Time for the epigenome

The complexity of genetic regulation is one of the great wonders of nature, but it represents a daunting challenge to unravel. The International Human Epigenome Consortium is an appropriate response.

Oncologist Ronald DePinho of Harvard University in Boston, Massachusetts, may be unaware of the impact of his innocent question at a meeting in 2004 on cancer epigenetics, a field then just starting to burgeon. But it was like the proverbial flap of a butterfly's wing in Brazil that caused a tornado in Texas. "Where do you want the field of epigenetics to be in five years?" he casually asked.

By 2004, large-scale genome projects were already indicating that genome sequences, within and across species, were too similar to be able to explain the diversity of life. It was instead clear that epigenetics — those changes to gene expression caused by chemical modification of DNA and its associated proteins — could explain much about how these similar genetic codes are expressed uniquely in different cells, in different environmental conditions and at different times.

That tiny perturbation — the curiosity of an interested outsider — focused the thoughts of a few of the scientists there, overwhelmed by an explosion of new data, and spurred them to action. The next year they held a workshop under the auspices of the American Association for Cancer Research. The participants focused on the concept of the 'epigenome' — a genome-wide map of epigenetic modifications — and developed a blueprint for a Human Epigenome Project. This would reach much further than the epigenetics of cancer. Rather, it would amount to a global activity involving all relevant areas of biology. After all, given the universality of the fundamentals of gene regulation across all organisms and all cell types, biologists needed to be speaking the same language from the beginning.

This blueprint has been refined, and was launched last week as IHEC, the International Human Epigenome Consortium, the metaphorical tornado (see page 596). If DePinho were to ask his question now, the answer would trip instantly off the tongue: within a few years, epigeneticists will have hundreds of reliable reference epigenomes freely available to them at a mouse-click.

Although some may baulk at the launch of yet another giant '-omics' project, particularly in a complex area where there is still so much uncertainty, IHEC is in fact setting off at just the right time to instil order. It could be said, indeed, that the train has already left the station: the US National Institutes of Health is already setting up reference-epigenome centres, and the European Commission is planning to do so soon. This is not the time to reopen a clichéd debate on big science versus small science — the latter will carry on regardless, all the better for the standards IHEC will provide. It is instead the time for funding agencies, and, it is to be hoped, industry also, to generate money for an initiative that serves a vision that is both grand and thoroughly worthwhile.

> "The project's most enthusiastic proponents refer to it as the hadron collider of biology."

IHEC's most enthusiastic proponents refer to it as the hadron collider of biology — an exuberant comparison that nonetheless bears consideration. CERN's Large Hadron Collider (LHC) near Geneva, Switzerland, is a global project searching for the most fundamental subatomic particles that must exist if the standard model of the physics underpinning our Universe is correct. If biologists were to refer to an equivalent standard model of the biological codes that underpin the diversity of life, the genetic code itself would be only a small part of it. Epigenetics, on the other hand, could contribute a very significant proportion of the rest.

In terms of sheer hardware, IHEC doesn't compare to the LHC. Nor will its costs be quite as huge. But its scientific vision is just as compelling. And, given that epigenetic coding will be orders of magnitude more complex than genetic coding, its requirement for data crunching may be similar. But IHEC will deliver information about health and disease as it progresses, particularly about intractable diseases such as cancer. For all of these reasons, and just as with the LHC, the Human Epigenome Project is a challenge whose time has come. ∎

Stand and deliver

Science has done well in the proposed US budget. Researchers need to justify the funding boost.

The surprises just keep coming. Even off the back of the worst economic crunch seen in many years, US President Barack Obama has evidently concluded that science should continue to be privileged. His US$3.8-trillion budget request to Congress, unveiled on 1 February, contains generous funding for basic-science research, including a proposed 8% boost for the National Science Foundation and the largest percentage increase since 2003 for the National Institutes of Health (see page 594). Nearly every research agency emerges as a winner in Obama's budget proposals. Even NASA, heavily curtailed in its plans for human exploration of space, is slated for an increased spend on science.

The rises are remarkable given the political and fiscal stresses being experienced by the United States. Obama's administration has been trying to prove that it cares about responsible spending, and one focus of the proposed budget has been reining in the deficit, which is projected to reach up to $1.6 trillion in the next year. The administration thus dutifully declared a freeze at present funding levels for most non-defence discretionary spending programmes — those that Congress has authority to change from year to year, which excludes major mandatory items such as social security and Medicare health-insurance programmes.

Because such discretionary spending accounts for about a seventh of the national budget, this freeze comes as something of a

hollow gesture. Yet it is remarkable that even so, scientific research is mostly exempt.

Whether Congress will approve the budget is the question, given how little the president and Congress have managed to get done together, even before the Democrats lost their overwhelming majority in the Senate following last month's Massachusetts election victory by a Republican candidate. However, if the budget goes through as something resembling Obama's proposal, funding levels for basic-research agencies would stay on track for a doubling over ten years. Before researchers begin rejoicing at the prospects of more money for grants, they would be wise to remember that the economic downturn continues to resonate in many areas.

Universities continue to suffer record drops in their endowments, with many researchers at public and private universities going on involuntary unpaid leave. Competition for graduate-student and postdoc places is fierce, with many qualified students being turned away. This is particularly worrying because training the next generation of scientists and engineers is one of the primary themes of the Obama administration. In announcing his budget this week, he argued again that investment in scientific research would lead to jobs and new industries down the line. That may well turn out to be true but, in the short term, students and young researchers are still feeling the pinch.

Future investment from the private sector also remains unknown. US venture-capital spending plunged by 31% from 2008 to 2009, according to market tracker Dow Jones VentureSource, although the biotech sector emerged the least scathed, with 'only' an 11% drop. It remains to be seen how venture capitalists and larger corporations that survive the economic downturn will restructure and rethink their investments in research and development.

Nevertheless, researchers can be relieved at the proposed increases in the presidential budget. But they need to be aware that expectations will be high for delivering on this largesse, and for accountability.

One form of accountability lies in the administration's laudable effort for government transparency. To this end, research agencies can be supportive by posting high-value data sets from their grantees on Data.gov. As of this week, there remained much opportunity for improvement; agencies were trickling out fascinating but stubbornly limited data sets, such as the number of Freedom of Information Act requests received by the National Science Foundation last year. Full transparency will help to inform the American people as to whether their tax money is being best spent.

Another more contentious form of accountability lies in the extent to which the science budgets are deployed on research that is intended to address short-term problems. Research communities can anticipate pressures to move the balance of research effort in that direction. However researchers respond, they will need collectively to go to unprecedented lengths to convince the nation that the president's trust was not misplaced. ∎

Back to books

Researchers should be recognized for writing books to convey and develop science.

As the era of the electronic book dawns, perhaps hastened by Apple's much-touted iPad, researchers should prime themselves to take advantage of the spacious book format. Unlike a tweet, blog or research paper, a good book offers space to breathe, to contemplate complex ideas and to convey a mode of thinking. But most scientists don't think of writing one, and, if they do, they do it in secret.

In the hope that this might change, this week we kick off a series of weekly interviews with science book authors in *Nature*'s Books & Arts section. Peter Atkins reveals the hard work behind a successful textbook (see page 612); Carl Zimmer will highlight how passion is essential for popular science; David Brin will reveal that criticism improves fiction writing; Georgina Ferry will share research tips for biographies; and Joanna Cole will explain how to convey science to children.

Such advice is timely. The role of textbooks in handing down the tenets of disciplines is changing as online components take over from printed text. Atkins acknowledges that the extra effort of producing layers of educational material for the web today makes writing a textbook daunting. Most researchers, he admits, would not be able to devote so much time to translating their work for students. Covering broad core subjects such as general chemistry would be nigh on impossible to do in snatched moments. He is lucky that, following his publishing success, his department supported his shift to full-time writing and teaching. Many would see their careers set back if their research was displaced.

Textbook workloads could be eased in the online environment by being written in modules by conglomerates, in a 'wiki' model of education. Such accumulations of knowledge are extensive, valuable and democratic. But they move away from the classic idea of a textbook as one person's view of a field, accumulated through personal experience, years of research and face-to-face interactions with students.

Beyond textbooks, the human side of research deserves exposure as much as it ever has, through popular science accounts, biography and fiction. Although the publishing markets today apply a narrow filter, future readers can expect to enjoy access to a wider range of topics through e-books, which are easier to distribute and lack the overheads of print. As a specialist area, science stands to benefit.

Rather than limit scientific discourse to curt journal papers, researchers should embrace the book as another means of expressing not only their insights but also their visions. Through the various styles of writing, all aspects of science can be explored and laid out for posterity and learning. The expansiveness of a book allows sophisticated arguments to be put forward and widely debated; new ideas that cross disciplinary boundaries can more readily be shared and worked through.

But if this exhortation is to have any traction, the effort and skill required to write a book needs to be rewarded in the career recognition of scientists who devote time to mastering the art to good effect — a recognition that is commonplace in the social sciences and humanities. It is time to bring the book back into the science mainstream. This needn't be a mass movement: just a dedicated few, but more of them, could fulfil the reasonable hope that their books will inspire a new generation. And they should be encouraged to do so. ∎

RESEARCH HIGHLIGHTS

Flower power

Proc. R. Soc. B doi:10.1098/rspb.2009.2163 (2010)

Europe's carnations have diversified faster than any plant group studied so far — a surprising finding that counters the idea that European plants form new species more slowly than tropical plants.

Luis Valente of the Royal Botanical Garden of Madrid and his colleagues analysed specific DNA sequences from plants belonging to 104 species of European carnation (*Dianthus* spp., pictured) to trace the relationships among them. They found that carnations have been accumulating new species at the speedy rate of 2.2–7.6 per million years. By comparison, lupins (*Lupinus* spp.) of the tropical Andes have generated only 1.3–3.8 new species per million years.

One possible explanation offered by the authors is that carnations, which flower during dry summer months when pollinators are scarce, have developed diverse floral arrangements to compete for pollinators.

ENGINEERING
Sticky when wet

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0914720107 (2010)

Inspired by leaf beetles' way of sticking to leaves, researchers have created a device that uses the surface tension of many water droplets to adhere to a variety of surfaces.

Michael Vogel and Paul Steen of Cornell University in Ithaca, New York, devised a system in which water is pumped through hundreds of tiny holes in a plate by an electric pulse. The surface tension of the droplets being pushed through these holes generates the adhesive force. When the plate comes close enough to another surface, the droplets form 'liquid bridges' with the object to hold it. To release, the water is pumped back into the device, breaking the bridges.

The duo show that the device can hold its own weight plus a payload (pictured right) for several minutes.

PHYSICS
Plasma pinch

Nature Phys. doi:10.1038/nphys1510 (2010)

Charged particles trapped in Earth's magnetic field form a plasma, and tend to clump together when disturbed by the turbulent solar wind. This 'pinching' effect has been replicated in the laboratory to create a hot, dense plasma. The achievement opens up the possiblity of developing a device to fuse deuterium and helium-3 — a process that could replace radioactive tritium-based reactions in proposed fusion reactors.

Michael Mauel of Columbia University in New York and his co-workers used a copper magnet to levitate a half-tonne superconducting dipole magnet. Microwave energy injected into a deuterium-gas-filled vessel containing the levitating magnet created the hot plasma; it also induced electric fluctuations that caused the ionized particles to pile up in the plasma's core.

ORGANIC CHEMISTRY
Methylene magic

Science **327**, 566–571 (2010)

Many organic molecules are made up of multiple methylene units ($-CH_2-$) in which the carbon–hydrogen bonds are inert to most reagents. Selectively converting just one of these methylenes into something else — to build a drug molecule, for example — is one of the biggest challenges in organic chemistry.

Mark Chen and Christina White at the University of Illinois at Urbana-Champaign have now developed an iron-based catalyst that can help hydrogen peroxide to oxidize specific methylene groups in a predictable manner, mimicking the actions of natural iron-containing enzymes. This process turns methylene into an oxygen-containing chemical group, which can be modified further in subsequent reactions.

The researchers demonstrated their technique on a series of terpenoids, the largest and most diverse set of naturally occurring organic compounds.

ASTROPHYSICS
Starlight versus dark matter

Astrophys. J. **710**, 236–247 (2010)

A surplus of high-energy electrons reported by instruments such as NASA's Fermi Gamma-ray Space Telescope has been interpreted as an indirect signal of dark matter, which is thought to make up around 85% of the Universe's matter. But starlight is enough to cause this excess of electrons, say Łukasz Stawarz and his colleagues at Stanford University in California.

They developed a model of electron propagation through the galaxy that they say more rigorously accounts for the effect of starlight on the electrons. By selectively scattering electrons of specific energies, starlight suppresses most electron energies in a way that gives the impression that there is

an excess of certain high-energy electrons.

The model casts doubt on theories suggesting that reported instrument results indicate signs of dark matter.
For a longer story on this research, see go.nature.com/FKwzJy

ATMOSPHERIC PHYSICS
Bolt from the blue

Geophys. Res. Lett. doi:10.1029/2009GL041753 (2010)
Gamma rays are bursts of high-energy radiation normally associated with powerful astronomical events, but they have also been observed on Earth. Electrons accelerated in lightning bolts to nearly the speed of light are believed to be behind these 'terrestrial' gamma-ray flashes. But the location of the lightning bolts and their flashes — a key fact needed to figure out how the flashes come about — is not known.

Morris Cohen of Stanford University in California and his colleagues have now built a map showing where lightning bolts and their associated flashes occurred. The researchers used satellite data on gamma-ray flashes, and a global network of very-low-frequency antennas to detect radio waves from lightning. They pinpointed the location of 36 lightning bolts that were associated with flashes and determined which came first. The order of the two events was not always the same.

BIOCHEMISTRY
Chemical gene switch

Proc. Natl. Acad. Sci. USA doi:10.1073/pnas.0911209107 (2010)
The ability to selectively control gene expression has a wealth of applications in biomedical research. Riboswitches are a promising tool; they are a portion of a messenger RNA (which is translated to generate a protein) that can bind to a small molecule to regulate the RNA's activity.

Jason Micklefield and his colleagues at the University of Manchester, UK, have developed riboswitches that, unlike naturally occurring counterparts, respond to synthetic small molecules but not to natural ones. This potentially enables greater gene control.

The authors introduced their switch into *Escherichia coli* and used it to induce greater gene expression by adding higher concentrations of a synthetic small molecule. They also generated mutated versions of their riboswitches that could induce low, medium or high levels of protein production.

EVOLUTIONARY ANTHROPOLOGY
Baby-like bonobos

Curr. Biol. doi:10.1016/j.cub.2009.11.070 (2010)
Adult bonobos are youthful in play and share food easily. Certain aspects of their behaviour and cognition seem to be developmentally delayed forms of those same traits in the closely related chimpanzee.

To test this hypothesis, Victoria Wobber of Harvard University and her colleagues studied groups of chimpanzees and bonobos of varying ages in the Republic of the Congo and the Democratic Republic of the Congo.

The team shows that whereas chimpanzees (pictured below, right) become less willing to share food as they approach adulthood, bonobos (left) maintain relaxed, juvenile levels of food sharing. Bonobos are also slower to acquire social inhibition. The bonobos' delayed behavioural development correlates with aspects of their cranial morphology, which has also retained juvenile traits. This suggests that these species share a common mechanism responsible for changes in behavioural and brain development.

IMAGING
Virus vision

Nature Biotechnol. doi:10.1038/nbt.1604 (2010)
The spread of hepatitis C in cultured human liver cells can now be viewed non-invasively with a fluorescent microscope.

Tracking this viral infection in cells has historically required destructive sampling methods. Charles Rice at the Rockefeller University in New York City and his colleagues took advantage of a protein that is anchored to key cell organelles and is normally cleaved by the virus to allow it to hide from the immune system. They fluorescently tagged the protein, added a segment that would target it to the cell's nucleus and introduced it into liver cells. As the virus infected cells, it clipped the protein, freeing it from its anchor and resulting in its visible migration into the nucleus.

The researchers monitored infection of live cells in real time and then watched the process reverse when they added a viral inhibitor to infected cells.

A bioengineer discusses how mechanical forces in tissues may promote malignancy.

The connective-tissue protein collagen has been considered to be a structural barrier against tumour invasion in tissues. Enzymes that cleave collagen and other extracellular matrix (ECM) molecules were thus thought to promote tumour progression, but inhibitors of these enzymes have failed in clinical trials. And paradoxically, increased collagen expression is associated with a greater incidence of cancer spread.

Working with mice, Valerie Weaver of the University of California, San Francisco, and her team show that other ECM-remodelling parameters regulate malignancy (K. R. Levental *et al.* *Cell* **139,** 891–906; 2009). They studied an enzyme that initiates collagen crosslinking and is often found in tissue around tumours. They reveal that the crosslinking increases the stiffness of collagen matrices, which upregulates growth-factor signalling and breast malignancy. This suggests that tumour progression depends on a tissue-remodelling process that is regulated by biochemical and mechanical factors.

Bioengineers developing implantable materials that promote tissue regeneration can also learn a lot from this paper. Dense collagen capsules typically form around implanted biomaterials, which has prompted a search for clues to how to engineer surfaces that promote blood-vessel formation and tissue regeneration rather than scarring.

Knowing which factors promote malignancy may also help us to engineer materials and tissues that tip the balance towards enhanced tissue regeneration. This paper might thus stimulate ideas on how to interfere with the interplay between ECM-crosslinking enzymes that enhance matrix stiffness and ECM-protein-cleaving enzymes. Doing so may affect mechanosensitive cell-signalling pathways, promoting regeneration.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Obama budget backs basic science

Climate and clean energy receive a boost but NASA's Moon mission bites the dust.

In the first federal budget proposal to fully reflect his priorities, US President Barack Obama has signalled strong support for science and technology. His fiscal year 2011 budget request, announced on 1 February, calls for $66 billion to be spent on non-defence research and development, a 5.9% increase over this year's budget.

The increase is striking, given Obama's vow to freeze overall non-security discretionary spending — one part of the budget, not including mandatory programmes such as social security, over which Congress and the president have control. The proposed budget would fulfil that pledge, with all the potential political heat that implies, by cutting elsewhere.

"It's very gratifying to see science treated this way in a time of very severe budgetary constraint," says John Marburger, who served as science adviser under former president George W. Bush.

Obama has called investment in science a driver of the American economic engine. His proposal includes a modest increase for the National Institutes of Health (NIH), which apart from a one-time economic stimulus bonanza has experienced stagnant funding in recent years. And it would see strong increases for the National Science Foundation (NSF), the National Institute of Standards and Technology (NIST) and the Department of Energy's Office of Science (see Table). Bush had started to double the collective budgets of these three agencies over 10 years, and Obama's request would keep that plan on track.

Congress has the final say over what the budget numbers for 2011 will be, and it is still unclear whether lawmakers will go along with his priorities. Last year, however, Congress did award more than $21 billion in research and development funding in the economic stimulus bill.

The new request puts a heavy emphasis on climate research and clean-energy development, allocating $2.6 billion to the multi-agency US Global Change Research Program, for example — an increase of 21%. It would tackle some long-neglected problems, including restructuring the troubled National Polar-orbiting Operational Environmental Satellite System to streamline coordination between the National Oceanic and Atmospheric Administration,

Barack Obama hinted at his spending priorities in his State of the Union address.

NASA and the defence department. Perhaps most significantly, the proposal would cancel NASA's flagship Constellation programme of rockets and capsules, which was inaugurated under Bush with the goal of returning astronauts to the Moon. That move is likely to face stiff opposition in Congress, as the programme produces many jobs in key members' districts.

Here's a closer look at what the proposals mean for key agencies.

NASA

NASA's overall budget would rise by 1.5% to $19 billion. But in a radical shift, Obama's administration would give up a government-developed replacement for the space shuttle and would abandon the goal of returning humans to the Moon.

