl-365N NMR samples were made from ligand stocks dissolved in perdeuterated dimethyl d_6-sulfoxide (DLM-10-10; Cambridge Isotope Labs). (*R*,*R*)-formoterol was a gift from Sepracor.

Preparation of carazolol-bound [^{13}C]methyl-β_2AR365 mutants for Lys 305 peak assignment. To facilitate the assignment of dimethyllysine peaks 1 and 2 (Fig. 2a), we developed a small-scale expression and purification technique that yields an NMR sample of carazolol-bound β_2AR365 (about 2–3 mg) from 1–2 litres of Sf9 cells. We modified β_2AR365 to include a carboxy-terminal hexahistidine tag for nickel affinity chromatography[31]. Six histidine codons were added to the β_2AR365 cDNA between the Gly 365 and STOP codons (β_2AR365-His).

To maximize the yield of β_2AR365-His, we added 200 nM carazolol directly to the culture medium at the time of infection with baculovirus. Buffers for all subsequent purification steps also included a minimum of 200 nM carazolol. We have observed that inverse agonists improve cell-surface expression and help to stabilize the receptor during solubilization. However, the affinity of β_2AR for carazolol ($K_d < 0.1$ nM) is sufficiently high for binding to be essentially irreversible. This precludes the use of alprenolol-Sepharose chromatography as a functional purification step. Instead, β_2AR365-His was first purified by Flag affinity chromatography. Reactive cysteine residues were alkylated as described above, and the protein was dialysed extensively against buffer A to remove Flag peptide. Reductive methylation was performed as described above. [^{13}C]Methyl-β_2AR365-His was then purified by nickel affinity chromatography (Chelating Sepharose Fast Flow; GE Healthcare) as described previously[31]. Final NMR dialysis and concentration were performed as described above. This method was used to prepare the samples shown in Fig. 2b, c and Supplementary Figs 10 and 11.

Crystallographic data collection and processing. Crystals of reductively methylated β_2AR365–Fab5 complex were generated essentially as described in ref. 3 and were isomorphous to the non-methylated receptor–Fab5 complex. Data collection was performed with the 10-μm collimated microfocus beamline 23ID-B of the Advanced Photon Source at Argonne National Laboratory. A data set comprising 316° of oscillation data was obtained from a single crystal (see Supplementary Table 2). Because of radiation damage, only 5–10° of data (1° per frame) could be measured before the crystal was translated to a new position. Data were processed with HKL2000 (ref. 35). Similarly to the previous β_2AR24/365–Fab5 complex data reduction, global post-refinement of the unit-cell parameters was not performed. Rather, the unit-cell parameters were obtained from indexing and refinement from one wedge of data, and were subsequently used for processing the remaining data without unit-cell constant refinement.

Structure solution and refinement. The structure of the methylated β_2AR365–Fab5 complex was solved by initially performing rigid-body refinement in CNS[36] using the unmethylated β_2AR24/365–Fab5 complex structure (PDB accession code 2R4S) as a single rigid body. This gave R and R_{free} values of 0.358 and 0.348, respectively. Multiple rounds of manual rebuilding, positional refinement, grouped temperature factor refinement and Translation–Libration–Screw (TLS) refinement were performed with the PHENIX package[37], bringing R and R_{free} values down to 0.233 and 0.274, respectively. As in the previous β_2AR24/365–Fab5 complex refinement[3], only those residues that could be unambiguously assigned were included in the final model. In addition to the residues present in the β_2AR24/365–Fab5 structure (PDB accession code 2R4S), receptor residues 35–36, 91 and 307–310 were included in the methylated β_2AR365–Fab5 model. The following receptor residues were modelled as alanine because of insufficient electron density to model full side chains: 36, 39, 42, 49, 53, 55, 63, 69, 77, 112, 114, 120, 122, 131, 147, 156, 209, 227, 232, 263, 279, 287, 308, 309, 321, 324, 326 and 332.

31. Kobilka, B. K. Amino and carboxyl terminal modifications to facilitate the production and purification of a G protein-coupled receptor. *Anal. Biochem.* 231, 269–271 (1995).

32. Yao, X. J. *et al.* The effect of ligand efficacy on the formation and stability of a GPCR–G protein complex. *Proc. Natl Acad. Sci. USA* **106,** 9501–9506 (2009).
33. Tota, M. R. & Strader, C. D. Characterization of the binding domain of the β-adrenergic receptor with the fluorescent antagonist carazolol. Evidence for a buried ligand binding site. *J. Biol. Chem.* **265,** 16891–16897 (1990).
34. Baker, J. G. The selectivity of β-adrenoceptor antagonists at the human β_1, β_2 and β_3 adrenoceptors. *Br. J. Pharmacol.* **144,** 317–322 (2005).
35. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Macromol. Crystallogr. A* **276,** 307–326 (1997).
36. Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
37. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).

The Dbf4–Cdc7 kinase promotes S phase by alleviating an inhibitory activity in Mcm4

Yi-Jun Sheu[1] & Bruce Stillman[1]

Eukaryotic DNA replication uses kinase regulatory pathways to facilitate coordination with other processes during cell division cycles and response to environmental cues. At least two cell cycle-regulated protein kinase systems, the S-phase-specific cyclin-dependent protein kinases (S-CDKs) and the Dbf4–Cdc7 kinase (DDK, Dbf4-dependent protein kinase) are essential activators for initiation of DNA replication[1–5]. Although the essential mechanism of CDK activation of DNA replication in *Saccharomyces cerevisiae* has been established[6,7], exactly how DDK acts has been unclear. Here we show that the amino terminal serine/threonine-rich domain (NSD) of Mcm4 has both inhibitory and facilitating roles in DNA replication control and that the sole essential function of DDK is to relieve an inhibitory activity residing within the NSD. By combining an *mcm4* mutant lacking the inhibitory activity with mutations that bypass the requirement for CDKs for initiation of DNA replication, we show that DNA synthesis can occur in G1 phase when CDKs and DDK are limited. However, DDK is still required for efficient S phase progression. In the absence of DDK, CDK phosphorylation at the distal part of the Mcm4 NSD becomes crucial. Moreover, DDK-null cells fail to activate the intra-S-phase checkpoint in the presence of hydroxyurea-induced DNA damage and are unable to survive this challenge. Our studies establish that the eukaryote-specific NSD of Mcm4 has evolved to integrate several protein kinase regulatory signals for progression through S phase.

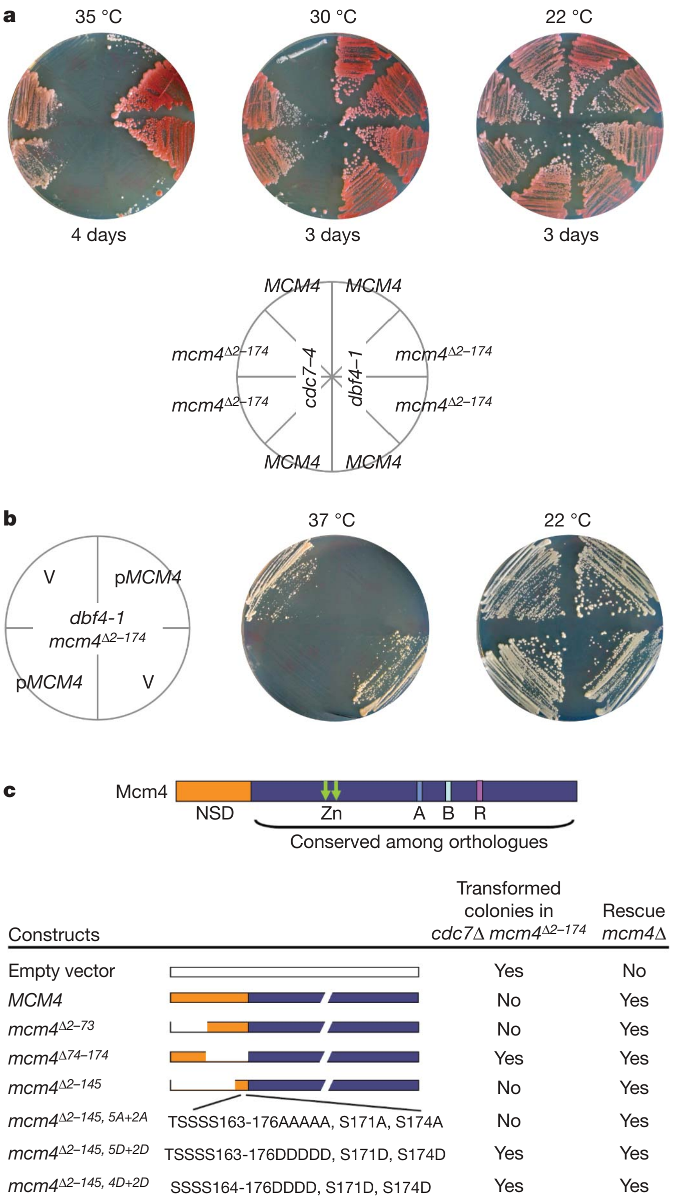
In the early 1970s, studies on fusion of human cells indicated that DNA in G1 nuclei was competent for initiation of DNA replication, but G1 cells lacked an activator(s) that was present in S phase cells[8]. The competent state has been defined as licensing of replication origins before S phase[1,5,9,10]. The process occurs at the M-phase exit through G1 phase, when a pre-replicative complex forms at each origin. Pre-replicative complex assembly begins with the binding of the origin recognition complex (ORC), which recruits more protein factors, and ultimately completes with the loading of the minichromosome maintenance (MCM) complex. Subsequently S-phase-specific kinases, S-CDKs and DDK, activate this competent state by promoting assembly of the CMG complex (Cdc45, MCM and GINS), the active replicative helicase[11–13]. The minimal set of S-CDK targets essential for initiation of replication has been identified[6,7]. S-CDKs phosphorylate Sld2 and Sld3, enabling them to bind to Dpb11[6,7,14]. Genetic and biochemical evidence indicated the MCM complex as one DDK target[3,4]. In budding yeast, DDK phosphorylates several MCM subunits and a mutation in *MCM5*, *mcm5-bob1*, can survive without DDK[15–19].

DDK binds to Mcm4 via a kinase-docking domain, allowing processive phosphorylation of several sites within the adjacent 174 amino acid NSD[18]. Because deletion of NSD does not prevent cells from initiating DNA replication, it is likely that the role of NSD is regulatory. One hypothesis is that the NSD of Mcm4 blocks the activation of licensed origins and phosphorylation of the NSD by DDK alleviates the inhibition. To test this idea, we replaced the chromosomal *MCM4* with $mcm4^{\Delta 2-174}$, which lacks the entire NSD, in the temperature sensitive (*ts*) DDK mutants *cdc7-4* and *dbf4-1*. Deletion of the Mcm4 NSD rescued the *ts* defect of *cdc7-4* or *dbf4-1* (Fig. 1a). Moreover, *cdc7*Δ $mcm4^{\Delta 2-174}$ cells were viable (Supplementary Fig. 1). The *cdc7*Δ $mcm4^{\Delta 2-174}$ cells, however, grow slowly, probably either due to (1) residues 2–174 harbouring a domain needed for optimal MCM functions, or (2) DDK having another function in addition to its essential role in regulating Mcm4. Nevertheless, removing the Mcm4 NSD allows cells to bypass the essential function of DDK.

The ability of $mcm4^{\Delta 2-174}$ to bypass DDK is recessive to *MCM4*. Reintroducing an *MCM4* vector into *dbf4-1* $mcm4^{\Delta 2-174}$ allows cells to grow better at the permissive temperature (22 °C), but they did not grow at 37 °C, in contrast to the empty vector (Fig. 1b). Moreover, the *MCM4* plasmid, unlike the empty vector, failed to yield transformed colonies in *cdc7-4* $mcm4^{\Delta 2-174}$ or *cdc7*Δ $mcm4^{\Delta 2-174}$ cells, whereas *CDC7* efficiently rescued *cdc7*Δ $mcm4^{\Delta 2-174}$ cells (Supplementary Fig. 2a, b). Together, these results indicate that the Mcm4 NSD contains an inhibitory activity that renders DDK essential for viability. Therefore, we used transformation of *cdc7*Δ $mcm4^{\Delta 2-174}$ cells as an assay to map the inhibitory activity. Whereas transformation of the $mcm4^{\Delta 74-174}$ plasmid or empty vector yielded numerous colonies, transformation of plasmids carrying either *MCM4*, $mcm4^{\Delta 2-73}$ or $mcm4^{\Delta 2-145}$ produced none (Fig. 1c and Supplementary Fig. 2c). Thus, the inhibitory activity resides within 74–174 of Mcm4 and residues 146–174 are sufficient for inhibiting transformation of *cdc7*Δ $mcm4^{\Delta 2-174}$ cells. We have previously demonstrated DDK target sites within the 146–174 region[18]. Here, we found that the phosphomimetic mutation constructs ($mcm4^{\Delta 2-145,\ 5D+2D}$ or $mcm4^{\Delta 2-145,\ 4D+2D}$) produced many transformed colonies, whereas a construct that could not be phosphorylated by DDK ($mcm4^{\Delta 2-145,\ 5A+2A}$) failed to produce transformants in *cdc7*Δ $mcm4^{\Delta 2-174}$ cells (Fig. 1c and Supplementary Fig. 2c). Thus, at least a portion of the Mcm4 NSD proximal to the DDK docking domain is inhibitory and phosphorylation of this region by DDK antagonizes the inhibitory effect.

We tested those *mcm4* alleles without the inhibitory activity for DDK bypass using a modified plasmid shuffle assay (Fig. 2a; see legend). This assay is stringent and relies on a centromere (*CEN*)-based plasmid with a single replication origin to carry the tested allele. Thus, only those *mcm4* alleles that can both fulfil the function of Mcm4 and bypass the requirement for DDK efficiently would allow the tester strain to survive on 5-fluoroorotic acid (5-FOA) medium. Plasmids carrying $mcm4^{\Delta 74-174}$, $mcm4^{\Delta 2-145,\ 5D+2D}$ and $mcm4^{\Delta 2-145,\ 4D+2D}$ allowed growth on 5-FOA media, indicating that these mutant alleles can cope with simultaneous loss of both *MCM4* and *CDC7* genes. In contrast, plasmids carrying *CDC7*, *MCM4*, $mcm4^{\Delta 2-174}$, $mcm4^{\Delta 2-73}$, $mcm4^{\Delta 2-145}$ and $mcm4^{\Delta 2-145,5A+2A}$ scored negative in this assay.