Constellation was deemed to be too far behind schedule and too expensive to justify adding to the $9 billion that has already been spent on it. "We are not on a sustainable path to get back to the Moon's surface," said NASA administrator Charles Bolden in a briefing on 1 February.

Instead, NASA would spend $6 billion over five years to stimulate the development of commercial rockets that would ferry not just cargo but also crew up to the International Space Station. These rockets could be ready by 2016, says the agency's deputy administrator Lori Garver.

Marcia Smith, a former director of the Space Studies Board at the US National Research Council, says the shift is another "topsy turvy" change in NASA's goals. "The big challenge for NASA is to convince everyone that, now, they have the right plan," says Smith.

Bolden has vowed to leave NASA science untouched by the shifting winds in the human-exploration programme, and in 2011, the agency's science budget would grow by 11% to $5 billion. Earth-science programmes would reap most of the gains, including more money to rebuild the Orbiting Carbon Observatory, a failed satellite that would have tracked global carbon dioxide levels. The astrophysics division saw a 2.6% decrease, which will hurt all the more because its biggest project, the James Webb Space Telescope due for launch in 2014, is vastly over budget.

Science may also benefit from the billions to be devoted to technology development in the human programme. Bolden says that $3 billion over five years would be spent on robotic precursor missions such as a Moon rover or an unmanned factory to mine the Moon or asteroids. Smith says these missions could be similar to the Lunar Reconnaissance Orbiter (LRO), whose costs have been shared by the human programme and the science division. However,

> **"The big challenge for NASA is to convince everyone that, now, they have the right plan."**

ROTTEN WORK
Interpretations of fossil record fail to account for decay.
go.nature.com/yickIS

she notes, most scientific destinations are chosen by peer review, whereas missions such as the LRO were dictated by the policy to return humans to the Moon.

National Institutes of Health

The NIH's budget would be boosted by $1 billion to a total of $32.1 billion, a 3.2% increase that matches the predicted rate of biomedical inflation.

"Of course I would wish for a more generous budget if that were possible," says Francis Collins, the agency's director. But, he says: "It could have been a lot worse if we didn't have an administration and a president who really values science."

For the NIH, the largest supporter of biomedical research in the world, the past few years have been a budgetary roller coaster. Following a doubling of the agency's budget between 1998 and 2003, the agency endured five years of effectively flat funding. Last year's economic stimulus bill brought a $10.4-billion windfall that significantly increased the number of successful grant applicants.

Collins said that the success rates for applications would certainly fall under the proposed budget, but he could not anticipate how far they would drop.

Despite the proposed funding increase, the agency estimates that it will give out 199 fewer competing research project grants than in fiscal year 2010, for a total of 9,052 new awards. But it would boost the payment levels on both new and existing multi-year awards by 2% to mitigate inflation. The budget would also raise stipends by 6% for the nearly 17,000 young scientists supported by the Ruth L. Kirschstein awards programme, which is the NIH's primary research-training programme for pre- and postdoctoral scientists.

Obama singled out two research areas for special attention: funding for autism studies would jump by 5% to $143 million, and cancer research funding would rise by 4.4%, to more than $6 billion.

Department of Energy

The energy department's budget would go up by more than 7% to $28.4 billion, providing significant increases for energy research and development, nuclear weapons and physical sciences.

The department's Office of Science, the nation's largest supporter of physical-sciences research, would get a 4.4% increase, to $5.1 billion. That would include a 12% hike for basic energy research, funding for which would grow to $1.8 billion.

Secretary Steven Chu's push to reorganize the department to answer pressing energy questions shaped three priorities in the budget proposal. It includes $300 million for the Advanced Research Projects Agency-Energy to promote innovative high-risk research, building on $400 million provided last year through stimulus funding; $107 million for a new 'energy innovation hub' focusing on batteries and electricity storage; and $140 million for a programme of small Energy Frontier Research Centers looking at new materials and basic energy research.

Chu acknowledges, however, that he must work to convince Congress to support these ideas: last year, he proposed setting up eight energy innovation hubs, but Congress provided funding for just three.

Spending on solar-energy research would rise by 22% to $302 million, and wind-energy research would jump by 53% to $123 million with an increased focus on offshore applications. Overall funding for fossil energy would drop by 20% after a huge increase in stimulus spending; Obama also proposes ending fossil-fuel tax subsidies worth $2.7 billion annually.

Perhaps the biggest changes come in nuclear power. In addition to cancelling funding for an underground nuclear-waste repository at Yucca Mountain in Nevada, Obama wants to restructure the nuclear-energy research programme by shifting $195 million into the development of small nuclear reactors and other advanced nuclear designs. The budget would also provide $36 billion in loan guarantees for new nuclear power plants, adding to the $18 billion that was authorized last year.

National Science Foundation

The NSF would see its budget rise from $6.9 billion to $7.4 billion. The bulk of that increase would go to the agency's core research activities, which would receive around $6 billion. This would be shared across all of the research disciplines it funds.

The budget for building major research equipment and facilities would rise by 41%, to $165 million. More than half of that boost would go to the Ocean Observatories Initiative, a network of sensors that is intended to provide unprecedented measurements of currents and other climate-influencing factors in the oceans. The agency would also spend $20 million to begin the construction of the National Ecological Observatory Network (NEON), another major facility that aims to monitor and forecast ecological changes at a continental scale over multiple decades. "NEON will collect data on the effects of climate change, changes in land use, and invasive species on natural resources and biodiversity," says NSF director Arden Bement.

As in last year's budget, Obama put a heavy emphasis on basic research into climate science, proposing $766 million for research into renewable-energy technologies and environmental- and climate-system processes. To attract and educate future scientists and engineers in clean-energy fields, a new programme known as RE-ENERGYSE would provide $19 million for up to 8,500 students pursuing clean-energy careers by 2015. Other educational initiatives include a 'cyberlearning transforming education' programme to bring advanced learning technologies such as networked computing to students at all levels. ∎

Reported and written by Janet Fang, Eric Hand, Richard Monastersky, Jeff Tollefson, Meredith Wadman and Mitchell Waldrop.
See Editorial, page 587.

OBAMA'S BUDGET REQUEST FOR 2011 (US$ million)

Agency	2009 (actual)	Stimulus	2010 (enacted)	2011 (president's request)
National Institutes of Health	30,396	10,400	31,089	32,089
National Science Foundation	6,490	3,002	6,873	7,424
Department of Energy's Office of Science	4,773	1,633	4,895	5,121
NASA	17,782	1,002	18,724	19,000
National Oceanic and Atmospheric Administration	4,454	830	4,853	5,554
National Institute of Standards and Technology (core funding)	820	580	862	922
Environmental Protection Agency	7,633	7,220	10,298	10,020
US Geological Survey	1,044	140	1,112	1,133
Department of Defense (basic and applied research)	6,830	0	7,204	6,475
Department of Homeland Security (science and technology)	933	0	1,006	1,018

IPCC flooded by criticism

Climate body slammed for errors and potential conflicts of interest.

Just over two years after winning the Nobel Peace Prize, the United Nations panel on climate change is undergoing a period of soul-searching.

The Intergovernmental Panel on Climate Change (IPCC) has always been a target for climate-change sceptics. In recent weeks, however, criticism has mounted and the panel admitted to a glaring error in its last comprehensive report, released in 2007, which says that Himalayan glaciers are likely to melt completely by 2035 (see *Nature* **463**, 276–277; 2010). On top of that, its chairman, Rajendra Pachauri, is under pressure to resign because the institute he directs, the Energy and Resources Institute in New Delhi, has ties with companies that could benefit from climate policies.

In response, many climate scientists — including a number already involved in the next IPCC assessment, due in 2014 — have been putting forward ideas about how the panel might reconsider its rules and procedures. The hope is to reduce errors in the final product and create policies on potential conflicts of interest.

The IPCC had already planned for the next assessment report, known as AR5, to follow the same basic outline as its last one, with three working groups to tackle three areas of

interest: the physical science of climate change, the impacts such change is likely to have and how these might be mitigated. There are two review editors for each chapter and the whole thing will be checked before publication, first by expert reviewers and then by governments. Governments and scientific organizations around the world are looking for people to serve as authors and reviewers; nominations are due by 12 March.

In addition to the Himalayan glacier error, the panel has been accused of saying that climate change has caused an increase in economic losses from extreme weather; the IPCC last week refuted that allegation, saying that a study in one part of the report had been taken out of context. Others have challenged the source of an estimate contained in the report that up to 40% of the Amazonian rainforest "could react drastically" to reduced rainfall.

Those involved in the process say it works well. "The IPCC is a bottom-up process in which hundreds of scientists dedicate their time on a voluntary, unpaid basis to provide an assessment of climate science," says Thomas Stocker of the University of Bern, co-chair of AR5's first working group, on the physical science of climate change. "This

The group's next assessment will focus on the socio-economic effects of climate change.

model has been extremely successful."

"In my view, the IPCC's institutional structures and processes are very strong," adds Pachauri, who says he will not step down as chairman. "What we have to do is to ensure full and complete compliance with the procedures while preparing IPCC's assessment reports."

Others say there is room for improvement. Kevin Trenberth of the National Center for Atmospheric Research in Boulder, Colorado, a lead author of the 2007 report, says that critics

Project set to map marks on genome

At one time, sequencing the whole human genome seemed almost impossible. But even as it was being completed, biologists were realizing that the genes encoded within it would reveal little about what makes each of us unique.

By 2003, it was clear that the true key to understanding our differences was more likely to be found in epigenomics — the reversible chemical modifications to DNA and its associated proteins that determine when genes can be expressed. And epigenomic studies could open up new research avenues in cancer and other complex diseases.

So it was with some fanfare that top biologists launched the
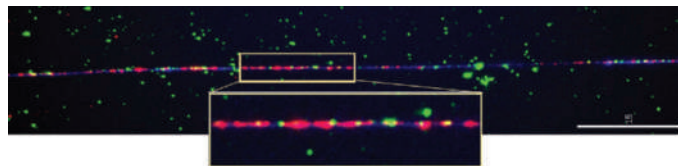
Chromatin fibres (red) can display epigenetic changes.

International Human Epigenome Consortium (IHEC) in Paris last week, a global consortium that aims, as its first phase, to map 1,000 reference epigenomes within a decade.

"Epigenomes are changeable, programmable and will feed us the bottom line on how the genome works," says Rob Martienssen of Cold Spring Harbor Laboratory in New York, a member of the consortium's steering committee. "IHEC is an exciting prospect."

IHEC is now recruiting members among funding agencies and other organizations to raise US$130 million for its first phase. Executive-committee members must commit $10 million over five years, and participants must agree to make their data freely available.

The initiative was spearheaded by the US National Institutes of Health, which launched its $190-million, five-year Roadmap Epigenomics Program in 2008, and

the European Commission, which in July will publish a call for proposals for a €30-million (US$42-million) epigenetics consortium, including a large component for reference epigenomes. Representatives from Canada, France, Germany and several Asian countries voiced interest at the launch in joining IHEC.

The consortium faces a much greater challenge than did the Human Genome Project when it was launched in 1990. "The human genome is singular and finite, but the human epigenome is almost infinite — the epigenome changes in different states and different tissues," says Philip Avner of the Pasteur Institute in Paris, a member of IHEC's steering committee.

Epigenomics is indeed a complex

WARMING SLOWDOWN
Drop in stratospheric water vapour counters greenhouse effect.
go.nature.com/7FxUgG

have grossly underestimated the rigour of the IPCC review process. However, he does say that the reports from the three working groups could be better coordinated.

The reports are done in parallel, even though the impacts working group is supposed to build on the results of the science working group. "Often a separate assessment is done in the [impacts] working group, and results can differ as the expertise is quite different," Trenberth says.

The 2007 error on Himalayan glaciers reflected the lack of coordination: it appeared in the impacts report even though the science

group had not found any peer-reviewed study that reached this conclusion.

Coordination will matter even more for the next assessment, which was set up to focus more heavily on assessing socio-economic aspects of climate change and its implications for sustainable development. As such, issues including water availability, ice-sheet shrinkage and sea level rise are likely to get more scrutiny — and will require more consistent treatment across working groups. The suggested approach, adopted in October 2009, is for each working group to assign a small group of authors the task of coordinating with their counterparts in the other working groups.

However, critics also say that, in the last round, IPCC lead authors ignored comments from reviewers and governments questioning the validity of the glacier claim. For instance, David Saltz of Ben-Gurion University of the Negev in Israel questioned why one sentence in the report said that the Himalayan glaciers would disappear completely, whereas the next said that they would shrink to 100,000 square kilometres. A later note from the writing team said only: "Missed to clarify this one".

Some errors will inevitably creep in, says Jürgen Willebrand, an oceanographer at the Leibniz Institute of Marine Sciences in Kiel, Germany, and a coordinating lead author of the 2007 report. "IPCC reports are written by humans," he says. "I have no doubt that similar

> "I have no doubt that similar errors could be found in earlier IPCC reports."

errors could be found in earlier IPCC reports, but nobody has bothered to look in detail because at the time of these reports the IPCC was less visible to society, politics and media." But he says the IPCC should have a more formal process for ensuring each flagged error is dealt with promptly.

He also calls for the IPCC to develop a policy on potential conflicts of interest.

Andrew Weaver, a climate scientist at the University of Victoria in Canada, wants more far-reaching procedural changes. Rather than carrying out "monolithic" assessments, he says, the IPCC should focus on more specific problems such as describing emissions pathways required to avoid a given temperature rise. The distinction between different working groups should also be revised, he suggests.

"If you have diverse interdisciplinary teams working on specific problems, then you can have scientists, economists and engineers all looking at a particular problem through the lens of their expertise," he says. "There is so much science out there to assess; it needs to be better focused."

Any such structural changes — or any change in Pachauri's leadership of the group — would need to be approved in the plenary session of the IPCC panel. The next such meeting will take place in October in Busan, South Korea. ■

Quirin Schiermeier

---

business. The human genome comprises strings of coiled DNA supercoiled around histone proteins into dense material called chromatin. Genes can be transcribed only when the chromatin is opened up to expose them, and this process depends heavily on chemical marks — the epigenome. Methyl groups, small non-coding RNAs and other small molecules can all serve to mark the DNA and the chromatin.

Whereas the hard-wired genome is virtually identical in all of a person's roughly 250 different types of tissue and is essentially stable during that person's lifetime, the epigenome changes during development such that each cell type has its own characteristic set of marks. These marks change with age and may also change, possibly in a heritable way, in response to environmental stress. Some diseases, such as cancer, may

be caused by harmful changes to the epigenome, or changes in the epigenome may be a consequence of disease.

The normal variation in the many epigenomes a single person will have in a lifetime, or even during the circadian day, is unknown, and technologies to detect the diverse types of epigenomic mark are still developing. Some scientists at the meeting were therefore concerned that IHEC might be premature.

"But this is exactly the point of launching IHEC now — to avoid confusion and move the field forward more quickly," says Peter Jones, director of the Norris Comprehensive Cancer Center at the University of Southern California in Los Angeles.

The number of publications on epigenomics is growing exponentially, he says, which is

why researchers need to agree now on how to standardize data. "We need a set of high-quality reference epigenomes that everyone can trust, we need to standardize technological approaches and we need to avoid unnecessary duplication," he told the meeting.

Already, the costs of some epigenome technologies are falling rapidly. The first genome-wide map of DNA methylation, or methylome, was published last November (R. Lister *et al. Nature* 462, 315–322; 2009) at a cost of $100,000, according to lead author Joseph Ecker of the Salk Institute for Biological Studies in La Jolla, California. "I expect the cost of a similar methylome will fall to $10,000 in the next six months," he says. So IHEC's target budget of $130 million could allow more than 1,000 epigenomes to be mapped.

At the meeting, project leaders decided that the vast majority of reference epigenomes should be from normal human tissue. "When we have all the references, we'll be able to compare diseased tissue against them," says Avner. The new consortium may ally itself with large genomics consortia such as the International Cancer Genome Consortium to share samples.

IHEC plans to have a final policy document by April, and nominations for the executive and scientific boards by June. The boards will choose the reference epigenomes and decide how to implement open-access policies. They will also ensure quality control and a realistic level of bioinformatics support, given that the scale of the data will dwarf that of the human genome. ■

Alison Abbott
See Editorial, page 587.

Biodiversity law could stymie research

Tighter rules on accessing and developing genetic resources may be counterproductive for conservation.

Scientists who study the world's biodiversity are facing a dilemma: proposals to regulate access to the riches of ecological hotspots may hamper the research needed to monitor these areas.

The warning comes as signatories to the Convention on Biological Diversity (CBD) begin negotiating ways to strengthen the treaty's legal framework and goals, which include conserving biodiversity and promoting the sustainable use of natural resources (see 'Key questions'). A crucial part of that effort, discussed last week at a meeting in Montreal, Canada, is the reform that is generally regarded as the least effective aspect of the CBD: deciding who can exploit valuable genetic resources, such as plants that produce potential pharmaceuticals, and who should benefit financially.

Many developing countries complain that this aspect of the CBD relies on voluntary guidelines, rather than clear international legislation. This makes it difficult to police how genetic resources are used, and to ensure that countries are paid fairly if commercial products are developed from their resources. Moreover, national regulations governing access and benefit sharing can vary enormously between countries, confusing both commercial and non-commercial researchers.

The negotiators aim to produce a legally binding framework that resolves all these problems, which will be considered for adoption under the CBD at a meeting of the convention's signatory countries in October in Nagoya, Japan.

But tougher regulation could come with a cost, warns David Schindel, an invertebrate
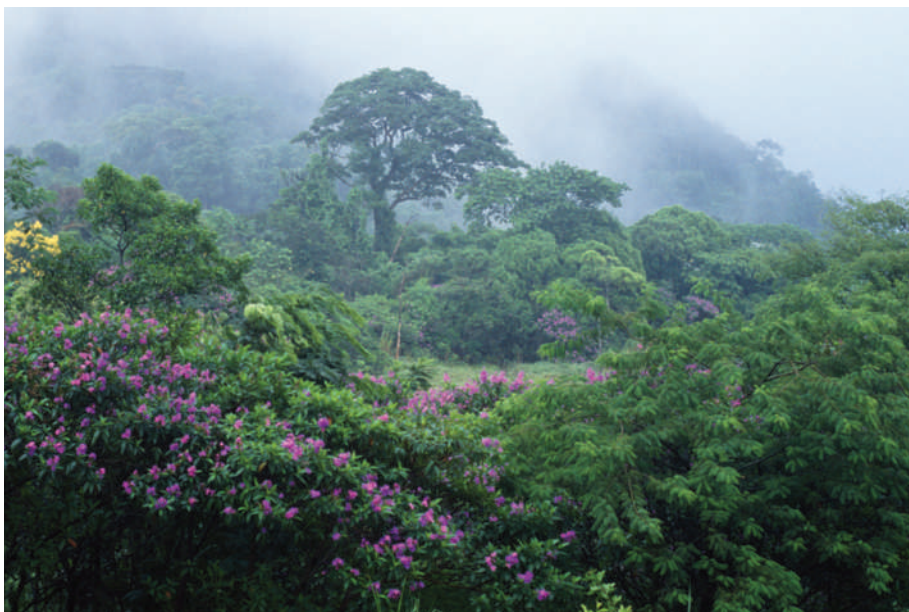
Efforts to regulate access to natural resources in Brazil have delayed research projects by years.

palaeontologist and executive secretary of the Consortium for the Barcode of Life, an international initiative to identify species using short genetic sequences, based at the Smithsonian Institution in Washington DC. "We are very concerned that it will become more restrictive," he says. In some cases, it can already take at least two years and reams of paperwork to agree the terms on which research can be conducted, specimens exported and profits shared. "You could go through a field season collecting specimens and then the government says they are going to hold on to them because you don't have the right permission," he says. "The specimens sit on a dock, rot and are lost."

David Oren, coordinator of biodiversity in the ecosystems management office of Brazil's Ministry of Science and Technology, agrees that the framework must strike a balance between protecting a nation's intellectual property and not impeding research. He says that access and benefits legislation introduced by Brazil in 2002 "essentially stopped" research on biodiversity in the country, as it restricted the exchange of specimens between institutions. The rules have since been relaxed to make a clear distinction between basic and commercial research, making it easier for academics to study biodiversity in Brazil, but many other countries, including

"If we get agreement, it will be a major breakthrough."

India, still have highly restrictive regimes, says Schindel. He hopes that the new CBD framework will clarify the situation, although some experts say that defining commercial research will not be easy — it is increasingly common for academic scientists to make patent applications based on their research, for example.

Ahmed Djoghlaf, executive secretary of the CBD, points out that an improved international framework should also benefit companies, giving them more confidence to invest the time and money required to develop products from natural resources. "What company would invest now if they don't know what a country's legislation will be like in 20 years' time?" he asks.

Other key sticking points in the negotiations include defining exactly what a genetic resource is, and whether the regime should be restricted to plants or should also include animals and pathogens.

Negotiators must reach a consensus on the framework at a meeting in March in Cali, Colombia, the deadline for finalizing the text for consideration at Nagoya. There's a lot at stake, says Djoghlaf. "If we get agreement, it will be a major breakthrough in the sustainable development and environmental movement." ■

Natasha Gilbert

Key questions

Five critical points for discussion as negotiators aim to strengthen the Convention on Biological Diversity:

● Is an ambitious long-term vision of halting all extinctions by 2050 realistic?
● Should action plans and targets to cut biodiversity loss be set at the national or international level?
● Can agricultural and other subsidies that are harmful to biodiversity be eliminated?
● Can every species facing extinction be allocated at least one protected area supporting a viable population?
● What is the best way of ensuring sustainable management of agriculture, forestry and aquaculture?

America pushes to overhaul chemical safety law

When it comes to commercial chemicals, the presumption of innocence may be coming to an end. The Toxic Substances Control Act (TSCA) allows the US Environmental Protection Agency (EPA) to test chemicals that pose a health risk — but only when it has evidence of harm. Since the law was passed in 1976, the agency has restricted just five chemicals, out of tens of thousands on the market. "It's a deeply flawed bill that needs to be rewritten," says Terry Davies, an environmental policy expert who worked on the act and is now at Resources for the Future, a nonprofit research organization in Washington DC.

Congress is likely to take up a bill this spring that would shift the burden of proof on to manufacturers; a hearing in a Senate subcommittee is scheduled for 4 February. Advocates for reform are pushing for potential legislation to include scientific advances in recent decades, such as how chemicals affect people at different ages and how multiple chemicals interact in the body.

The TSCA gave the EPA the authority to regulate chemicals and chemical mixtures not covered under laws for food, drugs, cosmetics and pesticides. But the 62,000 chemicals that were already on the market in 1976 were exempted. Of 21,000 chemicals registered since 1976, only 15% were submitted with any health-and-safety data, and the EPA has been able to require testing for only about 200 chemicals. About 95% of notices of new chemicals include confidential information, ostensibly to protect company trade secrets; this effectively prevents scientists outside the EPA from challenging the chemicals' safety.

With the TSCA essentially toothless, state officials have emerged as key watchdogs,

Many worry about toxic chemicals in toys.

banning chemicals in toys and elsewhere (see 'Patchy regulation') and supporting research into non-toxic alternatives.

But in December, Lisa Jackson, the head of the EPA, told Congress that the TSCA needed to be "updated and strengthened". She has also tapped into some unused regulatory tools available under current law. In late December, for instance, the EPA for the first time drew up a list of "chemicals of concern", including action plans that "outline the risks that each chemical may present". The list included phthalates — endocrine-disrupting plastic additives already banned in toys in Europe and several US states — along with three other classes of chemicals. And last month, Jackson announced that the agency would no longer keep confidential the names of chemicals that companies report as posing a substantial risk to health and the environment.

Further impetus for reform has come from an unexpected place: the chemical industry. Last year, the American Chemistry Council in Washington DC and other trade organizations announced their support for

"It's a deeply flawed bill that needs to be rewritten."

comprehensive legislation to overhaul TSCA. Mike Walls, vice-president for regulatory affairs at the council, says reforms are needed to "restore the public's confidence that chemicals can and will be used safely".

One problem with current law is that even as new methods arise for testing chemical safety, regulatory agencies must validate these techniques. Linda Birnbaum, head of the National Institute for Environmental Health Sciences in Research Triangle Park, North Carolina, notes that very low doses of endocrine disruptors can cause reproductive problems, cancers and other disorders, but they have eluded regulation. The EPA established an endocrine-disruption programme in 1998, but under current law it then had to validate new methods for testing their safety. Only in October 2009 was the agency able to request its first tests based on the new methods.

"It's important that the regulators have the flexibility to use all the available science," adds Birnbaum. For instance, the National Academy of Sciences has called for a shift away from animal-based studies to cell assays for testing the effects of chemicals on key pathways, then incorporating such data into computer models. In this and other areas, US regulators should benefit from comprehensive chemical testing being conducted under Europe's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation (see *Nature* 460, 1065; 2009), which came into force in June 2007.

The US chemical industry's appetite for reform has limits. Steve Risotto, a phthalate specialist at the American Chemistry Council, disagrees with the recent EPA decision to list as many as 8 phthalates as chemicals of concern. They all interfere with the male sex hormone androgen — hence the EPA's decision to list them together — but only some have been linked to cancer and developmental disorders. "That's just a dog's breakfast," Risotto says. "Maybe from an academic point of view it makes sense, but from a practical point of view, it doesn't lead you anywhere."

Davies will be watching the reform efforts closely. "I'm quite optimistic that something will happen," he says. "How good it will be or how far it will go, I'm not so sure." ∎

Brendan Borrell

E. THOMPSON/AP PHOTO

PATCHY REGULATION

Chemical	Application	Potential health links	Regulations
Bisphenol A	Plastic bottles, food cans	Prostate cancer, coronary heart disease	Banned from baby bottles in Connecticut (2009) and Minnesota (2010)
Phthalates	Plastic products	Reproductive disorders, obesity	Federal ban on several phthalates in toys (2008)
Perfluorooctanoic acid (PFOA)	Teflon and other non-stick cookware	Thyroid disease, fertility problems	EPA "chemical of concern" (2009)
DecaBDE	Fire retardant	Cancer, brain development problems	State bans, voluntary phase-out (2009)
Asbestos	Building insulation	Mesothelioma (asbestos-related cancer)	EPA total ban overturned (1991); many uses still banned
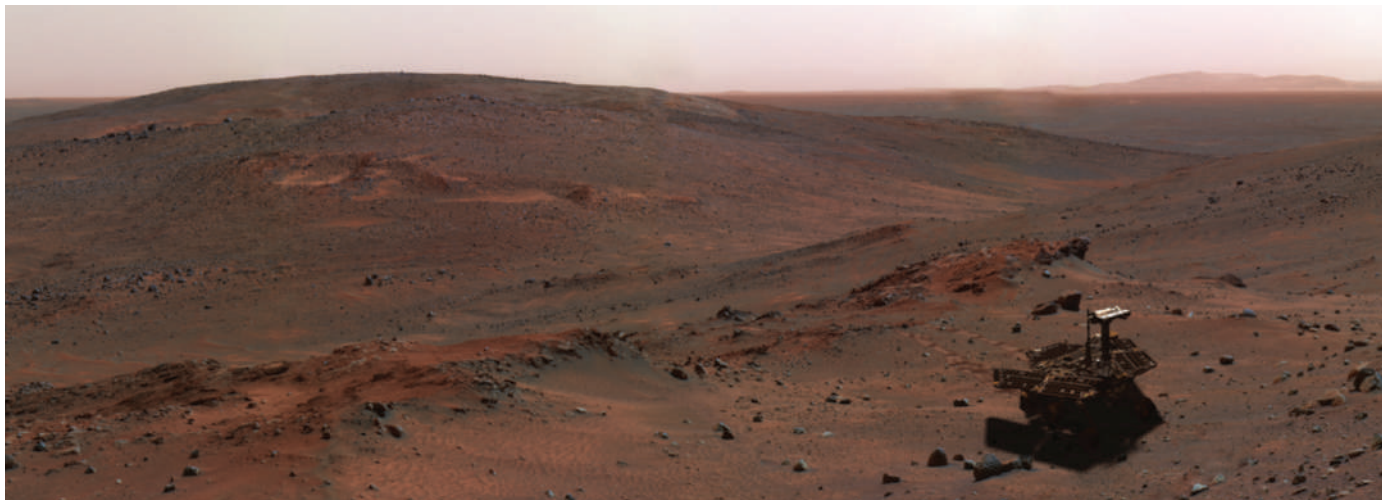
Mars rover Spirit (2003–10)

NASA commits robot explorer to her final resting place.

Spirit was born in 2003 to mission manager Mark Adler and Steven Squyres, a planetary scientist at Cornell University in Ithaca, New York. She was delivered at NASA's Jet Propulsion Laboratory in Pasadena, California, and it was there that she spent her formative months being schooled in rovering. Later, she moved to a finishing school at the Kennedy Space Center in Florida. Her graduation was epic: a 490-million-kilometre flight to Mars, where she and her twin Opportunity would pursue their destinies as roving geologists.

Her adult life began in January 2004, with an airbag-cushioned landing in the Gusev crater in January 2004 on the opposite side of the planet from her twin. The aim: to find evidence of water, and of environments that might once have been conducive to life.

With three spectrometers, an abrasion tool and panoramic and close-up cameras on board, the young rover quickly gained confidence. Her geological mettle was proved when just 32 days into her Martian voyage she picked out a rock, named Adirondack, swept it clean and drilled into it, confirming that it was the volcanic rock basalt.

Before long, she got her driving licence. She began controlling her own movements using her hazard-avoidance camera, rather than only following instructions from her large team of Earth-based mentors.

She went on to use her wire brush to uncover different-coloured layers in a rock in the Gusev crater that suggested multiple exposures to water — leading one Earthly scientist to declare the find a "miracle".

By the time the initial mission of 90 sols (Martian days) was complete, Spirit had driven 600 metres, but that was only the beginning. Well beyond her appointed days, she continued to gather valuable scientific information about Mars, sometimes with unexpected help, often against all odds. Dust was a constant nuisance, covering her life-giving solar panels. But in 2005, a dust devil happened to sweep the panels clean, giving her an energy boost.

In March 2006, Spirit's right front wheel stopped working. But she struggled on over soft ground towards McCool Hill, in the Columbia Hills region, dragging the broken wheel — and had another lucky stroke. The broken wheel churned up the soft soil, exposing dirt that Spirit analysed to show was unexpectedly rich in silicates, which need water to form.

Each winter, Spirit had to bed down on a north-facing slope to make the most of the low winter sun to charge her solar panels. A favourite spot was Home Plate, a sunny plateau that provided her with not only a winter home, but also a place to explore. While wintering here

in 2006, Spirit discovered a pair of iron-rich meteorites using her thermal-emission spectrometer. That same winter, in October, Spirit reached a milestone 1,000 sols on Mars and survived a technical hitch that support teams on Earth worried might be a Martian version of the millennium bug.

News reports back on Earth suggested that the rover's days were numbered, yet she constantly confounded any prophets of doom. But in late January 2009, Spirit's lucidity deteriorated. She had trouble moving around, couldn't identify the position of the Sun correctly, and her family on Earth had trouble understanding her. Cosmic rays were blamed.
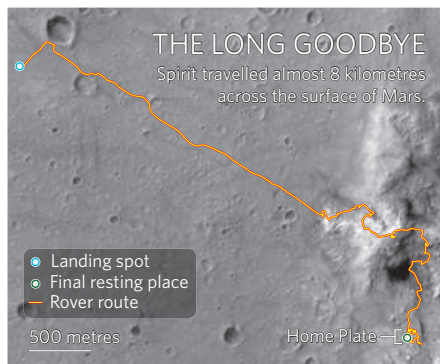
In April 2009, the ailing rover chose to reboot her computer twice. Worried controllers on Earth encouraged Spirit to press on, but more trouble lay ahead.

In a location called Troy, Spirit unwittingly crunched through the surface of a sandpit, and became entrapped. In November 2009 engineers on Earth, who had been testing a replica rover in a sand pit, tried to help her get out of her sticky situation — but to no avail. Even though the rover, by now suffering another broken wheel, did manage to climb up a few centimetres, Spirit finally gave up trying on 26 January 2010.

Her odometer read 7,730 metres. She will continue to radio back observations — of the atmosphere, of the planet's rotation — from her stationary position for as long as possible.

Spirit leaves behind her sister Opportunity — who is still active and is on her way to peer into a crater called Concepcion — and an extended family at NASA.

Katharine Sanderson

THE LONG GOODBYE
Spirit travelled almost 8 kilometres across the surface of Mars.

○ Landing spot
◎ Final resting place
— Rover route

500 metres

Home Plate

FOUR WAYS TO REINVENT THE INTERNET

The Internet is struggling to keep up with the ever-increasing demands placed on it. **Katharine Gammon** looks at ways to fix it.

The Internet is feeling the strain. Developed in the 1970s and 1980s for a community of a few thousand researchers, most of whom knew and trusted one another, the Internet has now become a crucial worldwide infrastructure that connects nearly two billion people, roughly a quarter of humanity. It offers up something like a trillion web pages, and transports roughly 10 billion gigabytes of data a month — a figure that is expected to quadruple by 2012. Moreover, those two billion users are exploiting the network in ways that its creators only dimly imagined, with applications ranging from e-commerce to cloud computing, streaming audio and video, ubiquitous mobile devices and the transport of massive scientific data sets from facilities such as the Large Hadron Collider, the world's highest-energy particle accelerator, based near Geneva, Switzerland.

To some extent, this rapidly rising flood of information has been dealt with by updating the software and expanding the size of the data

> **"Global connectivity means you have no way to prevent large-scale attacks."**
> — Felix Wu

pipes — a development that most Internet users experience through the proliferation of 'broadband' services provided through cable television connections, digital subscriber lines and wireless hot spots. Yet users continue to be plagued by data congestion, slowdowns and outages, especially in wireless networks. And, as dramatized in January when search-engine giant Google publicly protested against digital assaults coming from somewhere in China, everyone on the Internet is vulnerable to cyberattack by increasingly sophisticated hackers who are almost impossible to trace — security having been an afterthought in the Internet's original design.

The result has been a rising sense of urgency within the networking research community — a conviction that the decades-old Internet architecture is reaching the limits of its admittedly remarkable ability to adapt and needs a fundamental overhaul. Since 2006, for example, the Future Internet Design (FIND) programme run by the US National Science Foundation (NSF) has funded researchers trying to develop

wholesale redesigns of the Internet. And since October 2008, the NSF has operated the Global Environment for Network Innovations (GENI): a dedicated, national fibre-optic network that researchers can use to test their creations in a realistic setting. Similar efforts are under way in Europe, where the Future Internet Research and Experimentation (FIRE) initiative is being funded through the European Union's Seventh Framework research programme; and in

Japan, where in 2008 the National Institute of Information and Communications Technology launched JGN2plus, the latest iteration of its Japan Gigabit Network system.

Buoyed by these funding initiatives, researchers have been testing out a plethora of ideas for reinventing the Internet. It is still too early to know which will pan out. But the following four case studies give a sense of both the possibilities and the challenges.

Make the pipes adaptable

The problem with the bigger-and-bigger-data-pipe approach to dealing with the Internet's growth is that it perpetuates a certain dumbness in the system, says electrical engineer Keren Bergman of Columbia University in New York. Right now, there is no way for a user to say: "This ultrahigh-resolution video conference I'm in is really important, so I need to send the data with the least delay and highest bandwidth possible", or "I'm just doing routine e-mail and web surfing at the moment, so feel free to prioritize other data". The network treats every bit of data the same. There is also no way for the Internet to minimize redundancy. If 1,000 people are logged into a massively multiplayer role-playing game such as World of Warcraft, the network has to provide 1,000 individual data streams, even though most are close to identical.

The result is a lot of wasted capacity, says Bergman, not to mention a lot of wasted money for users who have to pay extra for high-capacity data connections that they will need only occasionally. If the Internet could just adapt intelligently to what its users are trying to do, she says, it could run much more data though the pipes than it does now, thereby giving users much more capacity at a lower cost.

This is easier said than done, however, because the dumbness is deliberate. In an effort to simplify the engineering, Bergman explains, the architecture of the Internet is carefully segregated into 'layers' that take one another for granted. This means that application programmers, for example, don't have to worry about physical data connections when they are developing new software for streaming video or online data processing; they can just assume that the bits will flow. Likewise, engineers working on the physical connections can ignore what the applications are doing. And neither has to worry about in-between layers such as TCP/IP (Transfer Control Protocol/Internet Protocol): the fundamental Internet software that governs how digital messages are broken up into 'packets', routed to their destination, then reassembled.

But this clean separation also stops the layers from communicating with one another, says Bergman, which is exactly what they need to do if the data flow is to be managed intelligently. Working out how to create such a 'cross-layer' networking architecture is therefore one of the central goals of Bergman's Lightwave Research Laboratory at Columbia. The idea is to provide feedbacks between the physical data connection and the higher-level routing and applications layers, then to use those feedbacks to help the layers adjust to one another and optimize the network's performance.

> "It's increasingly difficult for the public Internet to make progress."
> — Jonathan Turner

This kind of adaptability is not new in networking, says Bergman, but it has been difficult to implement for the fibre-optic cables that are carrying more and more of the Internet's traffic. Unlike standard silicon electronics, optical data circuits are not easily programmable. As a result, many of the dozen projects now under way in her lab aim to integrate optics with programmable electronic systems.

Bergman's lab is also a key member of the NSF-funded Center for Integrated Access Networks, a nine-university consortium headquartered at the University of Arizona in Tucson. Her group's efforts have helped to drive many of the technology development projects at the centre, which hopes to ultimately deliver data to users at rates that approach 10 gigabits a second, roughly 1,000 times faster than the average household broadband connection today. "The challenges are to deliver that information at a reasonable cost in terms of money and power," says Bergman.

Control the congestion

Meanwhile, however, some researchers are taking issue with TCP itself. Any time a data packet fails to reach its destination, explains Steven Low, professor of computer science and electrical engineering at the California

Institute of Technology (Caltech) in Pasadena, TCP assumes that the culprit is congestion at one of the router devices that switch packets from one data line to another. So it orders the source computer to slow down and let the backlog clear. And generally, says Low, TCP is right: whenever too many data try to crowd through such an intersection, the result is a digital traffic jam, followed by a sudden spike in the number of packets that get garbled, delayed or lost. Moreover, the tsunami of data now pouring through the Internet means that congestion can crop up anywhere, whether the routers are switching packets between high-capacity fibre-optic land lines carrying data across a continent, or funnelling them down a copper telephone wire to a user's house.

But more and more often, says Low, simple congestion is not the reason for lost packets, especially when it comes to smart phones, laptop computers and other mobile devices. These devices rely on wireless signals, which are subject to interference from hills, buildings and the like, and have to transfer their connection from one wireless hub to the next as the devices move around. They offer many opportunities for things to go wrong in ways that won't be helped by slowing down the source — a practice that just bogs down the network unnecessarily.