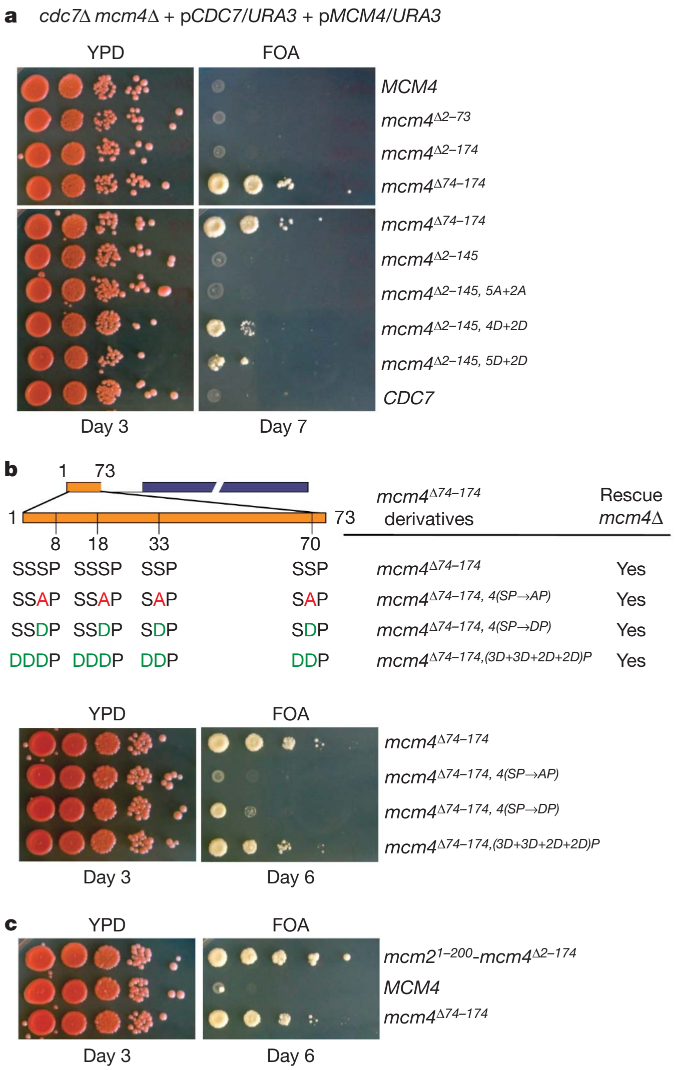
[1] Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.



Constructs		Transformed colonies in *cdc7*Δ *mcm4*$^{\Delta 2\text{–}174}$	Rescue *mcm4*Δ
Empty vector		Yes	No
MCM4		No	Yes
mcm4$^{\Delta 2\text{–}73}$		No	Yes
mcm4$^{\Delta 74\text{–}174}$		Yes	Yes
mcm4$^{\Delta 2\text{–}145}$		No	Yes
mcm4$^{\Delta 2\text{–}145,\ 5A+2A}$	TSSSS163-176AAAAA, S171A, S174A	No	Yes
mcm4$^{\Delta 2\text{–}145,\ 5D+2D}$	TSSSS163-176DDDDD, S171D, S174D	Yes	Yes
mcm4$^{\Delta 2\text{–}145,\ 4D+2D}$	SSSS164-176DDDD, S171D, S174D	Yes	Yes

Figure 1 | An inhibitory activity within the Mcm4 NSD is responsible for the dependency of cells on DDK for viability. a, Yeast strains were grown on YPD plates at permissive and non-permissive temperatures (30 °C and above for *cdc7-4* and 35 °C and above for *dbf4-1*). The *mcm4*$^{\Delta 2\text{–}174}$ allele was introduced to *cdc7-4* and *dbf4-1* cells by two-step gene replacement. Shown are parental strains (top sectors) and three isolates of the second-step homologous recombination products. **b**, The *dbf4-1 mcm4*$^{\Delta 2\text{–}174}$ cells transformed with empty vector (V) or vector carrying *MCM4* were streaked on selective medium and allowed to grow at 37 °C or 22 °C for 5 days. **c**, Diagram of Mcm4 and summary of transformation assay and complementation of *mcm4*Δ by the same plasmid constructs (Supplementary Fig. 2c). However, *mcm4*$^{\Delta 2\text{–}145,5A+2A}$ does exhibit growth defect even in the presence of DDK[18].

Although unphosphorylated *mcm4*$^{\Delta 2\text{–}145}$ seems to be sufficient for exerting the inhibitory effect (Fig. 1c), the inhibitory domain extends beyond residues 146–174 because *mcm4*$^{\Delta 147\text{–}174}$, *mcm4*$^{\Delta 123\text{–}174}$ or *mcm4*$^{\Delta 98\text{–}174}$ also fail to support DDK-independent growth (Supplementary Fig. 3). Importantly, alanine substitution of 11 potential DDK phosphorylation sites within 74–174 in the full length NSD is lethal even in the presence of DDK (Supplementary Fig. 4). It is possible that, when unphosphorylated, the proximal portion of the NSD exerts its inhibitory effect by imposing on the MCM complex a conformation that is not permissive for recruitment of activating factors, such as Cdc45 and GINS, and phosphorylation by DDK or removal of this domain allows the complex to assume a permissive state. The NSD inhibitory domain may also alter the MCM hexamer oligomeric state. Alternatively, the domain may directly block the access of activating factors.



8	18	33	70	*mcm4*$^{\Delta 74\text{–}174}$ derivatives	Rescue *mcm4*Δ
SSSP	SSSP	SSP	SSP	*mcm4*$^{\Delta 74\text{–}174}$	Yes
SSAP	SSAP	SAP	SAP	*mcm4*$^{\Delta 74\text{–}174,\ 4(SP\to AP)}$	Yes
SSDP	SSDP	SDP	SDP	*mcm4*$^{\Delta 74\text{–}174,\ 4(SP\to DP)}$	Yes
DDDP	DDDP	DDP	DDP	*mcm4*$^{\Delta 74\text{–}174,(3D+3D+2D+2D)P}$	Yes

Figure 2 | The status of the Mcm4 N terminus determines the efficiency of DDK-independent cell proliferation. a–c, A modified plasmid shuffle assay was used to identify *MCM4* alleles that allow cell growth in the absence of *CDC7* and *MCM4* when expressed from single-origin vectors. The tester yeast strain was constructed with both *MCM4* and *CDC7* genes deleted from its chromosomes while carrying both essential genes on plasmids with the *URA3* gene that can be counter-selected in the presence of 5-FOA. The tester cells were transformed with indicated plasmids and assayed for growth on YPD and 5-FOA media. All *mcm4* alleles used in the assay complement *mcm4*Δ. **a**, DDK-independent cell growth by the *mcm4* mutants described in Fig. 1. **b**, Top panel, diagram of *mcm4*$^{\Delta 74\text{–}174}$ derivatives with mutations at CDK phosphorylation sites and their preceding phospho-acceptor residues within the distal NSD (residues 2–73). Bottom panel, the ability of *mcm4* alleles described above to support DDK-independent growth. **c**, DDK-independent cell growth supported by *mcm2*$^{1\text{–}200}$-*mcm4*$^{\Delta 2\text{–}174}$.

The fact that *mcm4*$^{\Delta 2\text{–}174}$ scored negative in the stringent assay for DDK-independent cell proliferation is consistent with our previous finding that it executes the function of *MCM4* poorly[18]. In contrast, *mcm4*$^{\Delta 74\text{–}174}$ exhibited the best growth on 5-FOA medium in the same experiment (Fig. 2a), indicating that the distal part of the Mcm4 NSD (residues 2–73) has a positive role in supporting DDK-independent growth. A shorter version (residues 2–37) could function similarly (Supplementary Fig. 3). The distal NSD of Mcm4 is serine/threonine (S/T) rich and contains four CDK target (S/T-P) sites[20], all of which are preceded by additional S/Ts (Fig. 2b), which would become favourable phospho-acceptors for DDK following priming phosphorylation[21,22]. Converting all four CDK sites to alanines within *mcm4*$^{\Delta 74–174}$ (*mcm4*$^{\Delta 74–174,\ 4(SP\rightarrow AP)}$) had little effect on its ability to rescue *mcm4*Δ (Fig. 2b and Supplementary Fig. 5). However, this mutant failed to bypass DDK. In contrast, phosphomimetic substitution of these sites (*mcm4*$^{\Delta 74–174,\ 4(SP\rightarrow DP)}$) allowed DDK-independent growth and additional phosphomimetic substitutions of all the preceding S/Ts further improved the growth (Fig. 2b). One caveat is that constitutive phosphorylation or phosphomimetic substitution may compromise other aspects of Mcm4 function in DNA replication (see Supplementary Fig. 10). As a result, the phosphomimetic derivatives of *mcm4*$^{\Delta 74–174}$ do not support growth better than *mcm4*$^{\Delta 74–174}$ which is regulated by phosphorylation. Nevertheless, the positive function within the distal NSD may depend on phosphoregulation by CDK, and possibly by DDK. In the absence of DDK, CDK control of this region becomes essential. It remains to be addressed whether additional kinases also contribute to regulation of the Mcm4 NSD, which also contains several potential ATM/ATR target (S/T-Q) sites and many of these are also preceded by stretches of S/Ts.

Other MCM subunits such as Mcm2 and Mcm6 have also extended unstructured amino-terminal domains (NTDs) harbouring DDK target sites (Supplementary Fig. 6a). However, none of the N-terminal deletion mutants of *MCM2* or *MCM6* tested supported DDK-independent cell growth using analogous plasmid shuffle assays (Supplementary Fig. 6b and c). Thus, the inhibitory activity may be a unique feature of Mcm4. Yet, we have previously demonstrated that the NTD of Mcm2 can functionally replace the Mcm4 NSD in supporting normal cell proliferation and timely S phase progression[18]. The *mcm2*$^{1–200}$-*mcm4*$^{\Delta 2–174}$ fusion can function as an *mcm4* allele that supports DDK-independent cell proliferation to an extent that surpasses *mcm4*$^{\Delta 74–174}$ (Fig. 2c and Supplementary Fig. 7a). Therefore, the Mcm2 NTD has a positive role in activating DNA replication. This region contains >30% negatively charged aspartic acid (D) and glutamic acid (E) residues, reminiscent of phosphorylated S/Ts. Thus, the Mcm2 NTD may act like a phosphorylated distal NSD of Mcm4.

DDK-bypass alleles of *MCM4* were introduced into the endogenous locus in subsequent experiments. *mcm4*$^{\Delta 74–174}$, *mcm2*$^{1–200}$-*mcm4*$^{\Delta 2–174}$ and *mcm5-bob1* cells grew at the same rate as the wild type cells (Supplementary Fig. 7a and b). The proliferation rates of *cdc7*Δ *mcm4*$^{\Delta 74–174}$ and *cdc7*Δ *mcm5-bob1* were comparable. Consistent with earlier observations (Fig. 2a, c), *cdc7*Δ *mcm2*$^{1–200}$-*mcm4*$^{\Delta 2–174}$ cells proliferated faster than *cdc7*Δ *mcm4*$^{\Delta 74–174}$ cells, whereas *cdc7*Δ *mcm4*$^{\Delta 2–174}$ cells proliferated more slowly (Supplementary Fig. 7b). Cells without DDK grew slowly, entered S phase later and progressed through S phase at slower rates than their DDK positive counterparts (Fig. 3a). These results indicate that DDK has other non-essential roles in regulating S phase progression in addition to alleviating the inhibitory activity within the proximal NSD. For example, DDK may phosphorylate the distal NSD of Mcm4 or other substrates for efficient S phase progression.

One important consequence of DDK action during S phase is formation of a stable complex between Cdc45 and MCM at each origin as it is activated[18,23–25]. To determine if *mcm4*$^{\Delta 74–174}$ can bypass the requirement of DDK for Cdc45–MCM complex formation, co-immunoprecipitation of Cdc45 with Mcm2 antibodies in *cdc7*Δ *mcm4*$^{\Delta 74–174}$ cells, *mcm4*$^{\Delta 74–174}$ and wild type cells was examined. Cells were synchronized to allow progression through the cell cycle from G1 at 25 °C (Fig. 3a). The Cdc45–MCM complex was detected at similar intensity and kinetics in wild type and *mcm4*$^{\Delta 74–174}$ cells, with a peak at ~40 min after G1 release (Fig. 3b). In the *cdc7*Δ *mcm4*$^{\Delta 74–174}$ cells, the complex appeared at a later time (at ~60 min, peak at ~80 min) and at reduced levels. Nevertheless, these results demonstrated that eliminating the inhibitory domain in the Mcm4 NSD allows the Cdc45–MCM complex to form in the absence of DDK, but DDK is still needed for timely Cdc45–MCM association under this bypass condition.

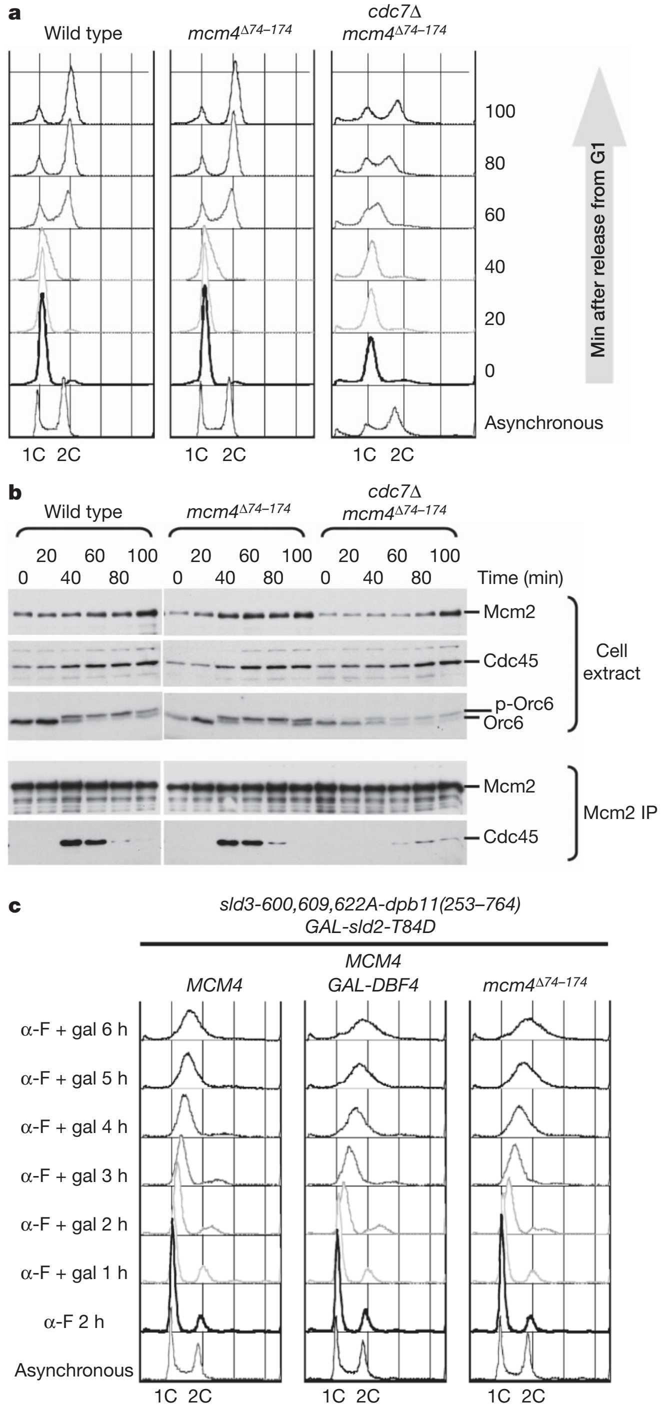


Figure 3 | Removal of the N-terminal inhibitory domain of Mcm4 allows DDK-independent initiation of DNA replication and S phase progression. a, Cell cycle progression of wild type, *mcm4*$^{\Delta 74–174}$ and *cdc7*Δ *mcm4*$^{\Delta 74–174}$ cells. Cells were synchronized by G1 arrest and released into fresh YPD medium at 25 °C, collected at indicated times, and analysed for DNA content by flow cytometry. **b**, Kinetics of Cdc45–MCM complex formation. Cell extracts were prepared from samples collected in **a** and subjected to immunoprecipitation (IP) using monoclonal antibody against Mcm2. Cell extracts and IP were analysed by immunoblotting. **c**, Flow cytometry analysis for DNA content in G1. CDK bypass cells containing *SD* fusion (*sld3-600,609,622A-dbp11(253–764)*) and *GAL-sld2-T84D*, with or without additional modification (*GAL-DBF4* or *mcm4*$^{\Delta 74–174}$) were synchronized and held in G1 using α-factor (α-F) and galactose (gal) was added to induce expression of sld2-T84D and Dbf4.