Researchers have been exploring several more-flexible ways to transmit data, says Low. One of these is FAST TCP, which he and his Caltech colleagues have been developing over the past decade, and which is now being deployed by the start-up company FastSoft in

Pasadena. FAST TCP bases its decisions on the delay faced by packets as they go through the network. If the average delay is high, congestion is probably the cause, and FAST TCP reduces speed as usual. But if the delay is not high, FAST TCP assumes that something else is the problem and sends packets along even faster, helping to keep the network's overall transmission rate high.

To test his FAST TCP algorithms, Low's lab teamed up with Caltech's high-energy physics community, which needed to transmit huge files to researchers in 30 countries on a daily basis. From 2003 to 2006, the team broke Internet world network speed records at the International Supercomputing Conference's annual Bandwidth Challenge, which is carried out on the ultrahigh-speed US research networks Internet2 and National LambdaRail. In the 2006 event, they demonstrated a sustained speed of 100 gigabits per second, and a peak

transfer speed of 131 gigabits per second — records that have not been substantially bettered by subsequent winners of the challenge.

Integrate social-networking concepts

What's great about the Internet, says computer scientist Felix Wu of the University of California, Davis, is that anyone with an address on the network can contact anyone else who has one. But that's also what's terrible about it. "Global connectivity means you have no way to prevent large-scale attacks," he says, citing as an example recent digital assaults that have temporarily shut down popular sites such as Twitter. "At the same time you are getting convenience, you are actually giving people the power to do damage."

In 2008, for example, security software maker Symantec in Mountain View, California, detected 1.6 million new threats from computer viruses and other malicious software — more than double the 600,000 or so threats detected the previous year — and experts say that these attacks will only get more common and more sophisticated in the future.

What particularly drew Wu's attention a few years ago was the problem of unsolicited junk e-mail, or 'spam', which accounts for an estimated 90–95% of all e-mails sent. What makes spam trivial to broadcast and hard to filter out, Wu reasoned, is the Internet's anonymity: the network has no social context for either the message or the sender. Compare that with ordinary life, where

people generally know the individuals they are communicating with, or have some sort of connection through a friend. If the network could somehow be made aware of such social links, Wu thought, it might provide a new and powerful defence against spam and other cyberattacks.

With funding from the NSF, Wu has created a test bed for such ideas, called Davis Social Links. The test bed has a messaging system that routes packets between users on the basis of the lists of friends that each person creates in social networking sites such as Facebook. This gives test-bed users the option of accepting only the messages that reach them through the paths or groups they trust, making it more difficult for them to be reached by spammers or attackers who lack the proper trusted paths.

These social relationships in the system don't have to be restricted to people, Wu notes. Websites are fair game too. Users of Davis Social Links can build social relationships with YouTube, for example. A search engine based on this social-network idea might pick up two sites that claim to be YouTube, one that is real and one that is cloned to look like the video site. The system would try one and if it didn't have the expected connections to other trusted contacts, the path would be designated as untrustworthy and the site dropped. "In today's routing you only give the IP address to the service provider, they do the rest," says Wu. "In social routing I don't have a unique identity. I have a social identity that supports different routing."

Davis Social Links is part of the GENI test bed and will soon start testing with up to 10 million network nodes. But even if this approach turns out not to be viable, says Wu, more types of social research need to be integrated into the future Internet. "We need to mimic real human communication," he says.

Break from reality

Computer scientist Jonathan Turner of Washington University in Saint Louis, Missouri, says that the basic packet-delivery service hasn't changed in more than 20 years not because no one has a better idea, but because new ideas can't get a foothold. "It's increasingly difficult for the public Internet to make progress," he says. The network's infrastructure is fragmented among many thousands of network providers who are committed to the Internet as it is, and who have little motivation to cooperate on making fundamental improvements.

This spectre of a rapidly ossifying Internet has made Turner a champion of data channels known as virtual networks. In such a network the bits of data flow through real, physical links. But software makes it seem as though they are flowing along totally different, fictitious pathways, guided by whatever rules the users desire.

In present-day commercial virtual networks, those rules are just the standard network protocols, says Turner. But it is possible to create virtual networks that work according to totally new Internet protocols, he says, making them ideal laboratories for researchers to experiment with alternatives to the current standards. His group, for instance, is working on virtual networks that enable classes of applications that are not well-served by the current Internet. "This includes person-to-person voice communication, person-to-person video, fast-paced multi-player games and high-quality virtual world applications," he says. "In general, any application where the quality of the user experience is dependent on non-stop data delivery and there is low tolerance for delay."

Moreover, Turner is just one of many researchers pursuing this approach. An academic–industry–government consortium known as PlanetLab has been providing experimental virtual networks on a worldwide collection of computers since 2002. The GENI test bed is essentially a collection of virtual networks, all of which run atop Internet2 and National LambdaRail. This allows the same physical infrastructure to handle multiple

experiments simultaneously — including many of the experiments mentioned in this article.

Looking farther down the road, says Turner, as the best of these non-standard, experimental protocols mature to the point of being ready for general use, virtual networks could become the mechanism for deploying them. They could simply be built on top of the Internet's existing, physical infrastructure, and users could start using the new functionality one by one. Different protocols could compete in the open marketplace, says Turner — and the era of the ossified Internet would give way to the era of the continuously reinvented Internet. ∎

Katharine Gammon is a freelance writer in Los Angeles.

Lyndon Emsley spent seven years building support to purchase the 1-gigahertz spectrometer for his institute in Lyon, France.

V. MONCORGE

Breaking the billion-hertz barrier

Researchers in France have switched on the world's most powerful nuclear magnetic resonance instrument. **Ananyo Bhattacharya** asks whether it will attract new life to NMR spectroscopy.

The 12-tonne, 4.5-metre-tall machine does little to betray the fact that it is working. No flashing lights break its smooth, white cylindrical surface. It makes no noise, not even a hum. But an incautious step over the yellow-and-black striped tape on the floor 12 metres from the world's most powerful nuclear magnetic resonance (NMR) spectrometer will erase the data on one's credit cards. Individuals with pacemakers are warned to go no closer.

It is the magnet that makes this spectrometer (pictured) at the European Centre for High Field NMR (CRMN) in Lyon, France, unique and hugely powerful. In a magnetic field, atomic nuclei with an odd number of protons or neutrons will resonate when pulsed with the right radio-wave frequency. The frequency at which they resonate depends on the chemical environment of the atom — the type of atoms that surround it, and their proximity. By gauging the resonance frequencies for a compound, researchers can map the relative positions of atoms and the structure and identity of the molecule. It is a technique that researchers have used to great effect since the first commercial 30-megahertz (MHz) machines became available in the 1950s.

The more powerful the magnet, the more spectroscopists can 'see' — picking up tiny traces of chemicals in complex mixtures of body fluids or environmental samples, for example. The magnet at the core of the Lyon spectrometer can generate a massive 23.5-tesla field. But they give it the designation 1 gigahertz (GHz) — the resonance frequency of a single hydrogen-1 nucleus in a field of this strength. As the most powerful of its kind, the machine represents a milestone.

Lyndon Emsley, the CRMN's scientific director who worked for seven years with the federal and regional governments and the CNRS, France's basic-research agency, to buy the instrument for €11.7-million (US$16.3-million), says that the machine will attract Europe's best chemists and structural biologists. A huge swathe of science is set to benefit from the added field strength, he says, which will reveal structures of bigger, more difficult to analyse proteins, and enable a clearer look into complex mixtures of chemicals and metabolites. But the machine has some work to do to prove that this milestone is more than symbolic. An increase of 50 MHz over other high-field machines hardly matches the vast improvements in power made during NMR

spectroscopy's heyday in the 1970s and 1980s, when superconducting magnet coils vastly enhanced field strength. Even Emsley's father, a chemist at the University of Southamphton, UK, who had long used NMR spectroscopy for his studies of liquid crystals, calls the step to 1 GHz "incremental". Nevertheless, Emsley the younger predicts that the behemoth will prove its mettle.

One reason for his optimism is the way in which biologists, particularly structural biologists, have produced demand for increasingly powerful machines. Emsley suspects that as much as half the time on the new machine could go to cracking the atomic structures of proteins.

Protein power

NMR spectroscopy provides advantages over its main rival, X-ray crystallography. Crystallography can resolve protein structures quickly — sometimes in hours or days, compared with months for NMR spectroscopy — but only after crystals of the protein have been grown, a trial-and-error process that can add months or years to a project. In addition, NMR spectroscopy can yield useful information about proteins without a well-defined structure, including those that are extremely difficult to

crystallize because they assume a shape only when they bind to another protein.

Charles Sanders, a biochemist at Vanderbilt University School of Medicine in Nashville, Tennessee, and his collaborators recently used an 800-MHz machine to determine the structure of diacylglycerol kinase, one of the largest membrane proteins solved by NMR spectroscopy to date (W. D. Van Horn *et al. Science* **324,** 1726–1729; 2009). But he says that the 1-GHz machine will appeal to those working on large proteins and membrane-bound proteins, which are extremely difficult to crystallize. For the latter, NMR-spectroscopy researchers can use lipid constructs called bicelles, which mimic cell membranes allowing bound proteins to assume a more natural shape. Some are even beginning to use nanodiscs, which corral lipid and protein together, maintaining them in their natural conformation (see image, right).

A technique developed in the late 1990s called transverse relaxation optimized spectroscopy (TROSY), has allowed NMR spectroscopy to deduce the structures of ever larger molecules by labelling the protein with different isotopes (K. Pervushin *et al. Proc. Natl Acad. Sci. USA* **94,** 12366–12371; 1997). The physical properties that TROSY exploits are predicted to give the sharpest possible peaks at fields of up to about 1.2 GHz, so the new machine could allow spectroscopists to crack proteins that are currently outside their grasp. Sanders, who until recently only had access to an 800-MHz machine, says he has travelled to use a 900-MHz spectrometer. "The same will be true for 1 GHz. People will travel for that."

But Jeremy Nicholson of Imperial College London says that grant-giving agencies may be losing interest in protein NMR spectroscopy. "It has been very successful in discovering the structures of a lot of very interesting, important proteins," he says. "But very few of those discoveries have translated into anything that has made a big mark in medicine."

Nicholson does foresee medical applications, however, particularly in metabolic profiling, which he helped to pioneer. Researchers generate complex chemical 'fingerprints' of body fluids such as urine, plasma or tissue — or even whole organisms such as nematode worms — which are analysed to tease out the effects of drugs, disease or environmental changes on the metabolism. Higher field strength helps to differentiate the thousands of different metabolites.

Benjamin Blaise, a PhD student and trainee medic at the University of Lyon, is carrying out metabolic profiling studies using samples from children with urinary problems. Blaise was one of the first people to use the new spectrometer. "It's amazing," he says. Formerly, he would run each of his samples about 100 times on the 900-MHz machine before getting a strong signal. Now, "you just need two experiments and you have a beautiful signal", he says.

Watching reactions

The metabolic profiling work also signals a return to NMR spectroscopy's roots in chemistry. Analytical chemists will be able to do similar kinds of chemical profiles for things such as water and soil samples to measure industrial effects. And high-field NMR spectroscopy could be used to analyse chemical reactions as they take place on the surface of large particles, such as chemical catalysts, which are often attached to grains of silica.

Here, the concentration of the active sample is often only around 1–2% of the total. The 1-GHz spectrometer "opens up new horizons", says chemist Christophe Copéret of the Lyon School of Chemistry Physics and Electronics. "We can study reactions that were not amenable to NMR at lower field."

There are still questions about whether the increase in field strength is worth the cost. Most NMR spectroscopy work doesn't require such large magnetic fields. Many protein structures can easily be deciphered at field strengths of less than 800 MHz, and 400 MHz is usually enough to check the purity and structure of compounds in chemistry during organic synthesis.

The CRMN already had a 900-MHz spectrometer before the 1-GHz machine arrived and it cost half as much. Given that for many applications the detail visible in NMR spectra improves in step with field strength, the extra €6 million seems steep for an improvement of around 10%.

Emsley sees the machine as a resource for the entire research community. With the European Union contributing towards the running costs of the facility, up to 200 days a year will be allocated to users from other labs — many from outside France. Even relatively well-equipped labs would be lucky to have a 600-MHz machine. An increase of 400 MHz is "an astonishing difference", Emsley says.

He points out that as magnetic field strength has increased in the past, new techniques have cropped up to exploit them. For example, TROSY came along just as NMR spectrometers with fields higher than 700 MHz were becoming more widely available. Sanders agrees. "In the past, every time there has been a big increase in the size of the magnet, NMR has leapt forward," he says. "Based on history, one might guess that there will be some breakthrough that results from this."

Kurt Wüthrich, who won the 2002 Nobel Prize in Chemistry for the development of NMR spectroscopy to determine protein structures, cautions that the step from 950 MHz to 1 GHz is small compared with the big hikes in field strength that have taken place since the early 1960s. For instance, fields had increased from around 60 MHz to 600 MHz by the mid-1980s. "There's now talk of 1.2- and 1.3-GHz machines," says Wüthrich, who is based at the Swiss Federal Institute of Technology in Zurich and the Scripps Research Institute, in La Jolla, California. "But in the past, relative increases were bigger."

The inroads that NMR spectroscopy has made into previously intractable problems hasn't just been driven by the increases in field strength. Other improvements in technology have contributed — such as the probe heads used to pulse samples with radio waves and measure resonances, and the computers that crunch the data. There have been big advances in sample preparation too, such as stereo-array isotope labelling (SAIL). In SAIL, proteins are made from specially labelled amino acids and these proteins produce much better spectra in NMR experiments.

Jim Emsley, Lyndon's father, says that the 1-GHz machine will provide a worthy challenge for his son. The technical advances of recent years and the funding secured by the Lyon centre are all signs that NMR spectroscopy remains a dynamic field. "As long as the science keeps saying 'Yes, it is worthwhile', that impetus will continue," he says. "So the onus now, obviously, is on Lyndon's laboratory." ■

Ananyo Bhattacharya is online news editor for *Nature.*
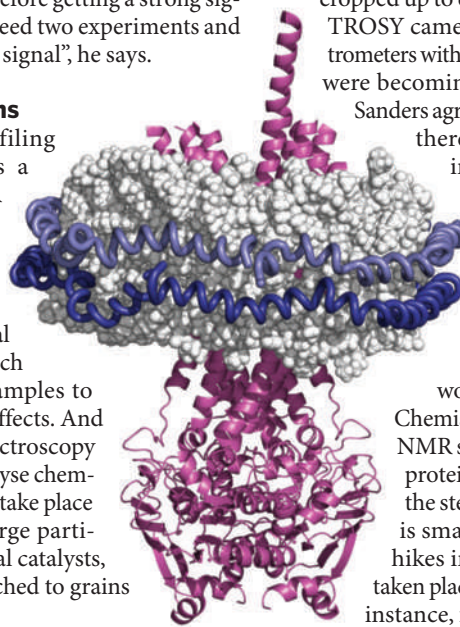
Nanodiscs made of protein (blue) and lipids (white) hold membrane proteins (pink) in their natural shape.

T.K. RITCHIE ET AL. METHODS ENZYMOL. **464,** 211–231 (2009)

> "You just need two experiments and you have a beautiful signal."
> — Benjamin Blaise

Better all the time

Innovation policies are more likely to be successful if they leverage existing capabilities, argues **Daniel Sarewitz**.

Scientists and policy-makers often insist that 'more research' is needed to solve a wide range of social challenges. But what is the most effective way to deploy our formidable knowledge-creating enterprise in the shadow of looming threats to human welfare, such as AIDS or climate change?

Last September, scientists and the media trumpeted a 'breakthrough' in AIDS vaccines after a 6-year clinical trial in Thailand seemed to show success in preventing HIV infection. Yet the results were almost immediately contested (see *Nature* **461,** 1187; 2009), and the trial's importance was soon being spun as a scientific advance rather than a practical step toward a successful vaccine. This latest frustration in the 25-year quest for an AIDS vaccine highlights a difficult but essential lesson for science policy: rapid and short-term progress in addressing urgent problems is much more likely to come from improvements on existing capabilities than from radical breakthroughs that emerge from basic science.

Major efforts to apply science and technology to AIDS have followed two main paths — the search for drugs to treat those already infected, and the pursuit of a vaccine to protect against infection in the first place. The differences in outcome of these two paths could not be more stark. People with AIDS who have access to combination drug therapies are experiencing radically improved health outcomes. Vaccines have failed to demonstrate any benefit.

The reason the drug path has succeeded is startlingly simple: when the AIDS crisis began in the early 1980s, chemicals already existed that could slow the disease, and the state of the science was sufficient to identify them.

Better to be lucky

The compound azidothymidine (AZT) was first described in 1964; it was supposed to fight cancer, but didn't. A decade later, it was found to inhibit the growth of retroviruses in mice. By the early 1980s, the drug firm Burroughs Wellcome (now GlaxoSmithKline) was exploring the drug's potential as an antibiotic. When scientists determined in 1984 that AIDS was caused by the HIV retrovirus, the US National Institutes of Health (NIH) in Bethesda, Maryland, began working with drug companies to identify existing chemicals that could act against the infection. AZT was tested against live HIV by the NIH National Cancer Institute,

WORLD VIEW

where researchers found that it suppressed the virus. Clinical trials quickly followed, and the drug was approved by the US Food and Drug Administration in March 1987.

Innovation is a social process, and the discovery of a drug that could act against a devastating and mysterious disease galvanized the diverse social actors necessary for rapid progress. Government science administrators recognized the need for a coordinated effort; scientists saw promising opportunities to confront an emerging epidemic; drug companies smelled big profits; and people with AIDS, willing to try almost anything that might help, pushed the regulatory system to bring new drugs to patients faster than ever before.

In the next 25 years, 25 drugs were approved for use in treating the disease (E. De Clercq *Rev. Med. Virol.* **19,** 287–299; 2009). Science didn't need to be driven by theory alone; rather, it could focus on the empirical problem of making things that worked better than AZT. By the mid-1990s, there had been a sudden and radical decrease in deaths from AIDS, at least in the United States and Europe. This inflection point was a true breakthrough, and it emerged not from new basic understanding but from the empirically discovered synergies achieved by combining different types of anti-HIV drugs into highly active antiretroviral therapies. In 1994, the 2-year survival rate for people diagnosed with AIDS was less than 50% in the United States; 3 years later it was nearly 75%. It continues to improve to this day.

The vaccine story, compellingly told by journalist Jon Cohen in his 2001 book *Shots in the Dark* (W. W. Norton & Company), couldn't be more different. In the absence of an even modestly effective way to stimulate immunity against HIV, vaccine research has lacked the focus and baseline for demonstrating the progress that is necessary to drive

> **"Chances of success in basic research should be understood as both unpredictable and long-term."**

rapid innovation. The same groups whose disparate interests converged around the quest for antiretrovirals contribute, instead, to disorder. Research has been carried out mostly by academics whose main motive is to advance knowledge and careers. Private-sector involvement has generally been limited to biotech firms looking for a game-changing new product; most of the big drug companies have stayed away, seeing little opportunity for profit and lots of risk of liability. In the face of technical uncertainties, governments have faced excruciating dilemmas about how best to allocate scarce resources. In the absence of evidence of efficacy, patient groups have been ambivalent even about participating in trials. These are not reasons to stop trying to find a good vaccine, but they are reasons not to expect significant progress any time soon.

Downstream policy

If a problem demands short-term technical progress, then innovation policies need to focus on accelerating the improvement of approaches that already show clear evidence of efficacy. Research is an important aspect of this process, but mostly insofar as it supports the knowledge needs of private-sector innovators seeking to develop new and better products. Much more important are the downstream aspects of innovation policy, especially those that encourage private-sector participation, and the cultivation of markets.

The lesson applies in domains as disparate as human health and energy supply. For example, a key role for the NIH in the development of antiretrovirals was running drug trials to demonstrate safety and efficacy. An analogous role for governments in the energy domain is running demonstration projects for unproven technologies such as carbon capture and storage, or modular nuclear reactors, to create confidence in technical performance and cost estimates.

Basic research — on AIDS vaccines, on new energy technologies — should be aggressively pursued as part of a portfolio of public investment in innovation, but the chances of success should be understood as both unpredictable and long-term. The AIDS story provides a clear message for policy-makers seeking to address urgent problems with new technologies: look for what already works, and make it work better. ∎

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

Our decadal research predictions ('2020 visions' *Nature* **463,** 26–32; 2010) provoked ideas — and ire.

Political instability may be a contributor in the coming decade

The next decade is likely to be a period of growing instability in the United States and western Europe, which could undermine the sort of scientific progress you describe in the Opinion collection of '2020 visions'.

Quantitative historical analysis reveals that complex human societies are affected by recurrent — and predictable — waves of political instability (P. Turchin and S. A. Nefedov *Secular Cycles* Princeton Univ. Press; 2009). In the United States, we have stagnating or declining real wages, a growing gap between rich and poor, overproduction of young graduates with advanced degrees, and exploding public debt. These seemingly disparate social indicators are actually related to each other dynamically. They all experienced turning points during the 1970s. Historically, such developments have served as leading indicators of looming political instability.

Very long 'secular cycles' interact with shorter-term processes. In the United States, 50-year instability spikes occurred around 1870, 1920 and 1970, so another could be due around 2020. We are also entering a dip in the so-called Kondratiev wave, which traces 40-60-year economic-growth cycles. This could mean that future recessions will be severe. In addition, the next decade will see a rapid growth in the number of people in their twenties, like the youth bulge that accompanied the turbulence of the 1960s and 1970s. All these cycles look set to peak in the years around 2020.

Records show that societies can avert disaster. We need to find ways to ameliorate the negative effects of globalization on people's well-being. Economic inequality, accompanied by burgeoning public debt, can be addressed by making tax rates more progressive. And we should not expand our system of higher education beyond the ability of the economy to absorb university graduates. An excess of young people with advanced degrees has been one of the chief causes of instability in the past.

Peter Turchin Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269, USA
e-mail: peter.turchin@uconn.edu

Could a boom in technologies trap Feynman's simulator?

I believe that quantum simulation is slated for a breakthrough by 2020. In a groundbreaking lecture (*Intl J. Theor. Phys.* **21,** 467–488; 1982), Richard Feynman pointed out that large quantum systems are impossible to simulate with a classical computer. He proposed that they might instead be simulated with a quantum-mechanical machine — in short, a quantum simulator.

Some promising technologies have emerged, but verifying a quantum simulation is not straightforward. Do the results demonstrate properties of the simulated model, or are they due to unrelated features of the simulator? At first, this quandary seems insoluble, as the same complexity that drives one to attack the problem on a quantum simulator will thwart any attempt, whether analytical or on a regular computer, to prove the quantum-simulation result is correct.

But if the same physics models are simulated on different quantum simulators based on different technologies, it is quite likely that common features of all the results will be due to the quantum-physical model and not to the systematics of the simulators. Scientists and funders should therefore support development of as many technologies as possible.

Dietrich Leibfried National Institute of Standards and Technology, Boulder, Colorado 80305, USA
e-mail: dil@boulder.nist.gov

Expanded view of universities would be more realistic

Tertiary education is poised for greater changes during the next decade than John Hennessey's vision implies (*Nature* **463,** 28–29; 2010). The rising demand for university education in the developing world could affect universities in developed countries.

Western universities, for example, could help fill huge gaps in faculties in China, India and countries in southeast Asia, Africa and South America. Greater international cooperation and more study programmes abroad are likely to feature.

Distance learning could become more popular, particularly in remote areas. Open-course material, such as that pioneered by the Massachusetts Institute of Technology, might warrant further development. And modified programmes for professional degrees, such as a one-year Master of Business Administration, could become more widespread.

Siddharth Dasgupta NSF Center for Chemical Innovation, Caltech, Pasadena, California 91125, USA
e-mail: sdg@caltech.edu

Women: diversity among leaders is there if you look

A disappointing myopia seems to have afflicted your '2020 visions', with just one female among the 20 contributors. This sends the wrong message at a time when women scientists are still striving for better representation.

Diversity among thought leaders is there if you look for it. You no longer have to look far among academics. Today, for example, women of vision are heads of the Massachusetts Institute of Technology; Harvard, Princeton and Brown universities; and the universities of Michigan, Pennsylvania and Cambridge.

Consider the Nobel prize. In 2009, it was awarded to five women (three of them scientists) and eight men, the narrowest gender gap since its inception.

The scientific community in 2020 should reflect the talent pipeline of 2010. Women are now well placed, if they stay the course, to enhance diversity in science. But they will need encouragement, support and opportunities if the barriers that have traditionally stymied diversity are to drop away.

Nancy C. Andrews, Sally Kornbluth, Doug Stokke Duke University School of Medicine, Box 2927, Medical Center, Durham, North Carolina 27710, USA
e-mail: nancy.andrews@duke.edu

Women: why just one to represent half the workforce?

In your prognostications about the future of science, you might have featured only women as authors, given that the ancient prophesying Sibyls were always female. However, there was just one woman among the twenty writers. We trust that ten years from now we shall not have to remind *Nature* that nearly half of working scientists (and *Nature* readers) are women.

Joan M. Herbers Association for Women in Science, 1442 Duke Street, Alexandria, Virginia 22314, USA
e-mail: herbers.4@osu.edu

Readers are welcome to comment on the next 10 years for science, at go.nature.com/htW8uM

BOOKS & ARTS

The woman behind HeLa

Steve Silberman enjoys a moving account that probes racial and ethical issues in medicine through the story of the young mother whose death from cancer led to the first immortal cell line.

The Immortal Life of Henrietta Lacks
by Rebecca Skloot
Crown: 2010. 320 pp. $26

One day in 1951, a young woman named Henrietta Lacks noticed blood on her underwear. She had been feeling pains for months that were not relieved by the birth of her fifth child, a boy. Finding what she described as a "knot" on her womb, Lacks was told by her physician to see a specialist.

Because Lacks was black, the nearest clinic that would admit her was 30 kilometres away. There, in the segregated ward of Johns Hopkins University hospital in Baltimore, Maryland, a gynaecologist discovered a lesion on Lacks's cervix that was "as big as a 25-cent piece … raised, smooth, glistening, and very purple." The chairman of the hospital gynaecology department, Richard TeLinde, thought that such lesions were a harbinger of invasive cervical cancer. He was right. Lacks's knot was an adenocarcinoma of formidable metastatic power. Eight months later she was dead.

In a strange twist of fate, Lacks's cancer has lived on to this day — as HeLa, the first mass-produced human cell line. In an impressive non-fiction debut, science writer Rebecca Skloot weaves the travails of the Lacks family with the endeavours of the researchers who created HeLa. A real-life detective story, *The Immortal Life of Henrietta Lacks* probes deeply into racial and ethical issues in medicine.

'Immortal' cells such as HeLa can proliferate indefinitely because they do not have the mechanisms that limit the number of times a cell can divide before disintegrating. Key medical advances — including the development of vaccines for polio and the human papillomavirus — were made by experimenting on HeLa cells. In recent years, they have been used to develop chemotherapy drugs, to study monoclonal antibodies and to examine the effects of exposing human tissue to nanomaterials. The plump, sticky cells have been frozen, cloned,

bombarded with γ-rays, infused with mouse DNA and launched into space. Yet many HeLa researchers know little about the woman behind them.

The one previous book-length account of HeLa history, Michael Gold's *A Conspiracy of Cells* (SUNY Press, 1986), largely ignored Lacks to focus on Walter Nelson-Rees, the cytogeneticist who published lists of cell lines that had been contaminated by aggressive HeLa strains. By bringing the Lacks family out of HeLa's shadow, Skloot reveals the societal forces and human passions at work in a scientific advance that has saved millions of lives.

The architects of Lacks's virtual immortality were TeLinde along with George and Margaret Gey, pioneering cell culturists who had tried for years to invent a method of sustaining human cells *in vitro*. They succeeded when they immersed a portion of Lacks's tissue in their home-made nutrient broth. Within days, "Henrietta's cells weren't merely surviving — they were growing with mythological intensity," writes Skloot, "doubling their numbers every twenty-four hours, stacking hundreds on top of hundreds, accumulating by the millions."

The Geys became HeLa evangelists, mailing out free vials of cells to labs worldwide. To furnish Jonas Salk with a way of testing his polio vaccine, other researchers scaled up the Geys's methods. The biotechnology firms now known as Invitrogen and BioReliance each got their start by marketing HeLa cells in industrial quantities. Meanwhile, Lacks's five children were living in Baltimore, too poor to afford health insurance and unaware that their mother's cells had taken on a life of their own. The notion of requiring informed consent before taking a tissue sample — particularly from a black patient on a public ward — was foreign to the medical establishment of the time.

The emotional impact of Skloot's tale is intensified by its skilfully orchestrated counterpoint between two worlds: the
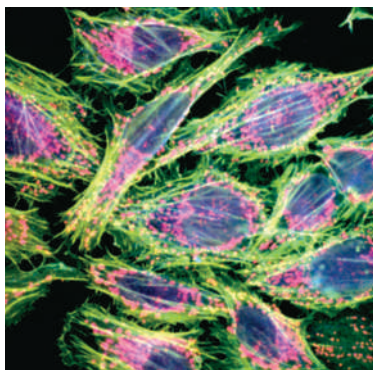

Henrietta Lacks: Twenty years after her death, her family learned that her cells lived on in a lab.

'Immortal' HeLa cells just keep on dividing.

heady pioneer days of cell research, and the hard existence of the Lacks family, beset by poverty, disease and a succession of people — including a professional con man — who tried to turn Henrietta's legacy to their own advantage.

The author forged a particularly tight bond with Lacks's feisty daughter Deborah, who learned from journalists in the 1970s that part of her mother was still alive. Half a century after Henrietta first walked into the segregated ward at Johns Hopkins, the two worlds of Skloot's book came together in a lab at the university when Deborah saw her mother's cells under a microscope for the first time. "They're beautiful," she whispered. ∎

Steve Silberman writes on health-related topics for magazines including *Wired* and *The New Yorker* from San Francisco, California.
e-mail: steve@stevesilberman.com

VISUALS UNLIMITED/CORBIS

SPL

In Retrospect: *Funes the Memorious*

When **Rodrigo Quian Quiroga** visited Jorge Luis Borges's private library, he found annotated books that bear witness to the writer's fascination for memory and neuroscience.

Funes the Memorious
by Jorge Luis Borges
First published in book form in 1944.

In 1944, the great Argentinean writer Jorge Luis Borges (1899–1986) published *Funes the Memorious*. It is the fictional story of Ireneo Funes, who, after falling off his horse and receiving a bad head injury, acquired the amazing talent — or curse — of remembering absolutely everything. "He knew the forms of the clouds in the southern sky on the morning of April 30, 1882, and he could compare them in his memory with the veins in the marbled binding of a book he had seen only once," Borges relates.

Borges is known for his fascination with mathematical and philosophical concepts, from infinity to history paths. Through Funes's story, he explores the vast labyrinths of memory and the consequences of an infinite recollection. His choice of character reveals the author's long-standing interests in psychology, memory and neuroscience.

A quest for knowledge about Borges led me to visit his widow, María Kodama, last year in Buenos Aires. Borges himself, she explained, had an exceptional memory. He could quote passages in Spanish, English, German and other languages. Borges purposely enriched his memory from an early age, knowing that he had a congenital disease that would eventually leave him blind and unable to read. Ironically, he lost his sight in 1955, the year he was appointed director of the Argentinean national library, but he retained his interest by asking others to read to him. Kodama recalled that, in one of her first meetings with the writer, he asked her to recite a specific passage in a book. To her surprise, he quickly guided her to the exact page, even though he had been blind for many years.

Characteristically modest, Borges considered himself more of an extraordinary reader than an accomplished writer. Kodama kindly showed me his private library, hosted by the Jorge Luis Borges International Foundation in Buenos Aires, which she directs. From the many annotations that adorn his personal copies of books, it is clear that his prolific reading had a great impact on his works.

For example, a copy of *The Mind of Man,* a psychology textbook by Gustav Spiller from 1902, contains an intriguing note by Borges: "Memories of a lifetime, page 187." On this page, Spiller estimates how many memories a person has from different stages in a lifetime: around 100 for the first 10 years, 3,600 until 20 years, 2,000 more memories between the ages of 20 and 25, reaching about 10,000 in the first 35 years of life. He also states how much time it would take to recall these memories. For example, one does not remember every detail of a long trip, but instead certain landmark points — perhaps the moment of departure and arrival, or some stop in between. Borges says of Funes: "Two or three times he had reconstructed an entire day; he had never once erred or faltered, but each reconstruction had itself taken an entire day."
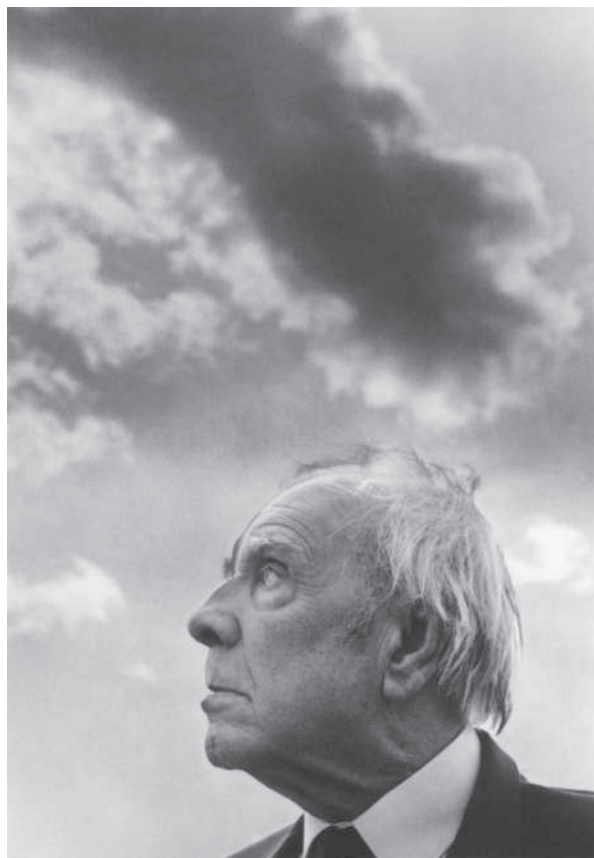
Jorge Luis Borges wrote of a protagonist who could remember everything — including the shapes of clouds on any particular day.

In the story of Funes, Borges described very precisely the problems of distorted memory capacities well before neuroscience caught up. We now know that memory function is linked to a particular brain area, the hippocampus, which lies at the end of the neural pathway that processes sensory information. Much of this knowledge came from the study of Patient H. M., who, in the 1950s, had his hippocampus surgically removed to cure him of epilepsy. Although he initially seemed to be normal after the surgery, it soon became apparent that he had developed anterograde amnesia: he could recall people and facts from before the surgery, but not recent events.

Evidence from Patient H. M. suggests that the hippocampus is crucial for the formation of new memories. In a study using electrodes to probe this brain region in epileptic patients for clinical reasons, we identified a type of neuron that fires in response to particular abstract concepts (R. Quian Quiroga *et al. Nature* **435,** 1102–1107; 2005). For example, one neuron in a patient fired only in recognition of different pictures of the actress Jennifer Aniston; another responded only to images of another celebrity, Halle Berry. It is thus possible that these neurons link perception and memory by creating the abstract encoding we use to store memories — especially considering that we tend to remember concepts and forget irrelevant details. If these neurons are lacking, the ability to generate abstractions may be limited, leading to pathologies such as autism or characters like Funes.

Even without this scientific knowledge, Borges's intuitive description is sharp: Funes, he wrote, was "virtually incapable of general, platonic ideas … His own face in the mirror, his own hands, surprised him every time he saw them … To think is to ignore (or forget) differences, to generalize, to abstract. In the teeming world of Ireneo Funes there was nothing but particulars." ∎

Rodrigo Quian Quiroga is professor and head of bioengineering in the Department of Engineering at the University of Leicester, Leicester LE1 7RH, UK.
e-mail: rqqg1@leicester.ac.uk

F. SCIANNA/MAGNUM PHOTOS

Q&A: Peter Atkins on writing textbooks

The success of **Peter Atkins**'s classic textbook *Physical Chemistry* led him to trade research for full-time writing and teaching in the 1980s. In the first of a series of five interviews with authors who each write science books for a different audience, Atkins explains how the rewards for textbooks can be great, but the effort needed can affect your research.

What is the purpose of a textbook?
To inspire a mode of thought. Ideally, a book then becomes a companion for life.

What is a good science textbook?
One that is considerate to the students — one that tries to put itself in their position and to explain and interpret what is going on. The author has to imagine himself as a student, and think and then write down what the student might be puzzling over. It also needs a logical flow.

Did your own teaching experiences influence you?
Wholly. Being face-to-face with students grappling with concepts is a real education. They say "Yes, I understand"; but there's a dullness about their eyes. So you have another go, and a light goes on. That's the bit you remember when you come to write.

How do textbooks come into being? Who decides on the subject and scope?
You are constrained by the market. The publisher is alert to what the market needs. In my case, the publisher had received a not-very-good proposal for a physical chemistry book, and came to me and said, can you do better? Sales reps are frequently on the lookout for potential authors; instructors often approach a publisher with a proposal.

What makes you write?
It is an urge to communicate and share my vision of the world. I've never taken a course on writing. I have a lucid style, and I think I can trace that to two things: one, I read a lot of books at university; two, it is important to have a deeper understanding than the level one wishes to exposit, so that you have the intellectual muscle and confidence.

What are the main challenges?
Getting the level right. Seeking that balance between completeness and succinctness so the text can be used over a broad spectrum of classes and isn't too long. Major chemistry books have about 1,000 pages, and they grow. Not because of factual content, but because more worked examples are put in.

Your book is in its ninth edition — why so many revisions?
Revisions have a rhythm that's decided by the market. Higher level texts are on a 4-year cycle; lower level texts on a 3-year cycle. To revise a book takes a year of rewriting and a year of production. When I say rewrite, I mean rewrite — not just tart up. To maintain one's premier position, one has to take revision seriously.

Peter Atkins sees textbooks as a way to capture a mode of thinking.

Does the second-hand market reduce sales?
Yes — it undermines income streams. In the second year, purchases fall, and in the third year there are none at all. If the second-hand market could be eliminated, books would last longer than 3 years and could be cheaper. There is a way — to produce electronic books [e-books] and kill them after a year.

How has the online shift affected textbooks?
There are many advantages in the e-book universe. The interactivity is compelling. You have cheaper distribution. Publishers can set the licence to expire to thwart the second-hand market. But there are disadvantages. You lose the serendipity of a real book because browsing is difficult. In an e-book, you're out in a forest and don't know where the edge is, or even if there is an edge.

Is adding online content difficult?
To produce an e-book you have to be more of an impresario than an author. You have the pictures, the unfolding of different depths of information. It's an extraordinarily demanding task.

Does this make it harder to find good writers?
It is increasingly difficult for publishers to get people to write textbooks. Heads of department are adamant that their faculty members focus on research, so publishers are finding it difficult to commission people of the quality they need. That becomes a bigger problem when you've got the e-book dimension added in.

What are the rewards for so much effort?
There are financial rewards. Beyond that, there are real pleasures in knowing you are helping to mould peoples' attitudes. It's heart-warming that wherever you go in the world, there's a group of people who already know you.