Recent studies reported conditions that allow yeast cells to replicate DNA in the absence of S-CDKs[6,7]. For example, the requirement

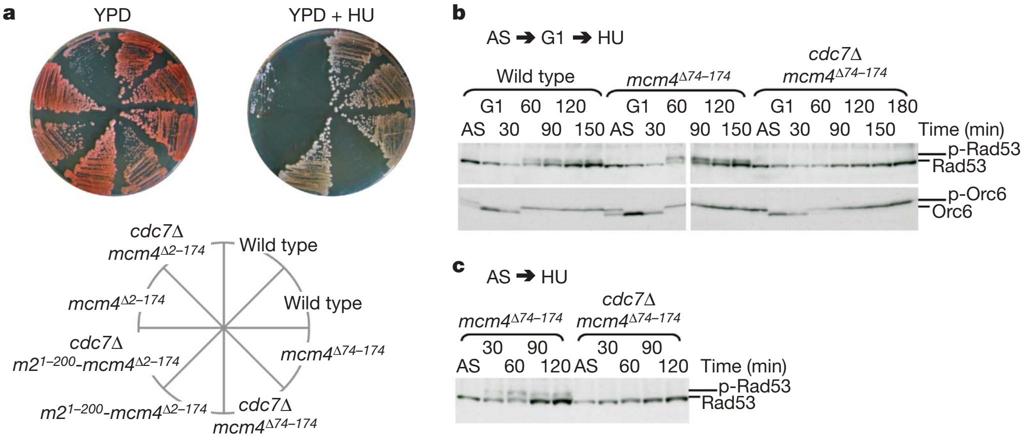


Figure 4 | Deletion of the N-terminal inhibitory domain of Mcm4 does not bypass the requirement for DDK for cell survival and checkpoint activation in the presence of hydroxyurea. **a**, Yeast strains of indicated genotypes at endogenous loci were grown on YPD medium with or without 100 mM hydroxyurea (HU). **b** and **c**, Immunoblot analysis for Rad53 and Orc6 phosphorylation status in wild type, *mcm4*$^{\Delta74–174}$ and *cdc7Δ mcm4*$^{\Delta74–174}$ cells. **b**, Protein samples were prepared from cells synchronized in G1 and released into 200 mM hydroxyurea for the indicated time. **c**, Protein samples were prepared from log-phase cells treated with hydroxyurea for the indicated time.

for S-CDK for DNA synthesis can be bypassed by combining an *sld3–dpb11* fusion (*SD* fusion) and overexpression of a phospho-mimetic *sld2-T84D* mutation[7]. Under this S-CDK bypass condition, DDK is limiting for DNA replication and overexpression of *DBF4* is necessary for extensive DNA synthesis in α-factor arrested, G1 cells. Instead of Dbf4 over-production, replacing the chromosomal *MCM4* with *mcm4*$^{\Delta74–174}$ in this S-CDK bypass system allowed a similar extent of DNA replication in G1 (Fig. 3c and Supplementary Fig. 8). We also observed a modest but consistent DDK-independent increase of DNA content in G1 by introducing *mcm4*$^{\Delta74–174}$ to a different S-CDK bypass condition[6] (Supplementary Fig. 9). Furthermore, unlike another DDK bypass mutation *mcm5-bob1*, *mcm4*$^{\Delta74–174}$ does not exhibit synthetic lethality with *SD* fusion. Thus, DDK bypass is not necessarily synthetically lethal with *SD* fusion as previously suggested[7]. This result indicates that DDK bypass by *mcm4*$^{\Delta74–174}$ is different from DDK bypass by *mcm5-bob1*. Moreover, accumulating biochemical evidence indicates that Mcm4, Mcm2 and Mcm6, but not Mcm5, are substrates of DDK[15,17–19,21,22,24,26]. Because *MCM4* does not cause lethality in *mcm5-bob1* cells lacking DDK, *mcm5-bob1* is epistatic to *MCM4* in this condition. Thus, DDK bypass by *mcm5-bob1* is probably downstream of the inhibitory function of the Mcm4 NSD.

In the presence of DDK, *mcm4*$^{\Delta2–174}$ cells, but not *mcm4*$^{\Delta74–174}$ or *mcm2*$^{1–200}$-*mcm4*$^{\Delta2–174}$ cells, were sensitive to the ribonucleotide reductase inhibitor hydroxyurea (Fig. 4a), indicating that the distal Mcm4 NSD or its functional equivalent (for example, NTD of Mcm2) is required under DNA-damaging conditions. Like *cdc7Δ mcm5-bob1* cells[19], *cdc7Δ mcm4*$^{\Delta2–174}$, *cdc7Δ mcm4*$^{\Delta74–174}$ and *cdc7Δ mcm2*$^{1–200}$-*mcm4*$^{\Delta2–174}$ cells were not viable in hydroxyurea (Fig. 4a). Thus, *mcm4*$^{\Delta74–174}$ and *mcm2*$^{1–200}$-*mcm4*$^{\Delta2–174}$ do not bypass the requirement of DDK for growth in the presence of hydroxyurea. We examined checkpoint activation under synchronous G1 to hydroxyurea release by monitoring Rad53 hyper-phosphorylation (Fig. 4b and c). Although checkpoint activation in *MCM4* and *mcm4*$^{\Delta74–174}$ cells was efficient, Rad53 hyper-phosphorylation was not detectable in *cdc7Δ mcm4*$^{\Delta74–174}$ cells over the course of 3 h in 200 mM hydroxyurea. A similar defect in checkpoint activation in S phase was also found in *cdc7Δ mcm5-bob1* cells[27]. Although it is conceivable that insufficient initiation in *cdc7Δ mcm4*$^{\Delta74–174}$ cells would evade the checkpoint[28], hydroxyurea treatment of the asynchronous *cdc7Δ mcm4*$^{\Delta74–174}$ culture, which accumulates a large population of S phase cells, still failed to elicit robust Rad53 phosphorylation, unlike the response of the asynchronous *mcm4*$^{\Delta74–174}$ culture (Fig. 4c). Thus, it remains possible that DDK is required for the checkpoint response through Rad53 under replication stress (Supplementary Figs 11 and 12). Overall, our results demonstrate that *mcm4*$^{\Delta74–174}$ can bypass the requirement for DDK in an unperturbed S phase, but it cannot bypass the requirement for DDK in proper intra-S-phase checkpoint response.

Cell fusion experiments indicated that "certain substances which are present in the S component probably migrate into G1 nucleus and cause initiation of DNA synthesis"[8]. The results presented here for DDK and elsewhere for CDK[6,7,14] have uncovered essential targets for such activators that must act on the competent pre-replicative complex. One surprising finding is that the essential DDK activity is to inhibit an intrinsic inhibitor of initiation of DNA replication. Our study shows that the unstructured Mcm4 NSD is a multi-function domain that may serve to integrate various signals to regulate eukaryotic DNA replication.

METHODS SUMMARY

Yeast genetic methods and strain construction, cell extract preparation, immunoprecipitation, immunoblot analysis and antibodies are described in detail in Methods. Conditions of cell growth, cell cycle block and synchronization, and flow cytometry analysis are similar to published methods[6,7,18,19].

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

Received 19 August; accepted 6 November 2009.

1. Bell, S. P. & Dutta, A. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71,** 333–374 (2002).
2. Diffley, J. F. Regulation of early events in chromosome replication. *Curr. Biol.* **14,** R778–R786 (2004).
3. Masai, H. & Arai, K. Cdc7 kinase complex: a key regulator in the initiation of DNA replication. *J. Cell. Physiol.* **190,** 287–296 (2002).
4. Sclafani, R. A. Cdc7p-Dbf4p becomes famous in the cell cycle. *J. Cell Sci.* **113,** 2111–2117 (2000).
5. Stillman, B. Cell cycle control of DNA replication. *Science* **274,** 1659–1663 (1996).
6. Tanaka, S. *et al.* CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. *Nature* **445,** 328–332 (2007).
7. Zegerman, P. & Diffley, J. F. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature* **445,** 281–285 (2007).
8. Rao, P. N. & Johnson, R. T. Mammalian cell fusion: studies on the regulation of DNA synthesis and mitosis. *Nature* **225,** 159–164 (1970).
9. Forsburg, S. L. Eukaryotic MCM proteins: beyond replication initiation. *Microbiol. Mol. Biol. Rev.* **68,** 109–131 (2004).
10. Laskey, R. A., Fairman, M. P. & Blow, J. J. S phase of the cell cycle. *Science* **246,** 609–614 (1989).
11. Gambus, A. *et al.* GINS maintains association of Cdc45 with MCM in replisome progression complexes at eukaryotic DNA replication forks. *Nature Cell Biol.* **8,** 358–366 (2006).
12. Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45/Mcm2–7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc. Natl Acad. Sci. USA* **103,** 10236–10241 (2006).

13. Pacek, M., Tutter, A. V., Kubota, Y., Takisawa, H. & Walter, J. C. Localization of MCM2-7, Cdc45, and GINS to the site of DNA unwinding during eukaryotic DNA replication. *Mol. Cell* **21,** 581–587 (2006).
14. Masumoto, H., Muramatsu, S., Kamimura, Y. & Araki, H. S-Cdk-dependent phosphorylation of Sld2 essential for chromosomal DNA replication in budding yeast. *Nature* **415,** 651–655 (2002).
15. Francis, L. I., Randell, J. C., Takara, T. J., Uchima, L. & Bell, S. P. Incorporation into the prereplicative complex activates the Mcm2–7 helicase for Cdc7–Dbf4 phosphorylation. *Genes Dev.* **23,** 643–654 (2009).
16. Hardy, C. F., Dryga, O., Seematter, S., Pahl, P. M. & Sclafani, R. A. *mcm5/cdc46-bob1* bypasses the requirement for the S phase activator Cdc7p. *Proc. Natl Acad. Sci. USA* **94,** 3151–3155 (1997).
17. Lei, M. *et al.* Mcm2 is a target of regulation by Cdc7–Dbf4 during the initiation of DNA synthesis. *Genes Dev.* **11,** 3365–3374 (1997).
18. Sheu, Y. J. & Stillman, B. Cdc7-Dbf4 phosphorylates MCM proteins via a docking site-mediated mechanism to promote S phase progression. *Mol. Cell* **24,** 101–113 (2006).
19. Weinreich, M. & Stillman, B. Cdc7p–Dbf4p kinase binds to chromatin during S phase and is regulated by both the APC and the RAD53 checkpoint pathway. *EMBO J.* **18,** 5334–5346 (1999).
20. Devault, A., Gueydon, E. & Schwob, E. Interplay between S-cyclin-dependent kinase and Dbf4-dependent kinase in controlling DNA replication through phosphorylation of yeast Mcm4 N-terminal domain. *Mol. Biol. Cell* **19,** 2267–2277 (2008).
21. Cho, W. H., Lee, Y. J., Kong, S. I., Hurwitz, J. & Lee, J. K. CDC7 kinase phosphorylates serine residues adjacent to acidic amino acids in the minichromosome maintenance 2 protein. *Proc. Natl Acad. Sci. USA* **103,** 11521–11526 (2006).
22. Montagnoli, A. *et al.* Identification of Mcm2 phosphorylation sites by S-phase-regulating kinases. *J. Biol. Chem.* **281,** 10281–10290 (2006).
23. Aparicio, O. M., Weinstein, D. M. & Bell, S. P. Components and dynamics of DNA replication complexes in *S. cerevisiae*: redistribution of MCM proteins and Cdc45p during S phase. *Cell* **91,** 59–69 (1997).
24. Masai, H. *et al.* Phosphorylation of MCM4 by Cdc7 kinase facilitates its interaction with Cdc45 on the chromatin. *J. Biol. Chem.* **281,** 39249–39261 (2006).
25. Zou, L. & Stillman, B. Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science* **280,** 593–596 (1998).
26. Jiang, W., McDonald, D., Hope, T. J. & Hunter, T. Mammalian Cdc7–Dbf4 protein kinase complex is essential for initiation of DNA replication. *EMBO J.* **18,** 5703–5713 (1999).
27. Ogi, H., Wang, C. Z., Nakai, W., Kawasaki, Y. & Masumoto, H. The role of the *Saccharomyces cerevisiae* Cdc7–Dbf4 complex in the replication checkpoint. *Gene* **414,** 32–40 (2008).
28. Shimada, K., Pasero, P. & Gasser, S. M. ORC and the intra-S-phase checkpoint: a threshold regulates Rad53p activation in S phase. *Genes Dev.* **16,** 3236–3252 (2002).

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

Acknowledgements We thank A. Stenlund and H.-G. Wendel for critical reading of the manuscript, J. Li, H. Araki and J. Diffley for plasmids and yeast strains. This work was supported by a grant from the US National Institute of General Medical Sciences.

Author Contributions Y.-J.S. and B.S. designed this study, analysed the data and wrote the paper. Y.-J.S. performed the experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.S. (stillman@cshl.edu).

METHODS

Yeast strains. Yeast strains generated in this study were derived from W303-1a (*MAT**a** ade2-1 can1-100 his3-11,15 leu2-3,112 trp1-1 ura3-1*) and are described in Supplementary Table 1, except for YS2041 (see below). A PCR-based gene deletion strategy was used for deletion of *CDC7* and *BAR1*. The deletion cassettes *cdc7Δ::KanMX6*, *cdc7Δ::HIS3* and *bar1Δ::TRP1* are from the Yeast Knockout Collection (distributed by Open Biosystems), YB514 (ref. 19) and y2007 (ref. 7) were as described. All deletions were confirmed by PCR in combination with phenotypic assessment and gene complementation. A two-step gene replacement method was used to replace the endogenous *MCM4* with *mcm4* mutants. Plasmid constructs for two-step gene replacement were generated by sub-cloning the SphI/MluI fragments from pRS415-based constructs carrying Mcm4 N-terminal deletion mutants into the same restriction sites of a pRS306-based integration plasmid (a gift from J. J. Li) containing *MCM4*. *CEN*-based plasmid constructs carrying Mcm4 N-terminal deletion and site-directed mutants were created using fusion PCR strategy. The *mcm4* mutant PCR products were digested with SphI/StuI and cloned into the same sites of pEM54.3 (pRS415/*MCM4*). Constructs were confirmed by DNA sequencing. Some constructs used in this work have been described in our previous work[18].