Is it true that authors can make millions from textbooks?
The financial rewards can be considerable, especially if they lead to other textbook commissions. The largest potential market is the freshman population in North America, around 200,000 students, and authors of textbooks for basic courses can make serious amounts of money. More specialist and graduate markets are much smaller, in the thousands, so earnings are proportionately less.

For those intent on writing a textbook, what is your advice?
Of course, I'd like to be competition-free. But when I'm gone, the prospective textbook writer should realize that there are immense rewards, not only intellectual. ∎
Interview by **Nicola Jones**, a commissioning editor for *Nature*'s Opinion section.

Peter Atkins's *Physical Chemistry* was first published in 1978 by Oxford University Press.

L. CENDAMO/BLACK ARCHIVES, CAMERA PRESS

NEWS & VIEWS

BIOGEOGRAPHY

Washed up in Madagascar

David W. Krause

How, when and from where did Madagascar's unique mammalian fauna originate? The idea that the ancestors of that fauna rafted from Africa finds support in innovative simulations of ancient ocean currents.

The biogeographic origins of Madagascar's terrestrial mammals have befuddled natural historians for centuries. Are they relics left over from when Madagascar lay at the centre of the Gondwanan supercontinent 165 million years (Myr) ago? Or did they simply disperse to the island from mainland Africa at a later time by walking across a causeway exposed by low sea levels? Alternatively, did they raft or swim across the Mozambique Channel, a seemingly insuperable barrier now 430 kilometres wide at its narrowest point? Ali and Huber's sophisticated modelling[1] of Indian Ocean currents during the Eocene epoch, roughly 56–34 Myr ago, adds a big new piece to the puzzle. Their analysis (page 653 of this issue) complements evidence from molecular systematics, the fossil record and palaeogeography, and strongly favours a rafting scenario.

With the exception of a few species recently introduced by humans, such as bushpigs, rats and mice, all of Madagascar's fully terrestrial mammalian species are found nowhere else. They belong to only four groups of placental mammal: rodents, carnivorans, tenrecs and the island's signature group, lemurs.

The fossil record is the most logical place to start when addressing questions of biogeographic origins. Unfortunately, the record of identifiable Malagasy terrestrial mammals from the entire Cenozoic era, the past 65 Myr, is virtually non-existent; it dates back only a few thousand years. The small sample of Malagasy terrestrial mammals from pre-Cenozoic strata does not include any placental mammals and therefore no potential ancestors of the extant taxa. However meagre this negative evidence might be, it suggests that the ancestral stocks arrived on Madagascar well after that landmass had rifted from Africa, and is entirely consistent with recent molecular dating of the times of divergence and diversification of each of the extant Malagasy groups. Molecular systematists have concluded that each of the four groups was derived from a single ancestor and that their closest relatives are African. By using statistical methods for estimating divergence ages, they have also concluded that the founding populations of each group probably arrived on the island independently at different times in the
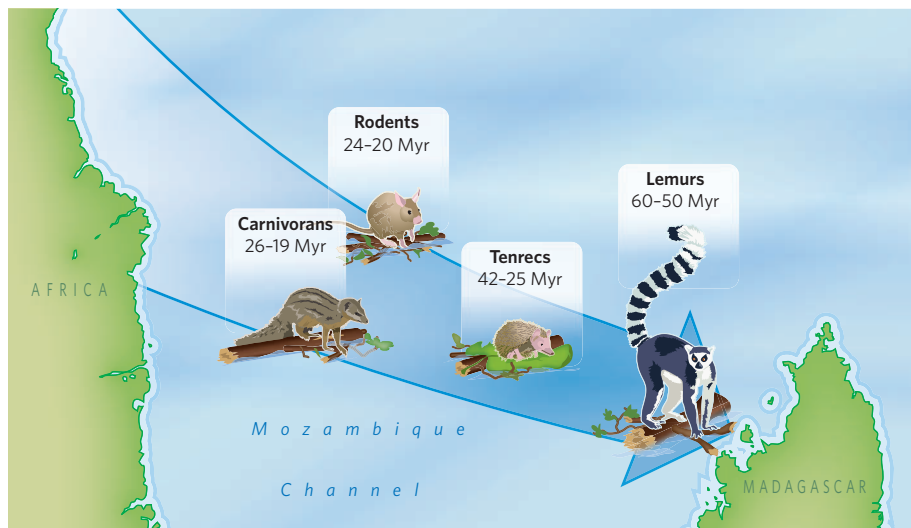
Figure 1 | By raft from Africa. On the evidence of molecular systematics, ancestors of the four endemic groups of placental mammal colonized Madagascar in single, separate events — lemurs 60–50 Myr ago, tenrecs 42–25 Myr ago, carnivorans 26–19 Myr ago and rodents 24–20 Myr ago[2]. Ali and Huber's ocean-current modelling studies[1] support Simpson's hypothesis[3] that they came by sea, on large rafts of vegetation. Examples of extant groups thought to have been unable to cross the Mozambique Channel in this way include elephants, zebras, antelopes and apes.

early–mid Cenozoic, between about 60 Myr and 20 Myr ago[2] (Fig. 1).

George Gaylord Simpson, arguably the most influential palaeontologist of the twentieth century, essentially surmised all of this 70 years ago[3]. Although he was steadfastly resistant to the 'mobilist' paradigm of plate tectonics, and was working well before the dawn of molecular systematics, Simpson noted the marked differences in composition between the extant mammals of Africa and Madagascar, and reasoned that Madagascar's mammals could not have used a land bridge to get there. Had a land bridge existed, he argued, Madagascar would have been populated by a greater diversity of 'typical' African mammals, both large and small. Instead, the Malagasy mammalian fauna is unique, taxonomically unbalanced, and decidedly on the small side.

But there have been persistent doubters. Some have argued that the ancestral stocks of Madagascar's terrestrial mammals must have been on the island while it was still connected to Africa in the Late Jurassic[4,5], some

100 Myr before the time at which molecular evidence tells us that lemurs diverged from other primates. In the early twentieth century, biogeographers argued for the existence of a land bridge, a view that recently resurfaced and gained considerable traction[6]. Others stipulated that the lengths of time necessary to cross the Mozambique Channel would have required ecophysiological specializations for hibernation or torpor[7]. And still others modelled modern-day ocean currents, explicitly assuming that they were the same throughout most of the Cenozoic, and claimed that it was impossible for mammals to have rafted from Africa to Madagascar because the currents could not have facilitated west to east transport across the Mozambique Channel[8].

Enter Ali and Huber[1]. Aware of the molecular and palaeontological evidence, and reviewing the lack of geological evidence for a Cenozoic land bridge between Africa and Madagascar, they provide a detailed model of surface-water currents in the ancient southwest Indian Ocean, controlling for large-scale winds,

palaeotopography, the positions of shifting landmasses and the opening of oceanic gateways. They discovered that the Eocene currents were not the same as those of today. Instead, during the Eocene (and probably throughout the ensuing Oligocene and into the Miocene), flow between Africa and Madagascar passed from northwest to southeast, from Tanzania and northeastern Mozambique towards northern Madagascar. Furthermore, these currents were episodically strong, conservatively estimated to have been more than 20 centimetres per second during certain time periods, a speed that would have transported large rafts of matted vegetation torn from the banks of East African rivers to Madagascar in a mere 25–30 days.

Based on admittedly preliminary data from another simulation at 15 Myr ago, Ali and Huber further argue that sometime during or after the early Miocene, Madagascar moved northwards into the equatorial gyre, and the regional current system shifted to its modern configuration. Current directions in the Mozambique Channel became directed essentially southwards, as today, thwarting further colonization of the island from Africa.

Doubts will persist, as they should. A red flag, for instance, is thrown up by studies of other taxa suggesting late Cenozoic dispersal from Africa to Madagascar[2]. In most cases, however, timings are uncertain — and timing is everything when testing biogeographic hypotheses. The discovery of additional Cenozoic fossils of terrestrial vertebrates and production of other molecular phylogenies, with associated divergence ages, will be essential. But so too will be the development of a stronger palaeogeographic and palaeo-oceanographic framework. Ali and Huber's modelling, for instance, was limited to a single slice of time in the Eocene. Acknowledging that each fully coupled simulation takes 1–3 years to run on a large, dedicated computer cluster, we can look forward to an increased density of palaeocurrent-direction simulations before, during and after the purported oceanic dispersals.

Nonetheless, Simpson's rafting hypothesis remains very much afloat. As unlikely as it might seem that rafts supporting terrestrial organisms could cross expansive marine barriers and establish founding populations, accumulating independent evidence from molecular dating, the fossil record and geology indicates that such feats probably explain an array of cases: the shared flora of New Zealand and Tasmania; flightless insects and birds on the Chatham Islands; mantellid frogs on the Comoros island archipelago; *Anolis* lizards in the Bahamas; iguanid lizards on the Fiji Islands; and platyrrhine monkeys in South America[9]. In this context, we must constantly remind ourselves of the immensity of geological time. As Simpson himself pointed out[10], "any event that is not absolutely impossible ... becomes probable if enough time elapses". ■

David W. Krause is in the Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794-8081, USA.
e-mail: david.krause@stonybrook.edu

1. Ali, J. R. & Huber, M. *Nature* **463,** 653–656 (2010).
2. Yoder, A. D. & Nowak, M. D. *Annu. Rev. Ecol. Evol. Syst.* **37,** 405–431 (2006).
3. Simpson, G. G. *J. Wash. Acad. Sci.* **30,** 137–163 (1940).
4. Masters, J. C., Rayner, R. J. & Tattersall, I. in *Creatures of the Dark: The Nocturnal Prosimians* (eds Alterman, L., Doyle, G. A. & Izard, M. K.) 31–44 (Plenum, 1995).
5. Heads, M. *Zool. Scripta* doi:10.1111/j.1463-6409.2009.00411.x (2009).
6. McCall, R. A. *Proc. R. Soc. Lond. B* **264,** 663–665 (1997).
7. Kappeler, P. M. *Folia Primatol. (Basel)* **71,** 422–425 (2000).
8. Stankiewicz, J., Thiart, C., Masters, J. C. & de Wit, M. J. *J. Biogeogr.* **33,** 221–235 (2006).
9. de Queiroz, A. *Trends Ecol. Evol.* **20,** 68–73 (2005).
10. Simpson, G. G. *Bull. Am. Mus. Nat. Hist.* **99,** 163–176 (1952).

PHOTOSYNTHESIS

Quantum design for a light trap

Rienk van Grondelle and Vladimir I. Novoderezhkin

The photosynthetic apparatus of cryptophyte algae is odd — its pigments are farther apart than is expected for efficient functioning. A study into how this apparatus works so well finds quantum effects at play.

It is common knowledge that plants, algae and certain bacteria use photosynthesis to convert solar energy into a form that can be used by the organisms to live and reproduce. But what is less well known is that the efficiency of photosynthesis might depend in part on quantum-mechanical processes. On page 644 of this issue, Collini *et al.*[1] report evidence suggesting that a process known as quantum coherence 'wires' together distant molecules in the light-harvesting apparatus of marine cryptophyte algae. This is the first time that this phenomenon has been observed in photosynthetic proteins at room temperature, rather than at much lower temperatures, bolstering the idea that quantum coherence influences light harvesting *in vivo*.

At a molecular level, the light-harvesting proteins (antennas) of photosynthetic organisms absorb solar photons, which excite electrons in pigment molecules. The antennas guide the resulting excitation energy to complexes of proteins known as reaction centres. The excited reaction centres then drive an ultrafast charge-separation process across a membrane, initiating a host of biochemical events that generate chemical energy. Remarkably, the energy transfer from the antennas to the reaction centre is almost perfectly efficient[2].

But why are light-harvesting antennas needed at all? Couldn't the reaction centres simply respond directly to incoming photons? There are several answers to these questions. First, photosynthesis must be able to operate at low light levels, such as those that generate less than one electronic excitation per molecule of chlorophyll per second. Yet the most important biochemical reactions associated with photosynthesis require several electron-transfer events. For example, water oxidation — a light-induced reaction in which water molecules split apart to generate protons, electrons and oxygen — requires the cumulative effect of four electronic excitations, all of which must occur within a certain time. Antennas overcome this problem by concentrating the available light energy, feeding the electronic excitations from hundreds of light-absorbing pigment molecules into a single reaction centre.

Another reason for having antennas is that they allow photosynthetic organisms to survive using fewer reaction centres. This is beneficial because reaction centres are 'expensive' — each one requires a large investment of resources from the host organism. Antennas also allow a broad range of the spectrum to be exploited for photosynthesis, because antennas that contain different pigments (and which therefore absorb different colours of light) can be connected to one reaction centre.

Finally, in a multi-protein light-harvesting antenna, the flow of excitation energy can be regulated by modulating the quenching properties of one of the constituent proteins. This provides a way of protecting plants from potentially harmful absorbed energy from excess sunlight[3].

So what is it that makes the light-harvesting process so efficient? There are many contributing factors. The first is that the pigments in light-harvesting complexes are optimally spaced — just closely enough to enable fast energy transfer[2], but far enough apart to prevent the molecular orbitals of the pigments from overlapping, which would quench their excited states (a phenomenon known as concentration quenching)[4]. The second factor is the supramolecular organization of the photosynthetic apparatus, which allows a multitude of energy-delivery pathways to connect to the reaction centre[5]. Collini *et al.*[1] now suggest that quantum coherence could be a third factor in optimizing energy-transfer efficiency.

The authors investigated two kinds of antenna complex taken from cryptophyte algae. Cryptophyte antennas exhibit exceptional spectral variation between species, largely because the structure of their main
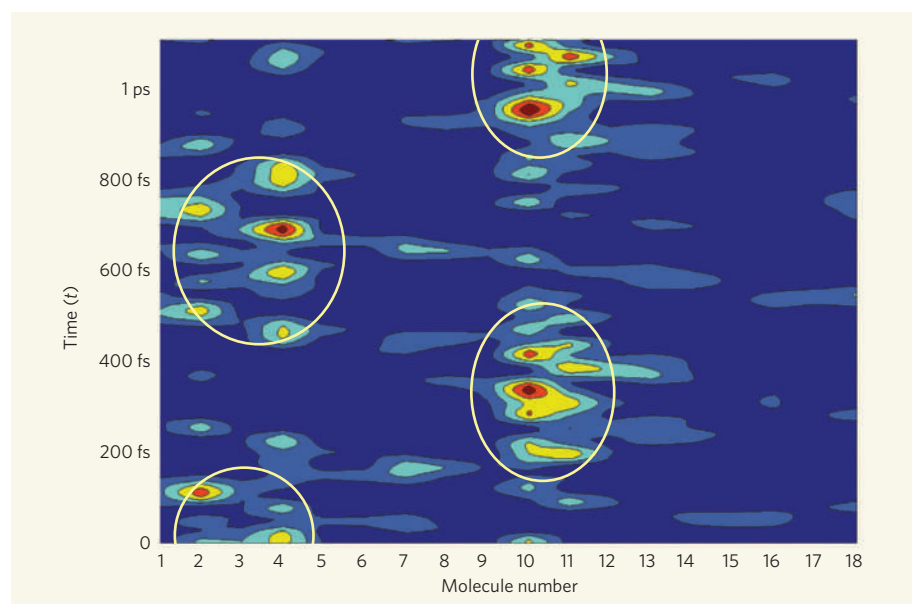
Figure 1 | Calculated dynamics of excited electronic states in a photosynthetic light-harvesting complex[6,7]. The LH2 light-harvesting complex from the bacterium *Rhodopseudomonas acidophila* contains 18 pigment molecules arranged in a circle[12], which enter excited states on absorption of photons. Here, the probability that an excitation (induced by a laser pulse) resides at a certain position within the complex is plotted against time (t); $t = 0$ corresponds to the time the excitation was induced, and the pigment molecules are numbered from 1 to 18. The colour scale ranges from dark blue (zero probability) to red (maximum probability). The excitation oscillates between small groups of molecules (ringed) in the complex with a half-period of about 350 fs, in a manifestation of a quantum effect known as coherence. Faster oscillations between individual molecules also occur. Collini *et al.*[1] observe similar oscillatory behaviour in the excitation dynamics of an algal light-harvesting system.

light-harvesting pigment (bilin) can be tuned to absorb different light frequencies. The authors excited the complexes using a short laser pulse (25 femtoseconds; 1 fs is 10^{-15} seconds), thus creating a superposition of excited electronic states known as a wavepacket, which evolves in time according to the laws of quantum mechanics. These laws predict that the wavepacket will behave in an oscillatory manner between the positions at which the excitation is localized, with distinct correlations and anti-correlations in phase and amplitude[6,7], analogous to a collection of oscillating pendulums connected by weak springs (Fig. 1). This coherent, collective behaviour would eventually disappear because of interactions between the individual excited pigments and their 'noisy' protein environment.

Such oscillatory behaviour in response to laser excitation has been observed before[8] in the light-harvesting complex of green sulphur bacteria. However, those experiments were performed at cryogenic temperatures, at which the interaction of excited pigments with their environment is generally much smaller than at room temperature.

Collini *et al.*[1] measured oscillating excitation dynamics in their algal antenna complexes using a technique called two-dimensional photon-echo spectroscopy. In this technique, the antenna under study is excited using a pair of short laser pulses, thus populating the different electronic levels of the system and creating electronic coherences (wave-like behaviour)

between the different levels. The evolution of the excited state is then measured by following the spectral response of the 'photon echo' — emitted light that is induced by a delayed third pulse. By plotting the emission wavelength as a function of the excitation wavelength, a two-dimensional spectrum is obtained. The extent to which exciting the system at one frequency provokes the appearance of another frequency in the emission spectrum directly reflects the degree of coupling between pigments.

Collini *et al.* not only observed pronounced oscillations in the diagonal peaks and off-diagonal cross-peaks of their spectra, as would be expected if quantum coherence occurred in the antenna, but they also found that the lifetime of these oscillations was surprisingly long (more than 400 fs; lifetimes of less than 100 fs were expected). What's more, the coherences occurred between pigments that are distant from each other, and so weakly coupled. Such an effect was also recently suggested to occur between the weakly coupled pigments of bacterial light-harvesting complexes[9,10].

Long-lived quantum coherences in a bacterial light-harvesting complex have previously been observed[8], but these were seen only at cryogenic (<77 kelvin) temperatures. By contrast, Collini and colleagues' experiments[1] were performed at room temperature, suggesting that long-lived quantum coherence might have a role in real-world light-harvesting processes.

If so, what could this role be? The correlated wave-like motion of an excitation that results

from quantum coherence allows the excitation to 'memorize' its earlier location. This is very different from the random hopping motion associated with classic mechanisms of energy transfer. Engel *et al.*[8] have speculated that quantum coherence allows antennas to search for the lowest-energy state of the complex more efficiently than they could using classical mechanisms; this, in turn, might enhance the efficiency of energy transfer towards a reaction centre. It has also been argued[11] recently that long-lived quantum coherence might help excitations to avoid local 'traps' (energy minima) that they encounter in the energy landscape on their way to a reaction centre.

But Collini *et al.*[1] suggest that their algal antenna complexes are special. Unlike most photosynthetic pigments, the eight bilin pigments in each cryptophyte antenna are covalently bound to the proteins of the complex. The authors speculate that this covalent attachment helps prevent the loss of quantum coherence. This, in turn, would explain why the lifetime of quantum coherence at room temperature is so long in cryptophyte antennas. The researchers also propose that quantum coherence might 'wire' together the final energy acceptors in cryptophyte antennas, thus compensating for the relatively weak electronic couplings between the pigments in the complex.

Much work clearly remains to be done. For example, it would be interesting to establish quantitatively whether or not photosynthetic light harvesting that involves quantum coherence is truly more efficient than it would be using classical mechanisms of energy transfer alone. But for now, Collini *et al.* have shown that quantum coherence can't be ignored in the search for explanations of the remarkable efficiency of photosynthetic energy-transfer processes. ■

Rienk van Grondelle is in the Department of Biophysics, Faculty of Sciences, VU University, de Boelelaan 1081, Amsterdam NL-1081 HV, the Netherlands. Vladimir I. Novoderezhkin is at the A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Leninskie Gory 1, Moscow 119992, Russia.
e-mails: rienk@few.vu.nl;
novoder@libro.genebee.msu.ru

1. Collini, E. *et al. Nature* **463,** 644–647 (2010).
2. van Grondelle, R., Dekker, J. P., Gillbro, T. & Sundstrom, V. *Biochim. Biophys. Acta* **1187,** 1–65 (1994).
3. Ruban, A. V. *et al. Nature* **450,** 575–578 (2007).
4. Beddard, G. S. & Porter, G. *Nature* **260,** 366–367 (1976).
5. Sener, M. K., Olsen, J. D., Hunter, C. N. & Schulten, K. *Proc. Natl Acad. Sci. USA* **104,** 15723–15728 (2007).
6. Novoderezhkin, V. I., Rutkauskas, D. & van Grondelle, R. *Biophys. J.* **90,** 2890–2902 (2006).
7. van Grondelle, R. & Novoderezhkin, V. I. *Phys. Chem. Chem. Phys.* **8,** 793–807 (2006).
8. Engel, G. S. *et al. Nature* **446,** 782–786 (2007).
9. Novoderezhkin, V., Wendling, M. & van Grondelle, R. *J. Phys. Chem. B* **107,** 11534–11548 (2003).
10. Zigmantas, D. *et al. Proc. Natl Acad. Sci. USA* **103,** 12672–12677 (2006).
11. Ishizaki, A. & Fleming, G. R. *Proc. Natl Acad. Sci. USA* **106,** 17255–17260 (2009).
12. McDermott, G. *et al. Nature* **374,** 517–521 (1995).

STEM CELLS

Big roles for small RNAs

Frank J. Slack

Embryonic stem cells can create copies of themselves, but can also mature into almost any type of cell in the body. Tiny gene regulators called microRNAs are now shown to have a role in directing these properties.

The single-celled, fertilized embryo is the source of all the trillions of specialized cells in our bodies, and the means by which this proliferation and specialization occur has fascinated developmental biologists for centuries. Moreover, the ability to artificially force mature, differentiated cells back into this naive state (called reprogramming) has huge potential in providing tools for regenerative medicine and for increasing our understanding of development. The work described on page 621 of this issue by Melton, Judson and Blelloch[1] indicates that small RNAs have a crucial role in stem-cell biology and therefore in future stem-cell-based therapies.

The cells that form soon after fertilization — embryonic stem cells — have a remarkable ability to divide rapidly, to make copies of themselves (self-renew) and to differentiate into any type of specialized cell — a property known as pluripotency. Once cells have differentiated, they express a set of genes encoding factors that inhibit self-renewal and that seal the differentiated state. Our understanding of the nature of this switch between pluripotent and differentiated states is still rudimentary, but several important regulatory genes are known to be required for the process. Moreover, when artificially expressed in differentiated cells (for instance, skin cells), a subset of these regulatory genes, which encode proteins known as 'stemness' factors, can reprogram the cells back to the pluripotent state[2,3], generating what are known as induced pluripotent stem cells (iPSCs). In an interesting convergence of research, multiple lines of evidence point to a type of RNA, called microRNA, as an essential factor in this switch.

MicroRNAs, as their name suggests, are tiny RNA molecules that are encoded in our genome. MicroRNAs are not translated into protein; their function is to regulate gene expression[4] by binding to other RNAs, particularly messenger RNAs. Binding of microRNA to mRNA inhibits mRNA translation to protein. In humans, thousands of microRNAs regulate thousands of mRNAs in a complex network.

MicroRNAs were first discovered in the nematode worm *Caenorhabditis elegans*, when mutations in the microRNA genes *lin-4* and *let-7* were found to result in defective stem-cell maturation[5,6]. Specifically, in the absence of these genes, *C. elegans* epithelial stem cells known as seam cells failed to exit from the self-renewing state and to differentiate.
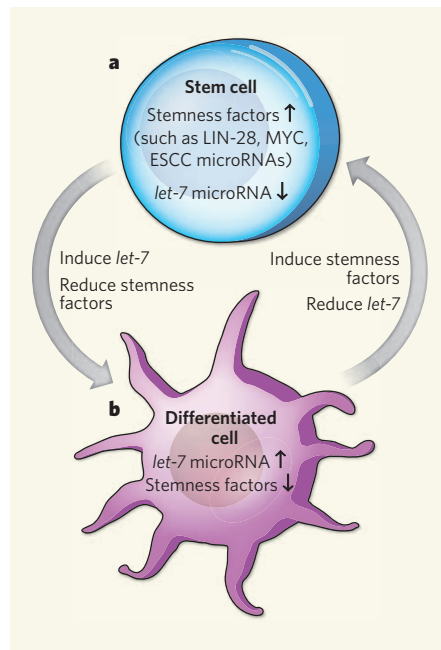
Figure 1 | Proposed roles of microRNAs in embryonic stem cells. a, Embryonic stem cells express high levels of 'stemness' factors, such as ESCC microRNAs and the proteins LIN-28 and MYC. Melton, Judson and Blelloch[1] now report that mouse embryonic stem cells also express *let-7* microRNAs at low levels. This microRNA must be induced to trigger cell differentiation, and seems to be essential for reducing the concentrations of the stemness factors. **b**, Differentiated cells, however, express low levels of the stemness factors and high levels of *let-7*. To dedifferentiate cells, stemness factors must be available and *let-7* must be inhibited.

Further work has revealed that microRNAs are involved in almost every biological process — including development, metabolism and ageing — in all multicellular organisms[4]. Moreover, microRNAs have been identified as being involved in many human diseases, most notably cancer[7,8].

The *let-7* microRNA is expressed in seam cells and mammalian stem cells[9] just before the cells differentiate. Thus, it has been proposed that *let-7* is a conserved anti-stemness and pro-differentiation factor. Its mechanism of action, however, had remained elusive. Melton *et al.* now demonstrate[1] that *let-7* is a key factor in inducing the differentiation of mouse embryonic stem cells, and it seems to be essential for depleting the cell of several stemness factors whose encoding genes and mRNAs are enriched in *let-7* binding sites. Decreasing the amounts of these stemness factors promotes differentiation (Fig. 1). In addition, the authors identify another family of microRNAs that prevents *let-7* from promoting differentiation. These findings form the basis of a model proposing that different types of microRNA have opposing effects on the fate of embryonic stem cells, one type promoting self-renewal and the other promoting differentiation.

In *C. elegans*, a mutation of the *lin-28* gene results in premature differentiation of stem cells[10], indicating that the LIN-28 protein is also a stemness factor. Recent work[11–15] has shown that LIN-28 is an RNA-binding protein that inhibits the processing and maturation of *let-7* microRNA in both mammals and *C. elegans* (microRNAs are modified by various proteins before they become fully functional). Moreover, LIN-28 is needed for efficient iPSC production[2]. If one 'joins the dots', it's not difficult to imagine that one of the roles of LIN-28 in iPSC production may be to suppress the maturation of *let-7* microRNA[16]. The current work[1] shows that inhibition of *let-7* promotes iPSC production from mature cells, supporting this model and highlighting the importance of this microRNA in maintaining the differentiated state (Fig. 1).

Melton and colleagues' observations[1] raise interesting prospects for regenerative medicine. First, they suggest that transient manipulation of microRNA levels might be a preferred route for generating iPSCs. The authors' findings will also interest workers in the cancer field, in which there is growing support for an emerging hypothesis that mutant stem cells are responsible for many cancers. The work[1] surely reinforces the strategy of manipulating the amounts of such microRNAs in tumours as anticancer therapy. These smallest of RNA molecules seem to have a big future ahead. ∎

Frank J. Slack is in the Department of Molecular, Cellular and Developmental Biology, Yale University, PO Box 208103, New Haven, Connecticut 06520, USA.
e-mail: frank.slack@yale.edu

1. Melton, C., Judson, R. L. & Blelloch, R. *Nature* **463**, 621–626 (2010).
2. Yu, J. *et al. Science* **318**, 1917–1920 (2007).
3. Wernig, M. *et al. Nature* **448**, 318–324 (2007).
4. Bartel, D. P. *Cell* **116**, 281–297 (2004).
5. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75**, 843–854 (1993).
6. Reinhart, B. J. *et al. Nature* **403**, 901–906 (2000).
7. Esquela-Kerscher, A. & Slack, F. J. *Nature Rev. Cancer* **6**, 259–269 (2006).
8. Yu, F. *et al. Cell* **131**, 1109–1123 (2007).
9. Wulczyn, F. G. *et al. FASEB J.* **21**, 415–426 (2007).
10. Moss, E. G., Lee, R. C. & Ambros, V. *Cell* **88**, 637–646 (1997).
11. Lehrbach, N. J. *et al. Nature Struct. Mol. Biol.* **16**, 1016–1020 (2009).
12. Viswanathan, S. R., Daley, G. Q. & Gregory, R. I. *Science* **320**, 97–100 (2008).
13. Rybak, A. *et al. Nature Cell Biol.* **10**, 987–993 (2008).
14. Newman, M. A., Thomson, J. M. & Hammond, S. M. *RNA* **14**, 1539–1549 (2008).
15. Heo, I. *et al. Mol. Cell* **32**, 276–284 (2008).
16. Nimmo, R. A. & Slack, F. J. *Chromosoma* **118**, 405–418 (2009).

EXTRASOLAR PLANETS

Fluorescent methane spotted

Seth Redfield

The atmospheric properties of distant worlds are becoming increasingly clear. The latest observations reveal fluorescent emission from methane in the upper atmosphere of a Jupiter-like extrasolar planet.

The first extrasolar planet found to be orbiting a Sun-like star was detected less than 15 years ago. As astronomers detect more and more planets orbiting stars other than the Sun in our Galactic neighbourhood, increasing attention is being paid to probing their atmospheres, and for good reason. Planetary atmospheres can be easily altered by geophysical[1], photochemical[2] and biological[3] processes. Owing to the relatively small amount of mass in an atmosphere, its properties can bear the signatures of processes driven from the planetary interior or surface that would not otherwise be observable. A notable example is the presence of the powerful greenhouse gas methane (CH_4) in planetary atmospheres.

In Earth's atmosphere, the dominant sources of non-anthropogenic methane are anaerobic bacteria and methanogens, which inhabit wetland and oceanic sediments, and the digestive tracts of some organisms (for example, ruminants and termites). Methanogens, like most archaeal microorganisms, can thrive in a wide range of conditions, including many that would be harmful to the bulk of complex life forms on Earth. Jupiter's methane, meanwhile, is photochemical in origin, and so is dependent on the radiation field incident on the atmosphere and the specific abundances of carbon, oxygen and hydrogen[4].

To disentangle the various processes that affect an atmosphere, it is crucial to obtain as much information as possible on a diverse sample of planetary atmospheres. On page 637 of this issue, Swain *et al.*[5] report the discovery of fluorescent emission from methane in the upper atmosphere of a nearby, Jupiter-mass extrasolar planet[6], HD 189733b (Fig. 1).

Fluorescence occurs when an atom or molecule absorbs a photon, is excited into a higher energy state and subsequently de-excites, emitting light at lower energies. It requires the relatively low particle densities that occur high in planetary atmospheres, where the time between collisions is longer than the time required for radiative relaxation. Methane, like all molecules, can only make such transitions among specific permitted electronic, rotational and vibrational energy pathways that are unique to its molecular structure. The

Figure 1 | Artist's impression of HD 189733b. Swain and colleagues' ground-based observations[5] of the Jupiter-mass extrasolar planet HD 189733b reveal fluorescent emission from methane in the planet's upper atmosphere. The work highlights the value of complementary ground- and space-based studies of such planets.

ESA/NASA/G. TINETTI/M. KORNMESSER

50 & 100 YEARS AGO

BIOMATERIALS
Dew catchers

Why did Incy Wincy Spider climb up the water spout? If he was after a drink, a report by Yongmei Zheng et al. in this issue suggests that he might have missed a trick — spiders don't need to look for water because the silk fibres that they spin are highly efficient at collecting it from moist air (Y. Zheng et al. Nature 463, 640–643; 2010).

To uncover the water-collection mechanism, Zheng et al. exposed silk fibres of the cribellate spider Uloborus walckenaerius to humid air, and monitored them with a scanning electron microscope. They observed that initial contact with water causes the hydrophilic fibres to re-structure, so that spindle-knots form periodically along the thread axis, separated by elongated joints roughly fourfold thinner.

Small water droplets then condense randomly onto the spindle-knots and joints, and begin to grow as they accumulate moisture from the air. When they reach a critical size, droplets that are attached to joints move towards the nearest spindle-knots, where they coalesce to form larger water drops. This leaves the joints free to start a new cycle of water condensation and collection.

Zheng et al. found that the structure of the silk fibres is crucial for water collection. They observed that the spindle-knots are made up of randomly arranged nanofibrils and have rough

C. VARNDELL/NATUREPL.COM

surfaces, whereas the joints consist of aligned nanofibrils and have smooth surfaces. The difference in roughness causes a surface-energy gradient that drives water towards the spindle-knots. This effect is boosted by the different shapes of the knots and the joints, which cause the drops to deform differently, thereby generating surface-tension forces that also drive water drops towards the spindle-knots. A final helping hand is provided by the different orientations of the nanofibrils in the fibres: water drops move easily along joint regions where the nanofibrils are aligned, but stick to the spindle-knots where the nanofibrils are randomly orientated.

The authors went on to make artificial fibres that mimicked the structure of spider silk, and found that these successfully reproduced the water-collection properties of the natural material. They speculate that their work will help in the design of fibres that could be used in devices that collect water from the atmosphere, or that remove liquid aerosols in manufacturing processes.

Magdalena Helmer

---

result is a characteristic spectrum of emission lines, and it is one of these emission lines, distinctive to methane, that has been detected by Swain and colleagues[5].

Fluorescent emission is relatively common in astrophysical environments. It has been detected, for example, in the accretion disks surrounding supermassive black holes[7], in the interstellar medium[8] and in comets[9]. Specifically, the fluorescent methane emission detected by Swain et al.[5] on HD 189733b has also been observed in the atmospheres of Jupiter, Saturn and Titan[10]. Fluorescence of other organic compounds has been detected on Venus and Mars[11]. These detections provide a probe of the physical structure of the upper atmosphere of these planets from the perspective of a minor atmospheric constituent. Methane is particularly important because it may help us to find and evaluate possible biological influences on extrasolar planetary atmospheres.

The upper atmosphere is a fascinating and important region where minor molecular constituents, such as methane, can play an indispensable part in establishing the overall heat budget of a planet, thereby altering the thermal profile of a considerable portion of the atmosphere. Powerful winds and vertical mixing of high-altitude atmospheric layers present the possibility for temporal and spatial variability of the fluorescent emission from such molecules. In addition, ionized particles in the upper atmosphere are affected by any global magnetic field — as is dramatically exemplified on Earth by another form of emission-line radiation, the aurorae. Because HD 189733b is extremely close to its host star (less than one-tenth the distance between Mercury and the Sun), energetic particles from its star will interact with any magnetic field the planet may have, possibly resulting in stronger emission-line displays than we see on Earth or Jupiter. Many other observed extrasolar planets are similarly close to their host stars, so variability in emission components because of magnetic effects may be common in these systems.

Swain and colleagues' detection[5] of a fluorescent emission line of methane on HD 189733b paves the way for future observations of alternative fluorescent emission lines of methane and other molecules in extrasolar planetary atmospheres. These observations will require sensitive, high-spectral-resolution instruments that can resolve the emission-line profile of fluorescent signals. In addition, this discovery highlights a crucial theme in modern astronomy: the necessity of making complementary observations from the ground and from space. Depending on the wavelengths of the relevant spectral features, Earth's own atmosphere may or may not interfere with detection. The authors' observations[5] were made with the 3.0-metre NASA Infrared Telescope Facility located at the summit of Mauna Kea in Hawaii. Our increasing understanding of this one extrasolar planet — from its discovery[6] to the growing list of known constituents of its atmosphere, which includes sodium[12], carbon monoxide[13], carbon dioxide[13], water vapour[13] and methane[14] — has been possible only with observations from both the ground and space.

During the past few years, we have made a transition that deserves some rumination. Rather than speculating about the possibility of other worlds in the cosmos, we can now identify them specifically and enumerate their various characteristics. Many stars that are readily visible to the naked eye, at least from relatively dark sites that are not heavily polluted by artificial light, have planets orbiting them, the masses and orbital characteristics of which we know. A number of other worlds, soon to be discovered, will be small and rocky like Earth, and will have atmospheres that we can detect, inventory and monitor. It is quite possible that, within our lifetimes, atmospheric studies of these extrasolar planets will provide the first evidence of biological life beyond Earth. Swain and colleagues' detection of the fluorescence of a hydrocarbon in the upper atmosphere of an extrasolar planet not only provides insight into the structure of the atmospheres of other worlds, but is also an important step in the far-reaching journey to uncover what may be below them. ■

Seth Redfield is in the Department of Astronomy, Van Vleck Observatory, Wesleyan University, Middletown, Connecticut 06459, USA.
e-mail: sredfield@wesleyan.edu

1. Kaltenegger, L. & Sasselov, D. Astrophys. J. 708, 1162–1167 (2010).
2. Gladstone, G. R. et al. Icarus 119, 1–52 (1996).
3. Segura, A. et al. Astrobiology 5, 706–725 (2005).
4. Yung, Y. L. & DeMore, W. B. Photochemistry of Planetary Atmospheres (Oxford Univ. Press, 1999).
5. Swain, M. R. et al. Nature 463, 637–639 (2010).
6. Bouchy, F. et al. Astron. Astrophys. 444, L15–L19 (2005).
7. Nandra, K. & Pounds, K. A. Mon. Not. R. Astron. Soc. 268, 405–429 (1994).
8. Shull, J. M. & Beckwith, S. Annu. Rev. Astron. Astrophys. 20, 163–190 (1982).
9. Schleicher, D. G. & A'Hearn, M. F. Astrophys. J. 331, 1058–1077 (1988).
10. Baines, K. H. et al. Earth Moon Planets 96, 119–147 (2005).
11. Deming, D. et al. Icarus 55, 347–355 (1983).
12. Redfield, S., Endl, M., Cochran, W. D. & Koesterke, L. Astrophys. J. 673, L87–L90 (2008).
13. Swain, M. R. et al. Astrophys. J. 690, L114–L117 (2009).
14. Swain, M. R., Vasisht, G. & Tinetti, G. Nature 452, 329–331 (2008).

APPLIED PHYSICS

Nanothermal trumpets

Rama Venkatasubramanian

The thermal process known as Joule heating, which often plagues electronic devices, has been turned to good use: making devices that can produce sound as well as reproduce music and speech.

James Prescott Joule, the nineteenth-century British physicist whose research on electricity and heat led to the first law of thermodynamics, is eponymous with Joule heating. This is the mechanism by which an electrical conductor converts electrical energy into heat. Joule heating has always been considered an inevitable loss term for the efficiency of electronic devices, except for domestic appliances such as electric heaters and cookers, which rely on it to function. Writing in *Applied Physics Letters*, Niskanen et al.[1] describe how Joule heating can be exploited to make efficient and controllable sources of sound.