Plasmid shuffle assay. The tester strain YS2041 (*MATa cdc7Δ::HIS3 mcm4Δ::TRP1 ade2 ura3 his3 leu2 trp1?* + pRS416/CDC7 pRS416/MCM4) is a meiotic product of a diploid strain obtained by crossing YB514 (*MATa ade2-101 ura3-52 lys2-801 his3Δ-200 leu2Δ-1 cdc7Δ*::HIS3 + pRS416/CDC7)[19] and YS958 (*MATα mcm4Δ::TRP1 ade2-1 can1-100 his3-11,-15 leu2-3,112 trp1-1 ura3-1* + pRS416/MCM4). Deletion of *CDC7* and *MCM4* were confirmed by selection for autotrophic traits, PCR and gene complementation. The pRS415-based test plasmids were used for transformation of YS2041. The transformed colonies were isolated and grown in SC-LEU medium overnight, and tenfold serial dilutions starting from 10^6 cells were spotted on 5-FOA plates to select for loss of URA3 plasmids carrying *CDC7* and *MCM4*. The same dilutions starting from 10^5 cells were spotted on YPD plates as control sets.

Cell extract preparation and immunoprecipitation. Yeast cell pellet containing $\sim 6 \times 10^8$ cells was resuspended in 150 μl of EB buffer containing 50 mM HEPES/KOH pH 7.5, 100 mM KCl, 2.5 mM $MgCl_2$, 2 mM NaF, 0.5 mM spermidine, 20 mM β-glycerophosphate, 0.1 mM $ZnSO_4$, 1 mM ATP, 1 mM DTT, 1 mM PMSF, protease inhibitor tablets (EDTA free, Roche). An equal volume of chilled 0.5 mm zirconia/silica beads (BiosSpec Products) was added to cell suspension and cells were lysed by vortexing for 15 cycles of 30 s on 30 s off at maximal strength. The efficiency of cell breakage here was ~ 50% after 12 cycles, as determined by visualization under a microscope. Cell lysates were collected in siliconized microcentrifuge tubes after micro-centrifugation for 10 min at 12,000 r.p.m. (15,300*g*) at 4 °C. Pellets of cell debris and beads were resuspended with 400 μl of EBX buffer (EB Buffer with 0.25% Triton X-100) supplemented with 1 mM $MnCl_2$ and 100 U ml^{-1} DNase I (Roche). The suspensions were allowed to mix for 30 min at 4 °C. Supernatants were collected and pooled with previous lysates after centrifugation for 10 min at 12,000 r.p.m. at 4 °C (15,300*g*). Pellets were resuspended again in the same buffer and allowed to mix for 15 min at 30 °C, before supernatants were collected and pooled with earlier preps. Cell extracts from this preparation were analysed for protein concentrations. Standard immunoprecipitation procedures were performed by mixing ~4 mg of total proteins and 2.5 μl of the mcm2-49 antibody at 4 °C for 30 min and precipitating the complex with GammaBindG Sepharose (GE Healthcare) after washing extensively with EBX buffer.

TCA precipitation of yeast proteins. Yeast cell pellet containing $\sim 5 \times 10^7$ cells was resuspended in 100 μl of TCA lysis buffer (1.85M NaOH and 7.4% β-mercaptoethanol), vortexed and left on ice. After 10 min, 100 μl of 20% TCA was added and gently mixed by inverting tubes. After incubation on ice for another 10 min, pellets were collected by centrifugation at 13,000 r.p.m. for 2 min (17,900*g*), washed with 1 ml ice-cold acetone and dried in vacuumed rotors. Dry pellets were resuspended carefully in 100 μl of 0.1 M NaOH. For analysis on SDS–PAGE, 100 μl of 2× sample buffer was added and samples were boiled for 10 min before loading.

Immunoblot analysis. Proteins from cell extract, immunoprecipitation or TCA precipitation were fractionated by SDS-10% PAGE and transferred to nitrocellulose membrane. Immunoblot analysis was performed as described previously using antibodies against Mcm2 (mcm2-39), Cdc45 (CS1485) and Orc6 (SB49)[18,19]. 12CA5 was used to detect 4HA-sld2-T84D and 2HA-dbf4 and 9E10 was used to detect sld2-11D-Myc. For detection of Rad53 from TCA precipitated yeast proteins, Rad53 (yC-19) antibody sc-6749 (Santa Cruz Biotechnology) was used at 1:1000 dilution and TBS with 0.1% Tween 20 was used for preparing blocking and washing solutions.

LETTERS

Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats

Bancinyane L. Sibanda[1], Dimitri Y. Chirgadze[1] & Tom L. Blundell[1]

Broken chromosomes arising from DNA double-strand breaks result from endogenous events such as the production of reactive oxygen species during cellular metabolism, as well as from exogenous sources such as ionizing radiation[1–3]. Left unrepaired or incorrectly repaired they can lead to genomic changes that may result in cell death or cancer. DNA-dependent protein kinase (DNA-PK), a holoenzyme that comprises the DNA-PK catalytic subunit (DNA-PKcs)[4,5] and the heterodimer Ku70/Ku80, has a major role in non-homologous end joining—the main pathway in mammals used to repair double-strand breaks[6–8]. DNA-PKcs is a serine/threonine protein kinase comprising a single polypeptide chain of 4,128 amino acids and belonging to the phosphatidylinositol-3-OH kinase (PI(3)K)-related protein family[9]. DNA-PKcs is involved in the sensing and transmission of DNA damage signals to proteins such as p53, setting off events that lead to cell cycle arrest[10,11]. It phosphorylates a wide range of substrates *in vitro*, including Ku70/Ku80, which is translocated along DNA[12]. Here we present the crystal structure of human DNA-PKcs at 6.6 Å resolution, in which the overall fold is clearly visible, to our knowledge, for the first time. The many α-helical HEAT repeats (helix–turn–helix motifs) facilitate bending and allow the polypeptide chain to fold into a hollow circular structure. The carboxy-terminal kinase domain is located on top of this structure, and a small HEAT repeat domain that probably binds DNA is inside. The structure provides a flexible cradle to promote DNA double-strand-break repair.

Knowledge of the three-dimensional structure of DNA-PKcs will enable a better understanding of its role in the events that take place in non-homologous end joining (NHEJ). However, owing to its size, crystallization of DNA-PKcs on its own or in complex with its interacting partners has proved challenging. We have now purified DNA-PKcs and complexed it with Ku80ct$_{140}$ and Ku80ct$_{194}$—C-terminal fragments of Ku80 of 140 and 194 amino acids, respectively. The uncomplexed form and both complexes crystallized but under slightly different conditions; the complex with Ku80ct$_{194}$ was used for most experiments and provides the observed structure factor amplitudes with which the electron density maps were calculated. A sodium dodecyl sulphate (SDS)–polyacrylamide gel of the dissolved crystals (see Supplementary Fig. 1) confirms the presence of DNA-PKcs and the Ku80ct$_{194}$ domain. The three-dimensional structure of this complex at 6.6 Å was determined using phases calculated by the multi-wavelength anomalous dispersion (MAD) method with the tantalum bromide heavy atom cluster (see Methods). These crystals have two molecules in the asymmetric unit (Fig. 1a). Although the electron density of most helical regions is clearly visible, the weak electron density in the loop regions connecting the helices (Fig. 1c) makes a reliable fitting of the whole polypeptide chain impossible at

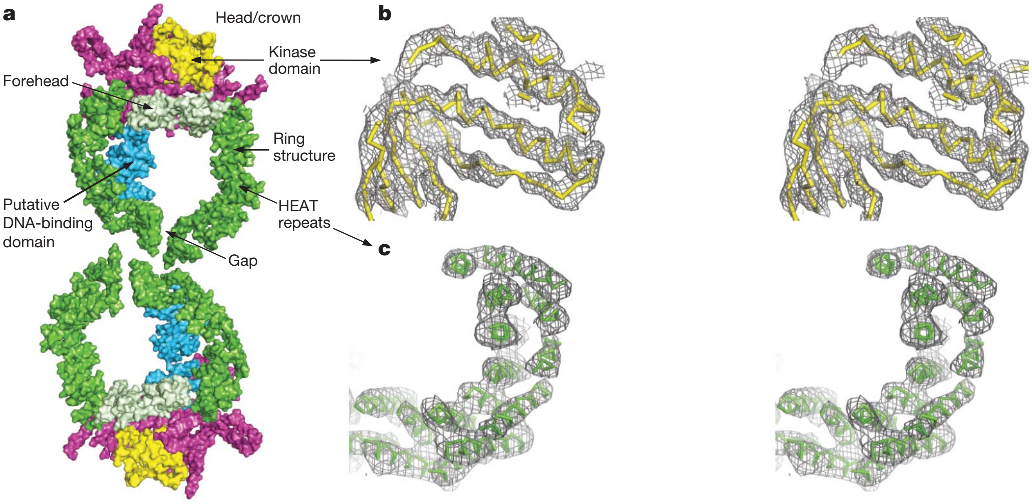


Figure 1 | Crystal structure of DNA-PKcs at 6.6 Å resolution. **a**, Molecular surface of the two DNA-PKcs molecules present in the asymmetric unit of the crystal that are related by the two-fold non-crystallographic (NCS) symmetry. The colour-coding of the molecule is as follows: green, the ring structure; light green, the forehead that is part of the ring structure; cyan, the putative DNA-binding domain; magenta, the larger C-terminal region that carries the FAT and FATC domains; and yellow, the kinase domain. **b**, **c**, Stereo diagrams of (**b**) a representative area of SigmaA-weighted $2F_o - F_c$ electron density map from final refinement in the kinase domain region (ribbon representation in yellow colour), and (**c**) experimental F_o electron density map calculated with phases obtained by the MAD method in the area of the N-terminal HEAT repeat ring, the final DNA-PKcs model shown in green. Both electron density maps shown are contoured at 1.0σ level.

[1]Department of Biochemistry, University of Cambridge, Old Addenbrooke's site, 80 Tennis Court Road, Cambridge CB2 1GA, UK.

this resolution. Moreover, it was also not possible to locate unambiguously the Ku80ct$_{194}$ fragment, which like DNA-PKcs also contains α-helical HEAT repeats[13,14]. The structure presented here gives an overall view of the molecule.

The DNA-PKcs structure, which is organized into several distinct domains (Fig. 2), measures 160 Å from the top of the kinase domain to the bottom of the ring structure, and the ring is about 120 Å in diameter (Fig. 2c). Going anticlockwise from the 'gap' about 66 helices are arranged as HEAT repeats in a ring structure (Fig. 2c). They are folded into a hollow circular structure, which has a concave shape rather like a cradle, when viewed from the side (Fig. 2b).

The region that is probably the head/crown identified in earlier electron microscopy studies[15–18] is shown in yellow and magenta (Figs 1a, b and 2a–c); the kinase domain can be docked into the yellow region (see later), indicating that the head/crown contains the C-terminal region of DNA-PKcs. Indeed, the unequivocal positioning of the kinase domain, together with the reasonably clear path of the HEAT repeats motifs, suggests that the amino terminus is in the ring structure, probably at the right-hand side of the gap as viewed in Fig. 2.

HEAT repeats are also found in the phosphatase 2A PR65/A subunit[19], importin-β[20], CAND1 (ref. 21) and many other proteins. In DNA-PKcs the repeats are structurally irregular (Fig. 2d), making it difficult to use known structures (Fig. 2e) to guide the modelling of the HEAT repeats. Consequently, alanine helices were initially placed in the clearly visible rods of electron density using COOT[22] and refined individually as rigid bodies in REFMAC[23]. The organization of the HEAT repeats leads to an inner and an outer layer of α-helices, giving a handedness to the overall fold of the polypeptide chain.

As noted earlier, the circular arrangement of the HEAT repeats has a gap. Most probably the polypeptide chain has its N terminus on one side of this gap, circumnavigates the ring and then reverses direction on the other (Fig. 2a). There are some particularly irregular regions; for example, at about 135° round the ring structure from the gap going anticlockwise (Fig. 2c) and again at around 225°. The portion of the structure on the opposite side to the gap in the ring structure lies between the two points of irregularity (135°–225°), and seems to be equivalent to the region identified as the forehead in electron microscopy studies[18]. The forehead and the residues around the gap flop towards one another, giving the ring structure a concave shape or cradle appearance when viewed from the side (Fig. 2b).

The two irregular helical regions are probably equivalent to the points of conformational flexibility suggested previously[24] on the basis of the electron microscopy studies. A conformational change could widen the gap, a movement resembling bent arms swinging apart, so providing a mechanism for DNA-PKcs release of DNA after ligation. The release is probably triggered by one of the two major phosphorylation clusters ABCDE or PQR that are thought to control DNA end processing[25]. The conformational changes would probably transmit to the head/crown that carries the kinase domain. Thus, the arrangement and size of the ring structure reflect the use of this part of the structure as a platform for proteins that engage in the repair of the broken DNA, and which together with Ku holds in place the DNA while it is being repaired.

After the chain reversal there is a much smaller globular domain, also organized as HEAT repeats, which represents a good candidate for DNA binding[18] (Fig. 2a, shown in cyan). DNA-PKcs can bind directly to DNA and become active in the absence of Ku70/Ku80 as demonstrated in earlier studies by electrophoretic mobility shift and atomic force microscopy studies[26]. This binding could take place through this domain. The region after the putative DNA-binding

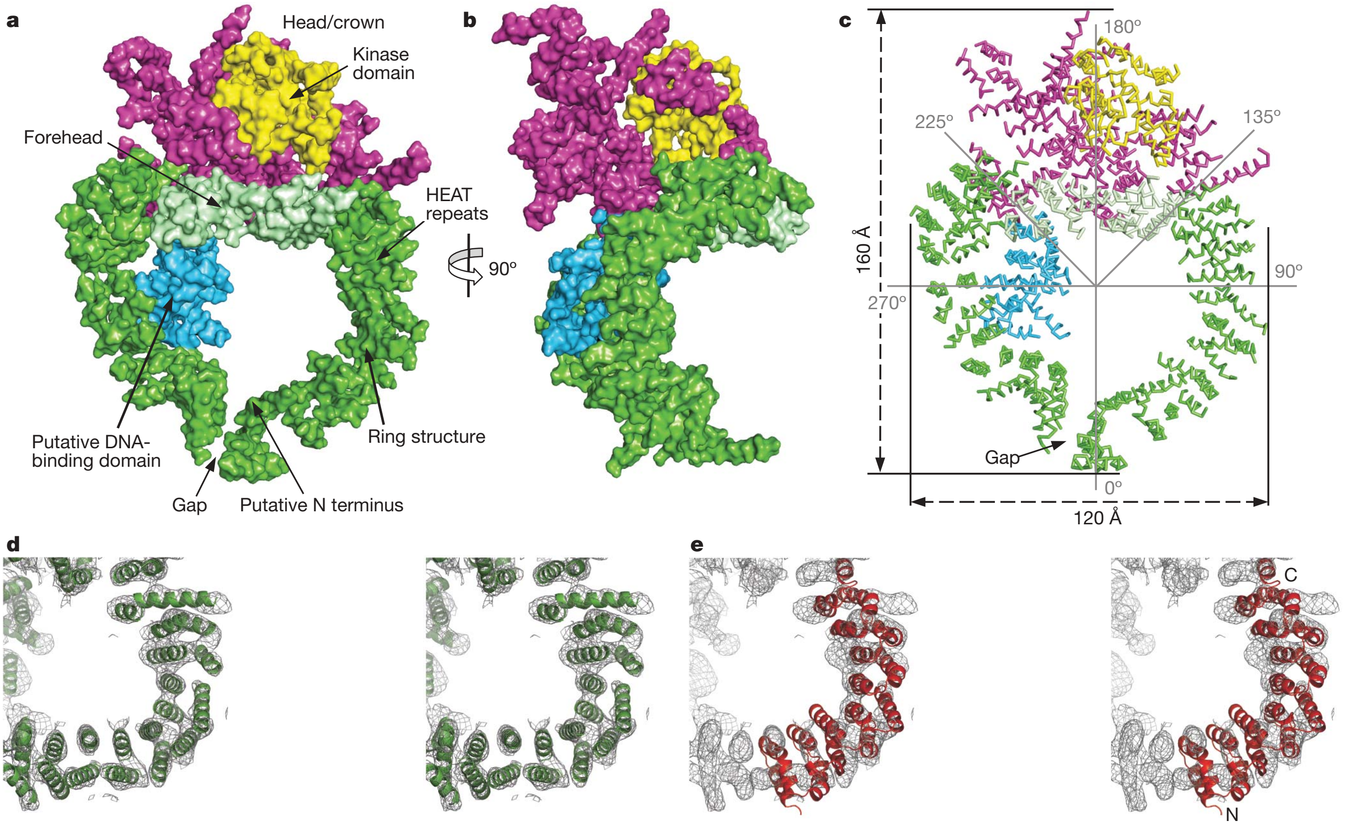


Figure 2 | Overall view of the DNA-PKcs structure. Molecular surface of DNA-PKcs. **a**, The front of the molecule. The colour-coding of the different parts of the molecule is as shown in Fig. 1a. **b**, The side view. **c**, Ribbon representation of the Cα positions of Fig. 2a, indicating the overall size and the sites that are potentially flexible at 135° and 225°. **d**, **e**, Stereo diagrams of (**d**) experimental F_o electron density map in the HEAT repeat area of the ring structure with the final DNA-PKcs model shown in green, and (**e**) HEAT repeats of the phosphatase 2A PR65/A subunit (residues 1–325; PDB accession code 1b3u) superposed onto the electron density of the DNA-PKcs structure. The N and C termini of the phosphatase 2A PR65/A subunit are shown. Both electron density maps are contoured at 1.0σ level.

domain and N-terminal to the kinase domain is likely to contain some of the sites that interact with other proteins. The electron density is predominantly rod like, indicating that this part is also α-helical with several of the helices organized into HEAT repeats—a feature that is observed throughout the DNA-PKcs structure.

We superposed the kinase structure from PI(3)Kγ[27] on the head/crown domain of the DNA-PKcs crystal structure, using COOT[22], and obtained a convincing fit to the β-strands of the N-lobe, which fit into the flat density present, as well as the helices that dominate the C-lobe as shown in Fig. 3. This provides unequivocal location of the kinase in the head/crown positioned above the ring structure. The FAT domain, a region of around 500 amino acids, N-terminal to the kinase domain[28] and named after the three main groups sharing this domain (FRAP (also known as MTOR) ATM and TRRAP), has been found only in PI(3)K-related kinases[28]. This family also has a much smaller, highly conserved domain at the extreme C terminus of their sequences that is 35 residues long called FATC[28], which is known to be α-helical[29]. These two domains are always found together, and they were proposed previously[28] to interact with the kinase domain wedged between them. This suggests that these are in the region shown in magenta with the kinase exposed at the very top accessible to substrates (Fig. 2).

As noted earlier, the crystal structure of DNA-PKcs was defined independently of the published electron microscopy structures[15–18,30], which differ quite radically between themselves. Nevertheless, our crystal structure has some features in common with the electron microscopy models, such as the head/crown as can be seen from the comparison in Supplementary Fig. 2. A further example is the putative DNA-binding domain, which was identified in the structure reported previously[18]. This can be seen in the crystal structure, although the base and further openings seen in some electron microscopy structures are not observed. The absence in the crystal structure of the extra openings that were thought to have a role in binding single-stranded DNA indicates that both double-stranded and single-stranded DNA may bind to DNA-PKcs through the putative DNA-binding domain.

Our crystal structure of DNA-PKcs brings into focus the distinct domains and their architectures, and reveals irregular regions of repetitive structures where conformational changes probably take place.

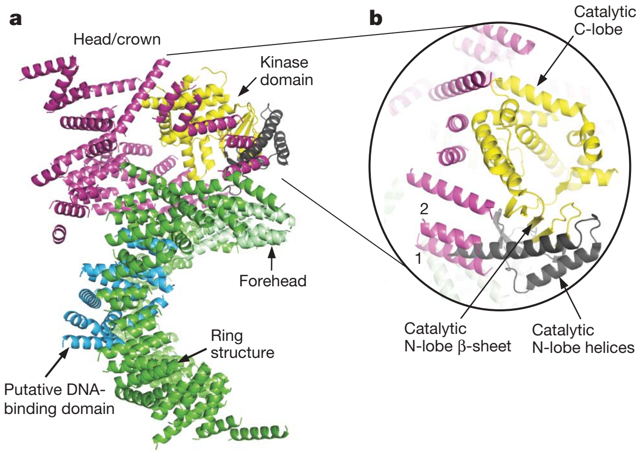


Figure 3 | Location of the DNA-PKcs kinase domain. **a**, The overall structure of DNA-PKcs showing the location of the kinase catalytic domain (yellow). The two helices of the N-lobe are shown in brown and were not built into the final structure of DNA-PKcs owing to unclear electron density in this area. The catalytic domain of DNA-PKcs was built on the basis of the crystal structure of PI(3)Kγ kinase (PDB accession code 1e8x). **b**, A close-up view of the DNA-PKcs kinase catalytic domain. Two helices of PI(3)Kγ kinase N-lobe could occupy the positions of helices 1 and 2 of DNA-PKcs shown in magenta.

METHODS SUMMARY

DNA-PKcs was isolated from HeLa S3 cells and purified using ion-exchange and size-exclusion chromatography. The Ku80ct_{194} and Ku80ct_{140} domains were over-expressed in *Escherichia coli* strain BL21 (DE3) and purified to homogeneity from the soluble lysates in four chromatographic steps. The vapour diffusion method using PEG 8000 as a precipitant was used to crystallize DNA-PKcs in complex with Ku80ct_{194} and Ku80ct_{140}. Diffraction data were collected at the European Synchrotron Radiation Facility (ESRF). The structure was solved using phases calculated by the MAD method with the tantalum bromide heavy atom cluster.

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

Received 29 July; accepted 6 November 2009.
Published online 20 December 2009.

1. Kemp, L. M., Sedgwick, S. G. & Jeggo, P. A. X-ray sensitive mutants of Chinese hamster ovary cells defective in double-strand break rejoining. *Mutat. Res.* **132**, 189–196 (1984).
2. Zdzienicka, M. Z., Tran, Q., van der Schans, G. P. & Simons, J. W. I. Characterization of an X-ray-hypersensitive mutant of V79 Chinese hamster cells. *Mutat. Res.* **194**, 239–249 (1988).
3. Biedermann, K. A., Sun, J., Giaccia, A. J., Tosto, L. M. & Brown, J. M. Scid mutation in mice confers hypersensitivity to ionizing radiation and a deficiency in DNA double-strand break repair. *Proc. Natl Acad. Sci. USA* **88**, 1394–1397 (1991).
4. Dvir, A., Stein, L. Y., Calore, B. L. & Dynan, W. S. Purification and characterization of a template associated protein kinase that phosphorylates RNA polymerase II. *J. Biol. Chem.* **268**, 10440–10447 (1993).
5. Carter, T., Vancurová, I., Sun, I., Lou, W. & DeLeon, S. A DNA-activated protein kinase from HeLa cell nuclei. *Mol. Cell. Biol.* **10**, 6460–6471 (1990).
6. Critchlow, S. E. & Jackson, S. P. DNA end-joining: from yeast to man. *Trends Biochem. Sci.* **23**, 394–398 (1998).
7. Gottlieb, T. M. & Jackson, S. P. The DNA-dependent protein kinase requirement for DNA ends and association with Ku antigen. *Cell* **72**, 131–142 (1993).
8. Ma, Y., Pannicke, U., Schwarz, K. & Lieber, M. R. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell* **108**, 781–794 (2002).
9. Hartley, K. O. *et al.* DNA-dependent protein kinase catalytic subunit: a relative of phosphatidylinositol 3-kinase and the ataxia telangiectasia gene product. *Cell* **82**, 849–856 (1995).
10. Hoekstra, M. F. Responses to DNA damage and regulation of cell cycle checkpoints by the ATM protein kinase family. *Curr. Opin. Genet. Dev.* **7**, 170–175 (1997).
11. Anderson, C. W. DNA damage and the DNA-activated protein kinase. *Trends Biochem. Sci.* **18**, 433–437 (1993).
12. Yoo, S. & Dynan, W. S. Geometry of a complex formed by double strand break repair proteins at a single DNA end: recruitment of DNA-PKcs induces inward translocation of Ku protein. *Nucleic Acids Res.* **27**, 4679–4686 (1999).
13. Harris, R. *et al.* The 3D solution structure of the C-terminal region of Ku86 (Ku86CTR). *J. Mol. Biol.* **335**, 573–582 (2004).
14. Zhang, Z. *et al.* Solution structure of the C-terminal domain of Ku80 suggests important sites for protein–protein interactions. *Structure* **12**, 495–502 (2004).
15. Chiu, C. Y., Cary, R. B., Chen, D. J., Peterson, S. R. & Stewart, P. L. Cryo-EM imaging of the catalytic subunit of the DNA-dependent protein kinase. *J. Mol. Biol.* **284**, 1075–1081 (1998).
16. Boskovic, J. *et al.* Visualization of DNA-induced conformational changes in the DNA repair kinase DNA-PKcs. *EMBO J.* **22**, 5875–5882 (2003).
17. Rivera-Calzada, A., Maman, J. P., Spagnolo, L., Pearl, L. H. & Llorca, O. Three-dimensional structure and regulation of the DNA-dependent protein kinase catalytic subunit (DNA-PKcs). *Structure* **13**, 243–255 (2005).
18. Williams, D. R., Lee, K.-J., Shi, J., Chen, D. J. & Stewart, P. L. Cryo-EM structure of the DNA-dependent protein kinase catalytic subunit at subnanometer resolution reveals α-helices and insight into DNA binding. *Structure* **16**, 468–477 (2008).
19. Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A. & Barford, D. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* **96**, 99–110 (1999).
20. Cingolani, G., Petosa, C., Weis, K. & Müller, C. W. Structure of importin-β bound to the IBB domain of importin-α. *Nature* **399**, 221–229 (1999).
21. Goldenberg, S. J. *et al.* Structure of the Cand1–Cul1–Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases. *Cell* **119**, 517–528 (2004).
22. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
23. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
24. Spagnolo, L., Rivera-Calzada, A., Pearl, L. H. & Llorca, O. Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. *Mol. Cell* **22**, 511–519 (2006).
25. Meek, K., Douglas, P., Cui, X., Ding, Q. & Lees-Miller, S. P. *trans* autophosphorylation at DNA-dependent protein kinase's two major autophosphorylation site clusters facilitates end processing but not end joining. *Mol. Cell. Biol.* **27**, 3881–3890 (2007).

26. Yaneva, M., Kowalewski, T. & Lieber, M. R. Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic force microscopy studies. *EMBO J.* **16,** 5098–5112 (1997).
27. Walker, E. H., Perisic, O., Ried, C., Stephens, L. & Williams, R. L. Structural insights into phosphoinositide 3-kinase catalysis and signaling. *Nature* **402,** 313–320 (1999).
28. Bosotti, R., Isacchi, A. & Sonnhammer, E. L. L. FAT: a novel domain in PIK-related kinases. *Trends Biochem. Sci.* **25,** 225–227 (2000).
29. Dames, S. A., Mulet, J. M., Rathgeb-Szabo, K., Hall, M. N. & Grzesiek, S. The solution structure of the FATC Domain of the protein kinase TOR suggests a role for redox-dependent structural and cellular stability. *J. Biol. Chem.* **280,** 20558–20564 (2005).
30. Leuther, K. K., Hammarsten, O., Kornberg, R. D. & Chu, G. Structure of DNA-dependent protein kinase: implications for its regulation by DNA. *EMBO J.* **18,** 1114–1123 (1999).

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

Acknowledgements We wish to thank G. Smith for providing pure DNA-PKcs that was used in the initial experiments as a marker and also for providing the DNA-PKcs and Ku70 antibodies, L. Pellegrini for help and advice at the beginning of the project, and S. Jackson for discussions. We are also grateful to R. Peat for preparing HeLa cells, L. Packman for help in identifying DNA-PKcs in polyacrylamide gels, N. Ban for providing $Ta_6Br_{12}{}^{2+}$ and C. Müller-Dieckmann at the ESRF for support during the diffraction data collection experiments. This work was funded by the Wellcome Trust and CR-UK.

Author Contributions T.L.B. and B.L.S. conceived the project. B.L.S. characterized, purified, crystallized and analysed the electron density for the DNA-PKcs–Ku80ct complex. D.Y.C. and B.L.S. carried out data collection and structure modelling. D.Y.C. carried out data processing, electron density calculations and refinement, with significant input from T.L.B. in the interpretation of the data. B.L.S. wrote the paper and all authors contributed and edited the manuscript.

Author Information The atomic coordinates and structure factors for the reported crystal structure have been deposited with the Protein Data Bank (PDB) under accession code 3kgv. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.L.S. (lynn@cryst.bioc.cam.ac.uk) or D.Y.C. (dima@cryst.bioc.cam.ac.uk).

METHODS

DNA-PKcs purification. DNA-PKcs isolated from HeLa S3 cells was purified to near homogeneity using a modified protocol described previously[31]. All steps were carried out at 4 °C. HeLa cells were purchased from Cancer Research UK in frozen-pellet form. HeLa cell nuclear extract was prepared as described[32], with the protease inhibitor tablets added at the nuclear extraction stage. The prepared nuclear extract was dialysed in 20 mM HEPES, pH 7.6, 100 mM NaCl, 10% glycerol, 0.5 mM EDTA, 2 mM $MgCl_2$, 5 mM dithiothreitol (DTT), 0.2 mM phenylmethylsulphonyl fluoride (PMSF) and fractionated using standard chromatographic procedures beginning with Q-sepharose followed by heparin agarose column. Fractions containing DNA-PKcs were further purified using Mono S and Mono Q ion-exchange columns. DNA-PKcs was eluted from these columns using a linear gradient of 0.1–1 M NaCl. DNA-PKcs-containing fractions were dialysed in the above buffer before each step. Superose-6 gel filtration column was used as the final purification step. Protein purity was judged by SDS–PAGE. Pure DNA-PKcs used as a marker was a gift from G. Smith. The purified protein was further confirmed as DNA-PKcs using mass spectrometry (proteins identified in polyacrylamide gels) by The Protein and Nucleic Acid Chemistry Facility (PNAC) in Cambridge. Pure protein was immediately used for crystallization experiments or stored in aliquots at −80 °C. All columns were purchased from Amersham Biosciences.

Ku80 C-terminal domain expression and purification. Two small Ku80 C-terminal domains spanning residues 539–732 ($Ku80ct_{194}$) and 593–732 ($Ku80ct_{140}$) were cloned into pGAT3 with a 6×His-tag at the N terminus and fused to glutathione *S*-transferase (GST). These were overexpressed in *E. coli* strain BL21(DE3), and the soluble lysates were purified to homogeneity in four chromatographic steps. In the first step, the cloned domain together with GST was isolated using Ni-NTA affinity chromatography. In the second step, the 6×His-tag and the GST were removed using tobacco ect virus (TEV) protease. The subsequent two steps were ion-exchange chromatography using a Mono Q column with a NaCl gradient of 0–1 M in 20 mM Tris buffer at pH 8.0, and size-exclusion chromatography using Superdex-200 in 20 mM Tris buffer at pH 8.0.

Crystallization. DNA-PKcs was crystallized using the vapour diffusion method in hanging drops. Extensive optimization was required to produce crystals suitable for X-ray diffraction analysis. The two buffer conditions that produced best crystals were: (1) 0.1 M Bis-Tris, 200 mM NaCl, 5 mM DTT in the presence of 8% PEG 8000 (w/v), and (2) 0.1 M Bis-Tris, 200 mM NaCl, 30% glycerol, 10 mM DTT, 5 mM EDTA in the presence of 18% PEG 8000 (w/v). The pH of these buffers varied from 6.2 to 6.7. Hanging drops were prepared by mixing a 6.4 mg ml^{-1} protein sample in a 1:1 ratio with the buffer solution containing PEG 8000 used in the reservoir. Improvement in crystal quality resulted upon forming complexes of DNA-PKcs with either of two Ku80 C-terminal fragments. These were mixed in the 1:3 ratio of DNA-PKcs–Ku80ct, and crystallized using conditions described in (1) and (2). The crystals using condition (1) required 26% ethylene glycol for cryo-protection, whereas those from condition (2) were directly flash-frozen with 30% glycerol present in the crystallization buffer providing cryo-protection. The complex with $Ku80ct_{194}$, which diffracted to 6.6 Å resolution (Supplementary Table) provided the data with which the density maps were calculated. An SDS–PAGE gel of the dissolved crystals (Supplementary Fig. 1) confirms the presence of DNA-PKcs and the $Ku80ct_{194}$ domain.

Heavy metal crystal soaks. For obtaining the phasing information, the crystals were soaked in a variety of conventional heavy metal atom solutions. Of these, dodeca-μ-bromo-hexatantalum dibromide ($Ta_6Br_{12}{}^{2+}$)[33] gave a deep green colour—a good indication of cluster incorporation into the crystal lattice. The soaking was carried out overnight at 1.1 mM concentration of $Ta_6Br_{12}{}^{2+}$ in the crystallization buffer. The crystals were then back-soaked for an hour before flash-freezing in liquid nitrogen. These crystals were used to obtain the phase information using the MAD method that led to identification of the structure of the DNA-PKcs–$Ku80ct_{194}$ complex.

Data collection, phasing and refinement. The diffraction data collection on $Ta_6\ Br_{12}{}^{2+}$ derivative crystals was carried out at 100 K using a Quantum 315R CCD area detector (Area Detector Systems Corporation) on the ID29 beamline at ESRF (Grenoble, France). X-ray fluorescence[34] (XRF) spectroscopy confirmed the presence of tantalum atoms in the crystals. The MAD X-ray diffraction data sets were collected at the absorption peak (wavelength 1.2549 Å) and at the inflection point of the Ta L3 edge (wavelength 1.2552 Å). The data sets were collected from one crystal using one degree oscillation, each of the data sets had 180 degrees of data, the appropriate data collection segments were calculated using MOSFLM's STRATEGY routine[35]. The data sets were processed and scaled using DENZO and SCALEPACK[36]. The crystal belongs to monoclinic $P2_1$ space group and has two molecules of the DNA-PKcs–$Ku80ct_{194}$ complex in the asymmetric unit resulting in crystal's solvent content of about 63%. The presence of the two molecules in the asymmetric unit was also confirmed by the presence of NCS two-fold axis clearly visible on self-rotation function plots calculated using POLARRFN from the *CCP*4 program suite[37].

The initial set of phases was determined by MAD method using tantalum atoms as anomalous scatterers. A total of ten $Ta_6Br_{12}{}^{2+}$ cluster sites were identified using PHENIX's AutoSol routine[38]. A resolution limit of 7.1 Å was used in the calculation—the highest resolution limit of the 'inflection' data set. This PHENIX's score of the solution was 75.6 and the figure merit of phases calculated by SOLVE was 0.54. Assuming the $Ta_6Br_{12}{}^{2+}$ cluster to be a single atom, the positional parameters and B factors were refined using SHARP[39]. The phases were refined and extended from 7.1 to 6.6 Å resolution by PHENIX, solvent flattening and histogram matching were carried out using DM from the *CCP*4 program suite[37]. Two-fold NCS electron density averaging was applied but did not improve the quality of the maps, probably because the electron density for one of the two molecules of the DNA-PKcs–$Ku\text{-}80ct_{194}$ complex in the asymmetric unit is less well defined possibly owing to thermal disorder.

A structural model of DNA-PKcs was built into the electron density maps calculated at 6.6 Å resolution. Highly idealized alanine α-helices of various lengths and curvatures could be fitted into the electron density maps using COOT[22]. The direction of α-helices was often ambiguous, although in the HEAT repeat areas, a consensus of helical directionality was sustained with helices built in one direction on the outer side of the HEAT repeat and in the opposite direction on the inner side. After the first round of rebuilding a total of 188 helices of various lengths were built (121 in one DNA-PKcs molecule and 67 in the other). Close visual inspection of the electron density maps in the area of kinase domain showed a set of three helices, the orientation of which was similar to that of the catalytic domain of PI(3)Kγ kinase[27] (PDB accession code 1e8x). A real space fitting of the catalytic domain of PI(3)Kγ kinase into the DNA-PKcs electron density positioned it exactly in the visually identified area. Using built helices of the two DNA-PKcs molecules in the asymmetric unit, the two-fold NCS operator was identified using LSQMAN of the RAVE program suite[40]. This was then used to reconstruct the incomplete model of the second DNA-PKcs molecule in the asymmetric unit using program XPAND of the RAVE program suite[40]. The final model produced an R/R_{free} of 52.0/53.8%, respectively, calculated against the absorption peak data set of 6.6 Å resolution.

The crystallographic refinement calculations were carried out using REFMAC5 of the *CCP*4 program suite[37]. Before the first run of the refinement, all atomic B factors of the model were set to 80 $Å^2$. A typical refinement protocol consisted of rigid body refinement procedure including the phases obtained from the MAD calculations. The rigid body groups included individual helices, and no NCS restraints or constraints were used. The model was then analysed visually in COOT and the helices were trimmed down and/or rigid-body translated/rotated to fit the resulting SigmaA-weighted $2F_o - F_c$, $F_o - F_c$ maps. A total of five such refinement/rebuilding stages were carried out, giving final R/R_{free} values of 44.2/44.1%, respectively (Supplementary Table).

All figures were generated using Pymol (http://pymol.sourceforge.net).

31. Gell, D. & Jackson, S. P. Mapping of protein-protein interactions within the DNA-dependent protein kinase complex. *Nucleic Acids Res.* **27**, 3494–3502 (1999).
32. Ausubel, F. M, *et al.* in *Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology* 5th edn, 12.3–12.6 (Wiley, 2002).
33. Schneider, G. & Lindqvist, Y. Ta6BrI4 is a useful cluster compound for isomorphous replacement in protein crystallography. *Acta Crystallogr. D* **50**, 186–191 (1994).
34. Leonard, G. A. *et al.* Online collection and analysis of X-ray fluorescence spectra on the macromolecular crystallography beamlines of the ESRF. *J. Appl. Crystallogr.* **42**, 333–335 (2009).
35. Leslie, A. G. W. Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EAMCB Newsletter on Protein Crystallography* **no. 26** (1992).
36. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
37. Collaborative Computational Project, Number 4. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
38. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
39. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D* **59**, 2023–2030 (2003).
40. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52**, 842–857 (1996).

ERRATUM

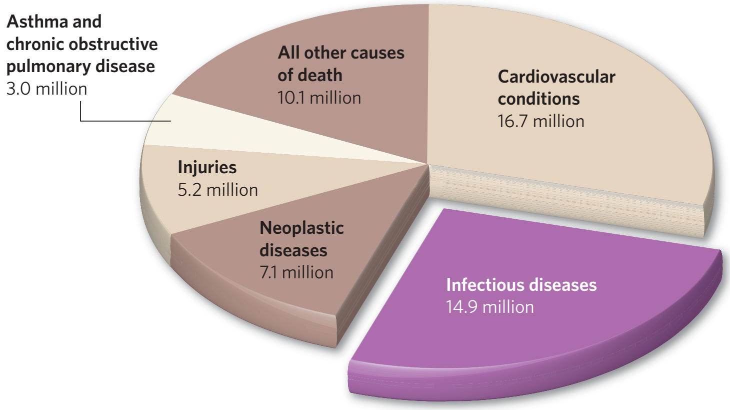
doi:10.1038/nature08554

The challenge of emerging and re-emerging infectious diseases

D. M. Morens, G. K. Folkers & A. S. Fauci

Nature 430, 242–249 (2004)

In Figure 2 of this Review, the pie chart was drawn incorrectly, with the wedge sizes not in proportion to the total size. The correct figure is shown below.



Infectious diseases	Annual deaths (millions)
Respiratory infections	3.96
HIV/AIDS	2.77
Diarrhoeal diseases	1.80
Tuberculosis	1.56
Vaccine-preventable childhood diseases	1.12
Malaria	1.27
STDs (other than HIV)	0.18
Meningitis	0.17
Hepatitis B and C	0.16
Tropical parasitic diseases	0.13
Dengue	0.02
Other infectious diseases	1.76

PROSPECTS

From geek to chic

Many stereotypes should be crushed, but some can prove beneficial to a fledgling scientist, says Peter Fiske.

In November, President Barack Obama held a news conference to announce a new national science fair. "Scientists and engineers ought to stand side by side with athletes and entertainers as role models, and here at the White House, we're going to lead by example," he said. "We're going to show young people how cool science can be."

The thought that scientists and engineers could one day be elevated in social stature to the level of pop-culture icons is a tantalizing prospect. Of course, one has only to turn on the television or watch any number of movies to see how society really imagines them. Although the mainstream media might at times portray scientists as powerful and villainous, a PhD at the end of a name tends to conjure up images of pallor and social awkwardness — people more to be pitied than feared. In his book *Mad, Bad and Dangerous? The Scientist and the Cinema* (Reaktion Books, 2005), Christopher Frayling notes that surveys of attitudes over the past 50 years have shown that the cultural stereotype surrounding 'scientist' has been largely consistent — and negative.

It is easy for scientists and engineers to resign themselves to some degree of social stigmatization or even to embrace an in-your-face uber-geek image as a badge of honour. Brisk sales of Garth Sundem's book *The Geeks' Guide to World Domination: Be Afraid, Beautiful People* (Three Rivers Press, 2009) imply the notion has broad appeal. But the bizarre image that science degrees can conjure in some people's minds affects more than just dating prospects: how people perceive scientists and engineers affects what they think people in these professions can do and, ultimately, what their value is to society. It certainly affects the range of career options available to them.

Perhaps, by unpacking the cultural stereotypes that surround 'scientist' or 'PhD', scientists will be able to understand how some stereotypes work against them — and that may actually work to their advantage. This can be useful for young scientists who hope to advance, especially those whose non-traditional career paths involve frequent interactions with non-science professionals.

Judging a book by its cover

The physical stereotype of the scientist (male or female) is hardly flattering. Comically unkempt, poorly dressed, wearing any number of geek accoutrements (pocket protector, slide rule, calculator), the stereotypical scientist is out of touch with his or her surroundings, and compulsively obsessed with the science.

This stereotype is the easiest to address. From my attendance at many science meetings and my business dealings with scientists in academia, government and the private sector, people with PhDs seem to be indistinguishable from the population at large when they are not wearing a lab coat. It is true that many research work environments have few if any dress codes, and attire, especially in graduate school, can become quite casual. But often it is not social ineptitude that prevents young scientists from dressing better; it is poverty.

SUNSET BOULEVARD/CORBIS

Still, young scientists sometimes fail to appreciate how their attire can limit their opportunities outside the lab. Fortunately, the rules of business dress and etiquette are much easier to master than the topics they encounter in graduate school. Polishing their appearance at conferences or when interacting with visitors can help young scientists to be taken seriously. Who knows? Potential contacts may even mistake graduate students for faculty members.

Scientists and engineers face more serious prejudices, especially in the business world. Many business managers fear that PhDs are simple-minded about money, impractical about time, have no sense of deadlines and are uncompromisingly idealistic. Where do these stereotypes come from? To be fair, obtaining a PhD does demand an intensity of study and a single-mindedness of focus, even a degree of obsessive compulsion. This may be helpful for a research career that focuses on a single topic. But in most other work environments, employers are looking for people who are leaders, not hermits; team players, not arrogant loners. In a job interview, scientists should be sure to work in stories about how they worked successfully in teams, have led activities and enjoy working with others.

Perhaps the most self-limiting stereotype about PhD scientists and what they are capable of comes from the scientists themselves. After spending so many years obtaining an advanced degree in a particular field of study, scientists understandably value their technical skills the most. And, having become so highly qualified in one field, they often feel totally unqualified to address issues that may exist in other fields or industries. When scientists look for employment they naturally frame every opportunity in terms of their field of study and their expertise: chemists look for jobs labelled chemist, biologists look for jobs labelled biologist.

The bigger picture

In truth, the training to become a scientist or an engineer comes with a long list of transferable skills that are of enormous value in the 'outside world'. Communication skills, analytical skills, independence, problem-solving skills, learning ability — these are all valuable. But scientists and engineers tend to discount these things because they are basic requirements of their profession. They tend to think of themselves as subject-matter experts rather than as broadly adaptable problem solvers. Unfortunately, the world needs a lot more of the latter than the former.

These more serious stereotypes deserve attention as scientists focus on advancing their careers. In networking opportunities, and especially during job interviews outside academia, it is particularly important to address the potential negative stereotypes that the wider world may harbour about a PhD. Instead of focusing the discussion on science and technical abilities, scientists should educate people about the leadership roles they have had, or the broad range of roles that they fulfil in the lab, or their interest in contributing to a bigger goal than just their own research.

A different sort of pervasive stereotype may prove much more beneficial — the scientist or engineer as a genius. Cartoons of the stereotypical science geek often show a thought bubble filled with equations and formulae hovering over their heads. People see a PhD after a name, and they assume that the person is a genius — a rocket scientist, even a reincarnation of Einstein himself.

This can be a good thing. By addressing the negative stereotypes, then dispelling them, scientists will leave their audience with the 'scientist as genius' label. As they strive to advance their careers, scientists should try not to dispel that misconception. This is one stereotype, after all, that fledgling scientists can use to great advantage. ■

Peter Fiske is chief technology officer of PAX Water Technologies in San Rafael, California, and author of *Put Your Science to WORK*.

Brief lullaby

A song from the stars.

Val Nolan

At first it is a whisper, a voice calling from the dark. The technician who records it considers it an aberration: a glitch, an unknown astronomical phenomenon or, at worst, a prank. Although she plays it to her colleagues for a day or two, no one expects the song to reappear. But it does.

It is a voice from another world, hundreds of thousands of years old if triangulated accurately. Meticulous investigation follows; months of study during which recordings of the song are bootlegged, leaked and then grudgingly confirmed. Governments issue statements. Papal edicts grapple with the implications. *The Times* proclaims four words that change our understanding of the Universe forever: WE ARE NOT ALONE.

It could not be kept a secret anyway, for as it strengthens the signal breaks through on ordinary radios, in the middle of the evening news or at a crucial moment in a drama. What panic ensues is soothed by the song itself, swelling to a dozen voices and then a hundred — an *a capella* polyphony out of which a soloist emerges sweet and soulful, lofting over his fellows. It is the music of a world bereft of instruments. Or more correctly, a world where voice itself is the instrument of choice. Who can forget their first time hearing how these hymns resolve from the sky's cacophony, those myriad cadences melting in and out of one another? "Where were you …?" becomes a question around dinner tables, a way of introduction.

Science meanwhile strives to find the song's significance, seeking answers in its harmony or line. Some go mad, convinced we are meant to understand these grand, mercurial and vaguely sentimental themes as blueprints for the Universe itself, although most agree it is an accidental broadcast. The song is deemed to be the remnants of a transmission bubble, radio waves emanating from far beyond the Solar System and which now wash over what we thought of once as all creation. Only in the deaf season, when Earth slinks guiltily behind the Sun, do we lose the signal, although every spring we re-emerge from the shadow of the star and there it is: deepened, developed somehow. Its variations happier or sadder as the years go by, freed from its earlier constraints as we imagine its society must be.

Decades pass and powerful movements take hold. Sects emerge believing this to be the work of God. They propagate like signals in the dark and their teachings transcend traditional interpretations of Jewish or Christian or Muslim faith; they believe the choir of angels sings to all, and it is difficult to argue. Humanity finds purpose and solidarity as old philosophies and politics begin to wither. Green parties gain ground. The first and third worlds strive to close the gap. Climate change is stabilized and people start talking of Utopia, although this remains a long way off.

JACEY

Beyond the reach of copyright, the alien signal airs on cheap wireless sets and on the Internet. A generation comes of age for whom it is the foremost influence. Songwriters and musicians begin to meld their styles to it, first sampling, then imitating, then incorporating freely. Musicologists become astronomers and physicists. Physicists write musicals and croon in deserted cabarets. Classical music, jazz, rock, pop and rap all listen and evolve. A hybrid form is born and what began as wonder now becomes passé. The signal is just another broadcast, just another channel on the dial. The song itself is followed less and less.

That is until it starts to grow chaotic, its strength diminishing and its voices starting to go silent. Fanatics blame our own complacency and immolate themselves. Sociologists and radio astronomers struggle to explain the change. Analysis abounds. The leading theory holds that these developments record a period of conflict in the song's society. Soon the aliens resume their earlier celebrity and their operatic struggle takes hold of our imaginations. Their war is reported as though it were our own, interpretations of a language we have never learnt to understand.

In the course of five short years, static comes to dominate the otherworldly frequencies until the song begins to fade completely. The liveliness of earlier is spent. War has consumed their world. Sporadic, forlorn dirges sing to us of this, although these meek voices have the ear of only telescopes and satellites. They are relayed to great gatherings in the cities of the world, vigils that are held until the song dies out. Candlelight illuminates whole continents. A period of mourning is declared. *The Times*, with an eye to symmetry, reports on all of this in terms that emphasize our existential loneliness.

Mere histories or monuments cannot convey our grief. We mourn the passing of these others with songs we sing ourselves, that amalgam we have made our own, that blend of human music and unearthly voice. Heartened for a hundred years, we might be all alone once more but we shall not forget. Our long-dead cousins will survive in us, in this melodic dispatch towards the stars. Somewhere, a hundred or a hundred thousand years from now, we hope this message is received. ■

Val Nolan is a graduate of the Clarion writing programme at the University of California, San Diego. He currently teaches in the Department of English at National University of Ireland, Galway.

Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a

Close supermassive binary black holes

Arising from: T. A. Boroson & T. R. Lauer *Nature* **458,** 53–55 (2009)

It has been proposed that when the peaks of the broad emission lines in active galactic nuclei (AGNs) are significantly blueshifted or redshifted from the systemic velocity of the host galaxy, this could be a consequence of orbital motion of a supermassive black-hole binary (SMBB)[1]. The AGN J1536+0441 (=SDSS J153636.22+044127.0) has recently been proposed as an example of this phenomenon[2]. It is proposed here instead that J1536+0441 is an example of line emission from a disk. If this is correct, the lack of clear optical spectral evidence for close SMBBs is significant, and argues either that the merging of close SMBBs is much faster than has generally been hitherto thought, or if the approach is slow, that when the separation of the binary is comparable to the size of the torus and broad-line region, the feeding of the black holes is disrupted.

Galaxies grow through mergers, and as all massive galaxies contain supermassive black holes, the formation of SMBBs will be common[1,3,4]. Close SMBBs, with orbital velocities ∼0.01c, are expected to last long enough to be observed[1,3]. It was proposed[1] that such close binaries could be detected from velocity shifts of their broad-line region (BLR) line profiles, and pointed out that velocity shifts of the peaks of low-ionization broad lines are common[1,4,5].

There are two testable predictions of this model: first, the radial velocities of the peaks in the line profiles will shift on the orbital timescale of the SMBB[6], and second, as all AGNs vary, if there are two separate BLRs, the line fluxes of the two peaks will vary independently[7]. The prototypical displaced BLR peak AGN 3C 390.3 fails both these tests. The radial velocity signature of orbital motion has been detected[6,8] but the velocity changes are incompatible with an SMBB and are instead consistent with orbital motion of features in a non-azimuthally-symmetric disk. The peaks in 3C 390.3 initially seemed to be varying independently on timescales longer than the light-crossing timescale[7], but an SMBB is strongly ruled out by better-sampled monitoring[9,10], which shows that on a light-crossing time the peaks vary simultaneously as expected for a disk. The longer timescale profile changes in these AGNs are consistent with orbital motion of clumps in a disk[11]. The profiles are consistent with theoretical line emission profiles expected from disks[12], but the disks are generally not azimuthally symmetric and it is common for one peak to be significantly stronger than the other[1,4,5,12].

The Balmer lines of J1536+0441 have a strong blueshifted peak[2]. The Hβ profile is shown in Fig. 1. Boroson and Lauer[2,13] interpret J1536+0441 as a SMBB. I argue here, however, that just as the Balmer line profiles in previous SMBB candidates have been shown to be due to disk emission, so too the Balmer line profile of J1536+0441 probably arises from disk emission. This has also been independently proposed[14,15].

Although J1536+0441 represents an extremum among AGNs selected as 'quasars' by the SDSS, its line profiles are not unique among AGNs in general. The AGN 0945+076, for example, has shown a nearly identical Hβ profile (compare Fig. 1 with the Hβ profile of 0945+076[1]). To qualitatively illustrate the non-uniqueness of J1536+0441, Fig. 1 includes part of the scaled Hβ profile of the well-known disk emitter Arp 102B. Fitting disk models to theoretical double-peaked profiles invariably shows that there is an extra component of gas at the systemic velocity[12], so the spectral region with a width corresponding to that of the high-velocity wings of the [O III] 5,007 Å line has been excluded around rest-frame Hβ. The width of the Arp 102B spectrum has been reduced by 34% to make the velocity of the blueshifted peak in J1546+0441 match that of the blueshifted peak in the mean Arp 102B spectrum. This is well within the range of line widths seen among disk-like emitters. The uncertainty in the width scaling is about ±5%. The flux has been scaled by minimizing residuals away from the peaks of the broad lines and away from the contaminating [O III] 5,007 Å lines. The profiles of broad disk-like lines vary strongly, so we do not expect a perfect match between any two objects a given time. The peaks of Arp 102B are particularly variable[11]. The differences in the peak fluxes are only about 1σ from the scaled Arp 102B mean profile (Fig. 1). More recent spectra of J1546+0441[13,15] (especially of Hα) show better agreement in the red peak.

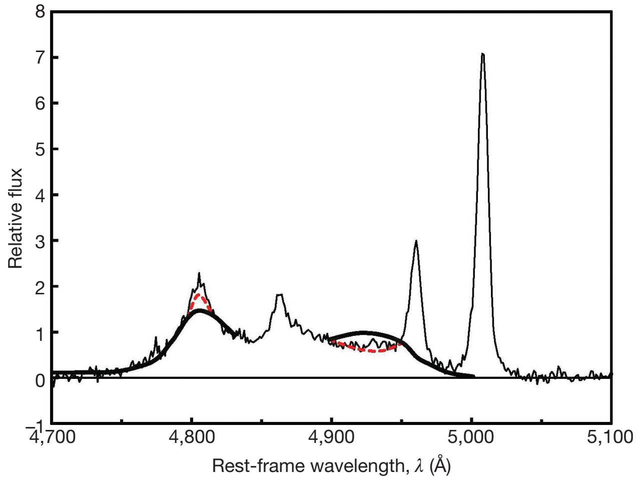


Figure 1 | Comparison of the continuum-subtracted SDSS spectrum of the Hβ region of J1536+0441 (thin black line) with the mean Hα profile of Arp 102B from 1992 to 1996 (smooth thick black line). The latter is taken from the mean spectrum in figure 3 of ref. 11 and scaled as described the text. The dotted red lines illustrate the effects on the peaks in the Arp 102B profile if they are changed by 1σ (based on the r.m.s. spectrum in figure 4b of ref. 11). The contribution of lower-velocity gas at the systemic velocity in Arp 102B has been omitted.

Clearly, the most rigorous test of the competing hypotheses is line profile variability[2,13]. If this verifies that the Hβ profile of J1536+0441 is the result of normal disk emission, it has significant implications for the evolution of SMBBs, as J1536+0441 is the only candidate so far for a sub-parsec SMBB out of ∼17,500 AGNs with $z < 0.70$ in the SDSS[2,13]. Because SMBB formation must be common, the absence of clear evidence for close SMBBs in AGNs needs to be explained. It suggests either that the lifetime of close SMBBs is considerable shorter than originally thought, or that they are long-lived and have their feeding interrupted so that activity is greatly reduced.

C. Martin Gaskell[1]
[1]Astronomy Department, University of Texas, Austin, Texas 78712, USA.
e-mail: gaskell@astro.as.utexas.edu

Received 25 March; accepted 8 October 2009.

1. Gaskell, C. M. Quasars as supermassive binaries. *Liége Int. Astrophys. Colloq.* **24,** 473–477 (1983).
2. Boroson, T. A. & Lauer, T. R. A candidate sub-parsec supermassive binary black hole system. *Nature* **458,** 53–55 (2009).
3. Begelman, M. C., Blandford, R. D. & Rees, M. J. Massive black hole binaries in active galactic nuclei. *Nature* **287,** 307–309 (1980).
4. Gaskell, C. M. in *Jets from Stars and Galactic Nuclei* (ed. Kundt, W.) 165–195 (Springer, 1996).
5. Gaskell, C. M. & Snedden, S. A. The optical case for a disk component of BLR emission. *ASP Conf. Ser.* **175,** 157–162 (1999).
6. Gaskell, C. M. Evidence for binary orbital motion of a quasar broad-line region. *Astrophys. J.* **464,** L107–L110 (1996).
7. Gaskell, C. M. in *Active Galactic Nuclei* (eds Miller, H. R. & Wiita, P. J.) 61–67 (Springer, 1988).
8. Eracleous, M., Halpern, J. P., Gilbert, A. M., Newman, J. A. & Filippenko, A. V. Rejection of the binary broad-line region interpretation of double-peaked emission lines in three active galactic nuclei. *Astrophys. J.* **490,** 216–226 (1997).
9. Dietrich, M. *et al.* Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. XII. Ground-based monitoring of 3C 390.3. *Astrophys. J.* **115** (Suppl.), 185–202 (1998).

10. O'Brien, P. T. *et al.* Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. XIII. Ultraviolet observations of the radio-galaxy 3C 390.3. *Astrophys. J.* **509,** 163–176 (1998).
11. Sergeev, S. G., Pronik, V. I. & Sergeeva, E. S. Arp 102B: variability patterns of the Hα line profile as evidence for gas rotation in the broad-line region. *Astron. Astrophys.* **356,** 41–49 (2000).
12. Eracleous, M. & Halpern, J. P. Completion of a survey and detailed study of double-peaked emission lines in radio-loud active galactic nuclei. *Astrophys. J.* **599,** 886–908 (2003).
13. Lauer, T. & Boroson, T. HST images and KPNO spectroscopy of the binary black hole candidate SDSS J153636.22+044127.0. *Astrophys. J.* **703,** 930–938 (2009).
14. Chornock, R. *et al.* SDSS J1536+0441: an extreme ''double-peaked emitter,'' not a binary black hole. *Astron. Telegr.* **1955** (2009).
15. Chornock, R. *et al.* The quasar SDSS J1536+0441: an unusual double-peaked emitter. Preprint at ⟨http://arXiv.org/abs/0906.0849⟩ (2009).

Competing financial interests: declared none.

doi:10.1038/nature08665

Boroson and Lauer reply

Replying to: C. M. Gaskell *Nature* 463, doi:10.1038/nature08665 (2010)

Gaskell[1] makes the point that the profile of the Balmer lines in the candidate binary supermassive black hole[2] AGN J1536+0441 bears some similarity to those that are thought to arise from disk emission in other objects. He further argues that two predictions of the competing binary black-hole hypothesis are (1) radial velocity changes that are interpretable in terms of the binary orbit, and (2) line fluxes in the two peaks that vary independently.

The similarity to disk emitters has been bolstered by subsequent spectroscopic observations[3,4], which show a red extension to the line profiles. However, the degree of similarity is debatable; the sharpness and strength of the blue peak and the presence of a central broad component make the line profiles in J1536+0441 unusual, if not unique. When sharp peaks are seen in objects thought to be disk emitters, these peaks are transient. As the simple disk emission models do not reproduce these characteristics at the extreme levels seen in J1536+0441[5], it is unclear whether their presence excludes the disk interpretation or not.

It is certainly true that the strongest evidence for determining the nature of this object will come from spectroscopic monitoring to detect radial velocity changes and flux changes within the line profiles. Initial evidence is inconclusive[3,4], in that no changes have been detected over a little less than one year. This finding excludes some of the orbital parameter space, but not that interpretation. Several more years of unchanging fluxes and velocities would push both explanations into uncomfortable areas, as the material in the disk must be orbiting as well.

It may be that J1536+0441 is a single AGN with an emission line profile resulting from a disk configuration of material. There are credible arguments for and against that hypothesis as well as the possibility that the system is a bound system of two supermassive black holes. Whichever the case, Gaskell is correct that the presence of at most one such object out of a sample of 17,500 has interesting implications for the process by which supermassive black holes merge.

Todd A. Boroson[1] & Tod R. Lauer[1]
[1]National Optical Astronomy Observatory, Tucson, Arizona 85719, USA.
e-mail: tyb@noao.edu

1. Gaskell, C. M. Close supermassive binary black holes. *Nature* **463,** doi:10.1038/nature08665 (2010).
2. Boroson, T. A. & Lauer, T. R. A Candidate sub-parsec supermassive binary black-hole system. *Nature* **458,** 53–55 (2009).
3. Chornock, R. *et al.* The quasar SDSS J1536+0441: an unusual double-peaked emitter. Preprint at ⟨http://arXiv.org/abs/0906.0849⟩ (2009).
4. Lauer, T. R. & Boroson, T. A. HST images and KPNO spectroscopy of the binary black hole candidate SDSS J1536.22+044127.0. *Astrophys. J.* **703,** 930–938 (2009).
5. Gezari, S., Halpern, J. P. & Eracleous, M. Long-term profile variability of double-peaked emission lines in active galactic nuclei. *Astrophys. J.* **169** (Suppl.), 167–212 (2007).

doi:10.1038/nature08666

Close supermassive binary black holes

Arising from: T. A. Boroson & T. R. Lauer *Nature* **458,** 53–55 (2009)

It has been proposed that when the peaks of the broad emission lines in active galactic nuclei (AGNs) are significantly blueshifted or redshifted from the systemic velocity of the host galaxy, this could be a consequence of orbital motion of a supermassive black-hole binary (SMBB)[1]. The AGN J1536+0441 (=SDSS J153636.22+044127.0) has recently been proposed as an example of this phenomenon[2]. It is proposed here instead that J1536+0441 is an example of line emission from a disk. If this is correct, the lack of clear optical spectral evidence for close SMBBs is significant, and argues either that the merging of close SMBBs is much faster than has generally been hitherto thought, or if the approach is slow, that when the separation of the binary is comparable to the size of the torus and broad-line region, the feeding of the black holes is disrupted.

Galaxies grow through mergers, and as all massive galaxies contain supermassive black holes, the formation of SMBBs will be common[1,3,4]. Close SMBBs, with orbital velocities $\sim 0.01c$, are expected to last long enough to be observed[1,3]. It was proposed[1] that such close binaries could be detected from velocity shifts of their broad-line region (BLR) line profiles, and pointed out that velocity shifts of the peaks of low-ionization broad lines are common[1,4,5].

There are two testable predictions of this model: first, the radial velocities of the peaks in the line profiles will shift on the orbital timescale of the SMBB[6], and second, as all AGNs vary, if there are two separate BLRs, the line fluxes of the two peaks will vary independently[7]. The prototypical displaced BLR peak AGN 3C 390.3 fails both these tests. The radial velocity signature of orbital motion has been detected[6,8] but the velocity changes are incompatible with an SMBB and are instead consistent with orbital motion of features in a non-azimuthally-symmetric disk. The peaks in 3C 390.3 initially seemed to be varying independently on timescales longer than the light-crossing timescale[7], but an SMBB is strongly ruled out by better-sampled monitoring[9,10], which shows that on a light-crossing time the peaks vary simultaneously as expected for a disk. The longer timescale profile changes in these AGNs are consistent with orbital motion of clumps in a disk[11]. The profiles are consistent with theoretical line emission profiles expected from disks[12], but the disks are generally not azimuthally symmetric and it is common for one peak to be significantly stronger than the other[1,4,5,12].

The Balmer lines of J1536+0441 have a strong blueshifted peak[2]. The Hβ profile is shown in Fig. 1. Boroson and Lauer[2,13] interpret J1536+0441 as a SMBB. I argue here, however, that just as the Balmer line profiles in previous SMBB candidates have been shown to be due to disk emission, so too the Balmer line profile of J1536+0441 probably arises from disk emission. This has also been independently proposed[14,15].

Although J1536+0441 represents an extremum among AGNs selected as 'quasars' by the SDSS, its line profiles are not unique among AGNs in general. The AGN 0945+076, for example, has shown a nearly identical Hβ profile (compare Fig. 1 with the Hβ profile of 0945+076[1]). To qualitatively illustrate the non-uniqueness of J1536+0441, Fig. 1 includes part of the scaled Hβ profile of the well-known disk emitter Arp 102B. Fitting disk models to theoretical double-peaked profiles invariably shows that there is an extra component of gas at the systemic velocity[12], so the spectral region with a width corresponding to that of the high-velocity wings of the [O III] 5,007 Å line has been excluded around rest-frame Hβ. The width of the Arp 102B spectrum has been reduced by 34% to make the velocity of the blueshifted peak in J1546+0441 match that of the blueshifted peak in the mean Arp 102B spectrum. This is well within the range of line widths seen among disk-like emitters. The uncertainty in the width scaling is about ±5%. The flux has been scaled by minimizing residuals away from the peaks of the broad lines and away from the contaminating [O III] 5,007 Å lines. The profiles of broad disk-like lines vary strongly, so we do not expect a perfect match between any two objects a given time. The peaks of Arp 102B are particularly variable[11]. The differences in the peak fluxes are only about 1σ from the scaled Arp 102B mean profile (Fig. 1). More recent spectra of J1546+0441[13,15] (especially of Hα) show better agreement in the red peak.

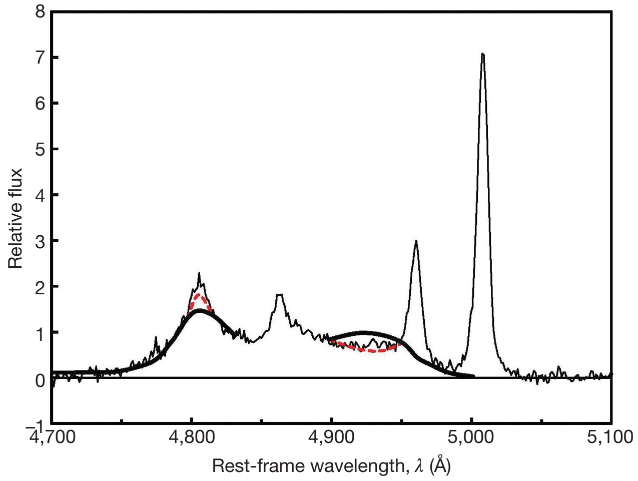


Figure 1 | Comparison of the continuum-subtracted SDSS spectrum of the Hβ region of J1536+0441 (thin black line) with the mean Hα profile of Arp 102B from 1992 to 1996 (smooth thick black line). The latter is taken from the mean spectrum in figure 3 of ref. 11 and scaled as described the text. The dotted red lines illustrate the effects on the peaks in the Arp 102B profile if they are changed by 1σ (based on the r.m.s. spectrum in figure 4b of ref. 11). The contribution of lower-velocity gas at the systemic velocity in Arp 102B has been omitted.

Clearly, the most rigorous test of the competing hypotheses is line profile variability[2,13]. If this verifies that the Hβ profile of J1536+0441 is the result of normal disk emission, it has significant implications for the evolution of SMBBs, as J1536+0441 is the only candidate so far for a sub-parsec SMBB out of ~17,500 AGNs with $z < 0.70$ in the SDSS[2,13]. Because SMBB formation must be common, the absence of clear evidence for close SMBBs in AGNs needs to be explained. It suggests either that the lifetime of close SMBBs is considerable shorter than originally thought, or that they are long-lived and have their feeding interrupted so that activity is greatly reduced.

C. Martin Gaskell[1]

[1]Astronomy Department, University of Texas, Austin, Texas 78712, USA.
e-mail: gaskell@astro.as.utexas.edu

Received 25 March; accepted 8 October 2009.

1. Gaskell, C. M. Quasars as supermassive binaries. *Liége Int. Astrophys. Colloq.* **24,** 473–477 (1983).
2. Boroson, T. A. & Lauer, T. R. A candidate sub-parsec supermassive binary black hole system. *Nature* **458,** 53–55 (2009).
3. Begelman, M. C., Blandford, R. D. & Rees, M. J. Massive black hole binaries in active galactic nuclei. *Nature* **287,** 307–309 (1980).
4. Gaskell, C. M. in *Jets from Stars and Galactic Nuclei* (ed. Kundt, W.) 165–195 (Springer, 1996).
5. Gaskell, C. M. & Snedden, S. A. The optical case for a disk component of BLR emission. *ASP Conf. Ser.* **175,** 157–162 (1999).
6. Gaskell, C. M. Evidence for binary orbital motion of a quasar broad-line region. *Astrophys. J.* **464,** L107–L110 (1996).
7. Gaskell, C. M. in *Active Galactic Nuclei* (eds Miller, H. R. & Wiita, P. J.) 61–67 (Springer, 1988).
8. Eracleous, M., Halpern, J. P., Gilbert, A. M., Newman, J. A. & Filippenko, A. V. Rejection of the binary broad-line region interpretation of double-peaked emission lines in three active galactic nuclei. *Astrophys. J.* **490,** 216–226 (1997).
9. Dietrich, M. *et al.* Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. XII. Ground-based monitoring of 3C 390.3. *Astrophys. J.* **115** (Suppl.), 185–202 (1998).

10. O'Brien, P. T. *et al.* Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. XIII. Ultraviolet observations of the radio-galaxy 3C 390.3. *Astrophys. J.* **509,** 163–176 (1998).
11. Sergeev, S. G., Pronik, V. I. & Sergeeva, E. S. Arp 102B: variability patterns of the Hα line profile as evidence for gas rotation in the broad-line region. *Astron. Astrophys.* **356,** 41–49 (2000).
12. Eracleous, M. & Halpern, J. P. Completion of a survey and detailed study of double-peaked emission lines in radio-loud active galactic nuclei. *Astrophys. J.* **599,** 886–908 (2003).
13. Lauer, T. & Boroson, T. HST images and KPNO spectroscopy of the binary black hole candidate SDSS J153636.22+044127.0. *Astrophys. J.* **703,** 930–938 (2009).
14. Chornock, R. *et al.* SDSS J1536+0441: an extreme "double-peaked emitter," not a binary black hole. *Astron. Telegr.* **1955** (2009).
15. Chornock, R. *et al.* The quasar SDSS J1536+0441: an unusual double-peaked emitter. Preprint at ⟨http://arXiv.org/abs/0906.0849⟩ (2009).

Competing financial interests: declared none.

doi:10.1038/nature08665

Boroson and Lauer reply

Replying to: C. M. Gaskell *Nature* **463,** doi:10.1038/nature08665 (2010)

Gaskell[1] makes the point that the profile of the Balmer lines in the candidate binary supermassive black hole[2] AGN J1536+0441 bears some similarity to those that are thought to arise from disk emission in other objects. He further argues that two predictions of the competing binary black-hole hypothesis are (1) radial velocity changes that are interpretable in terms of the binary orbit, and (2) line fluxes in the two peaks that vary independently.

The similarity to disk emitters has been bolstered by subsequent spectroscopic observations[3,4], which show a red extension to the line profiles. However, the degree of similarity is debatable; the sharpness and strength of the blue peak and the presence of a central broad component make the line profiles in J1536+0441 unusual, if not unique. When sharp peaks are seen in objects thought to be disk emitters, these peaks are transient. As the simple disk emission models do not reproduce these characteristics at the extreme levels seen in J1536+0441[5], it is unclear whether their presence excludes the disk interpretation or not.

It is certainly true that the strongest evidence for determining the nature of this object will come from spectroscopic monitoring to detect radial velocity changes and flux changes within the line profiles. Initial evidence is inconclusive[3,4], in that no changes have been detected over a little less than one year. This finding excludes some of the orbital parameter space, but not that interpretation. Several more years of unchanging fluxes and velocities would push both explanations into uncomfortable areas, as the material in the disk must be orbiting as well.

It may be that J1536+0441 is a single AGN with an emission line profile resulting from a disk configuration of material. There are credible arguments for and against that hypothesis as well as the possibility that the system is a bound system of two supermassive black holes. Whichever the case, Gaskell is correct that the presence of at most one such object out of a sample of 17,500 has interesting implications for the process by which supermassive black holes merge.

Todd A. Boroson[1] & Tod R. Lauer[1]
[1]National Optical Astronomy Observatory, Tucson, Arizona 85719, USA.
e-mail: tyb@noao.edu

1. Gaskell, C. M. Close supermassive binary black holes. *Nature* **463,** doi:10.1038/nature08665 (2010).
2. Boroson, T. A. & Lauer, T. R. A Candidate sub-parsec supermassive binary black-hole system. *Nature* **458,** 53–55 (2009).
3. Chornock, R. *et al.* The quasar SDSS J1536+0441: an unusual double-peaked emitter. Preprint at ⟨http://arXiv.org/abs/0906.0849⟩ (2009).
4. Lauer, T. R. & Boroson, T. A. HST images and KPNO spectroscopy of the binary black hole candidate SDSS J1536.22+044127.0. *Astrophys. J.* **703,** 930–938 (2009).
5. Gezari, S., Halpern, J. P. & Eracleous, M. Long-term profile variability of double-peaked emission lines in active galactic nuclei. *Astrophys. J.* **169** (Suppl.), 167–212 (2007).

doi:10.1038/nature08666