The idea of converting Joule heating into sound is nearly a hundred years old. It was first proposed[2] by Arnold and Crandall, who called their sound devices thermophones. But progress in turning such thermoacoustic machines into efficient sound sources has been hampered by the high thermal conductivity of the metal wires (the electrical conductors) and of the supporting structure on which the wires are hinged or laid. Under these limitations, any Joule heat produced by the metal wires would leak out and therefore not contribute to the large temperature variation in the device that is required to produce sound.

Niskanen and colleagues' sound devices[1] consist of an array of parallel aluminium wires — as many as 200,000 wires in an area of a few square centimetres — suspended over a silicon substrate. The wires are clamped at their ends to aluminium crossbars using standard silicon microelectronic processing techniques. Each wire is about 200 μm long, 3 μm wide and only 30 nm thick. Using electric-power levels of 17 W, the authors demonstrate sound levels of more than 100 dB at a frequency of 20 kHz and a distance of 7 cm from these microscale sound sources. Such sound levels are impressive for thermoacoustics but considerably smaller than those obtained in conventional home and rock-concert speakers, which can blare out sound at 100 dB for 1 W of electric power to a distance of 1 m. Even so, Niskanen et al.[1] show that their devices can reproduce music and speech.

The mechanism that underlies sound production in the authors' devices[1] is, although similar in fundamental origin, unlike that of a musical instrument such as a trumpet, in which sound is produced by an air-pressure wave set up by the musician blowing into the instrument's mouthpiece. Instead, their devices — let's call them 'nanothermal trumpets' —
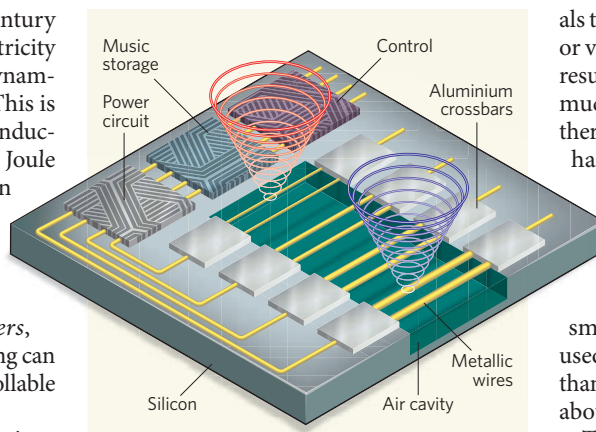

Figure 1 | A thermoacoustic integrated device. Niskanen and colleagues' sound device[1] is based on an array of parallel wires, nanometres thick, suspended over a silicon substrate and clamped at their ends to aluminium crossbars. It could be modified to incorporate further elements such as music-storage and control electronics. In this example, the array consists of several pairs of wires of different width, producing sound of different pitch (as denoted by the blue and red sound wavefronts). A power circuit supplies the array with electric power and causes temperature variations in the surrounding air that in turn generate air-pressure variations, and thus sound waves. A control element can either allow music to be stored or provide input directly to the power circuit. (Device not to scale.)

generate sound by means of temperature variations in the air surrounding the wire array. These air-temperature variations are induced by Joule heating in the wires when electric power is applied to the device, resulting in time-dependent air-pressure variations, and hence sound.

The low thermal conductivity required for the wires' support structure, which Niskanen and colleagues take advantage of in their experiment, was first achieved by Shinoda and colleagues[3]. These workers used an aluminium film on porous silicon, which had a thermal conductivity 1/168th that of crystalline silicon, to produce ultrasonic emission of frequencies up to 100 kHz. Later work by Xiao et al.[4] showed that ultrathin films of carbon nanotubes can be used to create loudspeakers over a wide range of frequencies. All these devices[1,3,4] have similar characteristics, being dependent on power and frequency and independent of crosstalk between adjacent wires. However, because it is fabricated on a silicon wafer, Niskanen and

colleagues' device will allow control and memory electronics to be integrated into it (Fig. 1). It could store music on a memory chip integrated on the same wafer, direct sound waves at a desired angle using 'phased arrays' of wires, and allow sound-wave interference in a particular pattern.

Progress in the field of thermoacoustics may be boosted by advances in the sister field of thermoelectrics, which seeks to use what is known as the Peltier effect to develop materials that efficiently convert heat into electricity, or vice versa. The past decade has witnessed a resurgence[5–8] of interest in this field. Although much of the work was aimed at reducing the thermal conductivity of semiconductors, there has also been growing interest in thermal transport in nanoscale metallic structures[9]. The thermal conductivity of ultrathin (less than 100 nm) metallic films has been predicted to be much smaller — more than a factor of five smaller for a 30-nm-thick film such as that used by Shinoda et al.[3] and Niskanen et al.[1] — than that of bulk structures (those thicker than about 1,000 nm).

The physics underlying the use of nanoscale metallic structures for thermoacoustics is not fully understood. Niskanen and colleagues mention the possible use of miniaturized thermoelectric coolers in their devices, presumably for additional temperature control. It would be worth considering making the metallic wires themselves out of a thermoelectric material, or cooling and heating the two ends of the wires with on-chip thermoelectric devices, thereby exploiting not only Joule heating but also Peltier heating and cooling. And the successful incorporation of nanoengineered thermoelectric devices onto silicon chips[8] indicates that such devices could be used to develop thermoacoustic integrated circuits.

Detailed knowledge of the thermal properties of nanoscale materials, complemented by modern techniques for fabricating materials and creating devices, will undoubtedly bolster research and development in thermoacoustics. Arnold and Crandall could hardly have imagined such achievements in the field nearly a century ago, and Niskanen and colleagues' sound source would undoubtedly have been music to James Prescott Joule's ears. ■

Rama Venkatasubramanian is at RTI International, Research Triangle Park, North Carolina 27709, USA.
e-mail: rama@rti.org

1. Niskanen, A. O. et al. Appl. Phys. Lett. **95,** 163102 (2009).
2. Arnold, H. D. & Crandall, I. B. Phys. Rev. **10,** 22–38 (1917).
3. Shinoda, H., Nakajima, T., Ueno, K. & Koshida, N. Nature **400,** 853–855 (1999).
4. Xiao, L. et al. Nano Lett. **8,** 4539–4545 (2008).
5. Venkatasubramanian, R., Siivola, E., Colpitts, T. & O'Quinn, B. Nature **413,** 597–602 (2001).
6. Hochbaum, A. I. et al. Nature **451,** 163–167 (2008).
7. Boukai, A. I. et al. Nature **451,** 168–171 (2008).
8. Chowdhury, I. et al. Nature Nanotech. **4,** 235–238 (2009).
9. Jin, J. S., Lee, J. S. & Kwon, O. Appl. Phys. Lett. **92,** 171910 (2008).

ARTICLES

Opposing microRNA families regulate self-renewal in mouse embryonic stem cells

Collin Melton[1,2], Robert L. Judson[1,2] & Robert Blelloch[1,2]

When embryonic stem cells (ESCs) differentiate, they must both silence the ESC self-renewal program and activate new tissue-specific programs. In the absence of DGCR8 (*Dgcr8*[−/−]), a protein required for microRNA (miRNA) biogenesis, mouse ESCs are unable to silence self-renewal. Here we show that the introduction of let-7 miRNAs—a family of miRNAs highly expressed in somatic cells—can suppress self-renewal in *Dgcr8*[−/−] but not wild-type ESCs. Introduction of ESC cell cycle regulating (ESCC) miRNAs into the *Dgcr8*[−/−] ESCs blocks the capacity of let-7 to suppress self-renewal. Profiling and bioinformatic analyses show that let-7 inhibits whereas ESCC miRNAs indirectly activate numerous self-renewal genes. Furthermore, inhibition of the let-7 family promotes de-differentiation of somatic cells to induced pluripotent stem cells. Together, these findings show how the ESCC and let-7 miRNAs act through common pathways to alternatively stabilize the self-renewing versus differentiated cell fates.

Mammalian development follows a carefully orchestrated unfolding of cell fate transitions leading to a complex set of highly specialized cell types. These cell fate transitions involve the silencing of previously active molecular programs along with the activation of new ones. miRNAs are small non-coding RNAs that are well suited to suppress previously active programs and, thereby, provide robustness to cell fate decisions[1,2]. miRNAs identify their targets by base pairing of nucleotides 2–8 of the miRNA (the seed sequence) with complementary sequences within the target mRNA's open reading frame (ORF) and 3′ untranslated region (UTR)[1]. This targeting is carried out in coordination with the RNA-induced silencing complex (RISC), and often results in both destabilization and translational inhibition of the targets. Although inhibition of any one target is usually only partial, each miRNA binds and suppresses hundreds of messenger RNA targets, resulting in large overall changes in the molecular constitution of cells.

Removal of genes required for maturation of all miRNAs has shown that miRNAs have essential roles in the proliferation and differentiation of ESCs[3–5]. For example, the loss of the RNA-binding protein DGCR8, which is required for the production of all canonical miRNAs, results in a cell cycle defect and an inability to silence the self-renewal program of ESCs when they are placed in differentiation-inducing conditions[3]. The introduction of individual members of a family of miRNAs, the ESCC miRNAs, into *Dgcr8*[−/−] ESCs can rescue the cell cycle defect[6]. These same miRNAs are able to enhance the de-differentiation of somatic cells to induced pluripotent stem cells[7]. Here we report the identification of another large family of miRNAs, the let-7 family, which performs the opposite role to the ESCC family. When introduced into *Dgcr8*[−/−] ESCs, let-7 silences self-renewal by suppressing many of the same downstream targets that are indirectly activated by the ESCC family. Indeed, co-introduction of the ESCC miRNAs inhibits the capacity of let-7 to silence self-renewal, and suppression of the let-7 family in somatic cells promotes de-differentiation.

ESCCs and let-7 regulate self-renewal

The let-7 miRNAs are broadly expressed across differentiated tissues[8,9] and are tightly regulated during ESC differentiation[10–14]. Therefore, we hypothesized that the let-7 miRNAs could rescue the capacity of *Dgcr8*[−/−] ESCs to silence ESC self-renewal when induced to differentiate. To test this hypothesis, we introduced mimics of a representative let-7 family member, let-7c, into the *Dgcr8*[−/−] ESCs (Fig. 1a). let-7c silenced the ESC self-renewal program even when the ESCs were maintained in ESC culture conditions. Three days after let-7c treatment, *Dgcr8*[−/−] cells downregulated ESC associated markers including alkaline phosphatase activity (Supplementary Fig. 1), Oct4 (also known as Pou5f1) immunofluorescence staining (Fig. 1b), and mRNA expression of *Oct4*, *Sox2* and *Nanog* (Fig. 1c). Furthermore, the transfected cells showed a diminished capacity to reform ESC colonies in replating assays, a functional test of ESC self-renewal capacity (Fig. 1d). Similar effects were observed with the introduction of let-7a, let-7b, let-7d and let-7g (Supplementary Fig. 2), and these effects were observed over a range of concentrations, including levels normally found in more differentiated cell types (Supplementary Fig. 3).

In contrast to the *Dgcr8*[−/−] ESCs, wild-type ESCs were resistant to let-7c-induced suppression of self-renewal (Fig. 1b–d and Supplementary Fig. 1). This finding indicated that other miRNAs normally expressed in wild-type ESCs inhibit the ability of let-7c to silence self-renewal. The ESCC miRNAs are probable candidates as they make up most of the miRNA molecules in mouse ESCs[15,16], they are rapidly downregulated after differentiation coincident with the upregulation of mature let-7 (Supplementary Fig. 4), and they promote the ESC fate[6,7,17,18]. Therefore, we introduced a representative member of this family, miR-294, to test whether it could block let-7c-induced suppression of *Dgcr8*[−/−] ESC self-renewal. Three days after co-introduction of miR-294 and let-7c, *Dgcr8*[−/−] ESCs retained alkaline phosphatase activity (Supplementary Fig. 1), Oct4 immunofluorescence staining (Fig. 1b), and mRNA expression of *Oct4*, *Sox2* and *Nanog* (Fig. 1c). Furthermore, miR-294 rescued the colony forming capacity of the *Dgcr8*[−/−] ESCs (Fig. 1d). Control miRNAs (miR-294 with a seed mutation and other ESC-expressed miRNAs, miR-291a-5p and miR-130b, which do not contain the ESCC miRNA seed sequence) did not antagonize the effects of let-7c (Fig. 1a–d), showing that the effect of miR-294 is not simply secondary to competition for RISC complexes.

[1]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, Center for Reproductive Sciences, Program in Biomedical Sciences, [2]Department of Urology, University of California San Francisco, San Francisco, California 94143, USA.
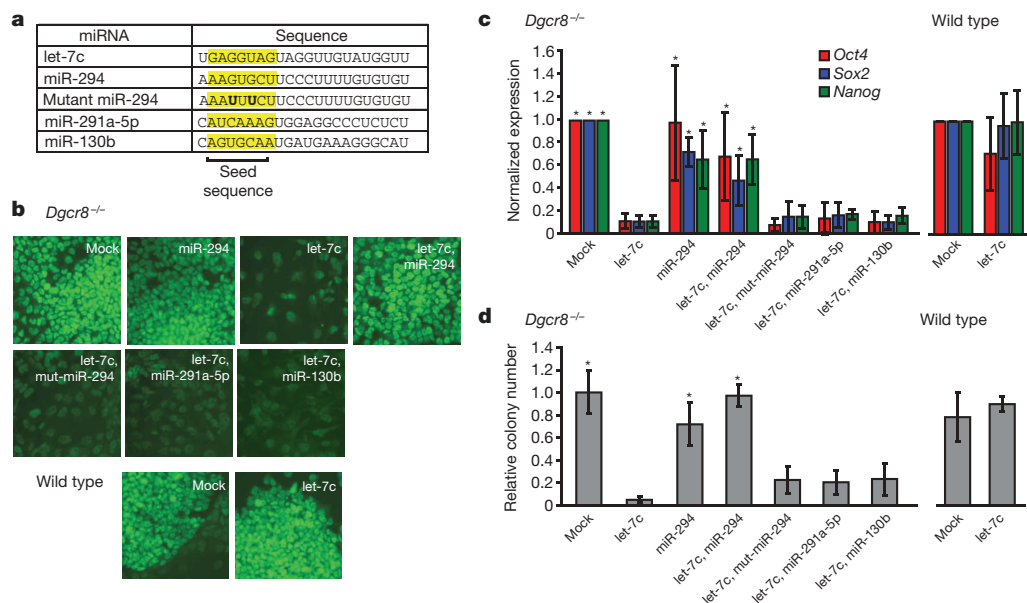
Figure 1 | The let-7 and ESCC miRNA families have opposing roles in regulating ESC self-renewal. a, Transfected miRNAs with the seed sequence highlighted. **b**, Oct4 immunofluorescence staining after transfection of let-7c, miR-294 and combinations of let-7c with miR-294, mutant (mut-) miR-294, miR-291a-5p, or miR-130b in *Dgcr8*[−/−] (top two rows) and wild-type (bottom row) ESCs. Representative images, *n* = 3. Original magnification,

×10. **c**, qRT–PCR for *Oct4*, *Sox2* and *Nanog* normalized to β-actin (*Actb*) after miRNA introduction as in **b**. *n* = 3–8. *$P < 0.02$. **d**, Colony reforming assays after miRNA introduction as in **b** and **c**. *n* = 3. *$P < 0.05$. All *P*-values were generated by Bonferroni-corrected *t*-test of comparisons to let-7c-treated. Error bars represent s.d.

Other members of the ESCC family miR-291a-3p, miR-291b-3p and miR-295 were similarly able to block the effects of let-7c (Supplementary Fig. 5). These data indicate that the let-7 and ESCC families of miRNAs have opposing roles in the maintenance of ESC self-renewal.

Targeting through ORFs and 3′UTRs

The functional antagonism between let-7c and miR-294 on ESC self-renewal suggested opposing roles for these miRNAs on downstream molecular targets. To test this prediction, we sought to identify these targets globally using mRNA microarrays after the introduction of let-7c or miR-294 into *Dgcr8*[−/−] ESCs. The introduction of the let-7c mimic led to downregulation of 693 and upregulation of 208 transcripts relative to mock treated cells, with a false discovery rate (FDR) of less than 5% (Fig. 2a and Supplementary Table 1). Of the 693 downregulated transcripts, 294 contained a let-7c 7-nucleotide seed match in the 3′UTR, 287 contained a 7-nucleotide seed match in the ORF, and 113 contained both 3′UTR and ORF 7-nucleotide seed matches (Supplementary Table 1). The presence of these seed matches in the downregulated transcripts was highly enriched compared to the entire gene set (Fig. 2b and Supplementary Fig. 6a). Similarly, the introduction of miR-294 led to a large number of upregulated and downregulated transcripts (Fig. 2c and Supplementary Table 1). Again, downregulated transcripts were enriched for seed matches in the 3′UTR and ORF. In contrast, upregulated transcripts were depleted for seed matches in the 3′UTR and ORF (Fig. 2d and Supplementary Fig. 6b). These findings suggest that miR-294 and let-7c functionally act through the downregulation of many targets by binding their ORF and/or 3′UTR.

Impact on ESC transcriptional network

To investigate further the mechanism for the opposing roles of let-7c and miR-294 on ESC self-renewal, we performed pathway analysis on the miRNA-regulated transcript sets. Specifically, we searched for overlaps between the miRNA-regulated transcripts and genes identified by chromatin immunoprecipitation (ChIP) of pluripotency associated transcription factors[15,19]. This analysis measures whether there is any influence of the let-7 or ESCC miRNAs on the transcription factors themselves (a and b in Fig. 3A, and Supplementary

Methods) or the transcripts originating from the genes bound by the transcription factors (c in Fig. 3A, and Supplementary Methods).

In ESCs, two Myc family members, N-Myc and c-Myc, are highly expressed and have largely overlapping ChIP target genes[19]. *c-myc* (also known as *Myc*) has previously been identified as a let-7 target in cancer cells[20], and we find that *N-myc* (also known as *Mycn*) is significantly downregulated by let-7c in our array data (Supplementary Table 1). Consistent with let-7 directly targeting the Myc family, overlapping of let-7c-regulated transcripts with Myc-bound genes showed an enrichment of Myc target genes in the let-7c-downregulated transcript set and a depletion in the let-7c-upregulated transcript set (see a in Fig. 3B). Furthermore, the enrichment was independent of the presence of seed sequence matches within the ORF or 3′UTR. This finding suggests that let-7 is acting directly through inhibition of Myc (*c-myc* and/or *N-myc*) rather than through inhibition of Myc's downstream target genes (see a in Fig. 3A).

Performing a similar analysis overlapping miR-294-regulated transcripts and Myc target genes showed the exact opposite pattern as the analysis with let-7c-regulated transcripts. There was depletion of Myc targets in the miR-294-downregulated transcript set and enrichment in the miR-294-upregulated transcript set (see b in Fig. 3B). This pattern suggests that miR-294 upregulates Myc activity (b in Fig. 3A). Indeed, microarray data showed that miR-294 markedly increased *c-myc* levels (Supplementary Table 1). As miR-294 itself suppresses its downstream targets (Fig. 2d), the upregulation of *c-myc* must be indirect, through an unknown intermediate repressor (b in Fig. 3A). These data show that the let-7 and ESCC families of miRNAs have opposing effects on Myc activity.

Overlap of the let-7c-regulated transcripts with ChIP target genes for the pluripotency transcription factors Oct4, Sox2, Nanog and Tcf3 once again showed enrichment among the let-7c-downregulated transcript set (see c in Fig. 3B). However, this enrichment was limited to the downregulated transcripts with seed matches in their ORF or 3′UTR. These data suggest that rather than directly regulating the pluripotency transcription factors, let-7 targets transcripts originating from the genes bound by them (c in Fig. 3A). This pattern of enrichment is most clear for the ChIP target genes bound by Tcf3, cobound by Oct4, Sox2 and Nanog, or bound by the pluripotency cluster of
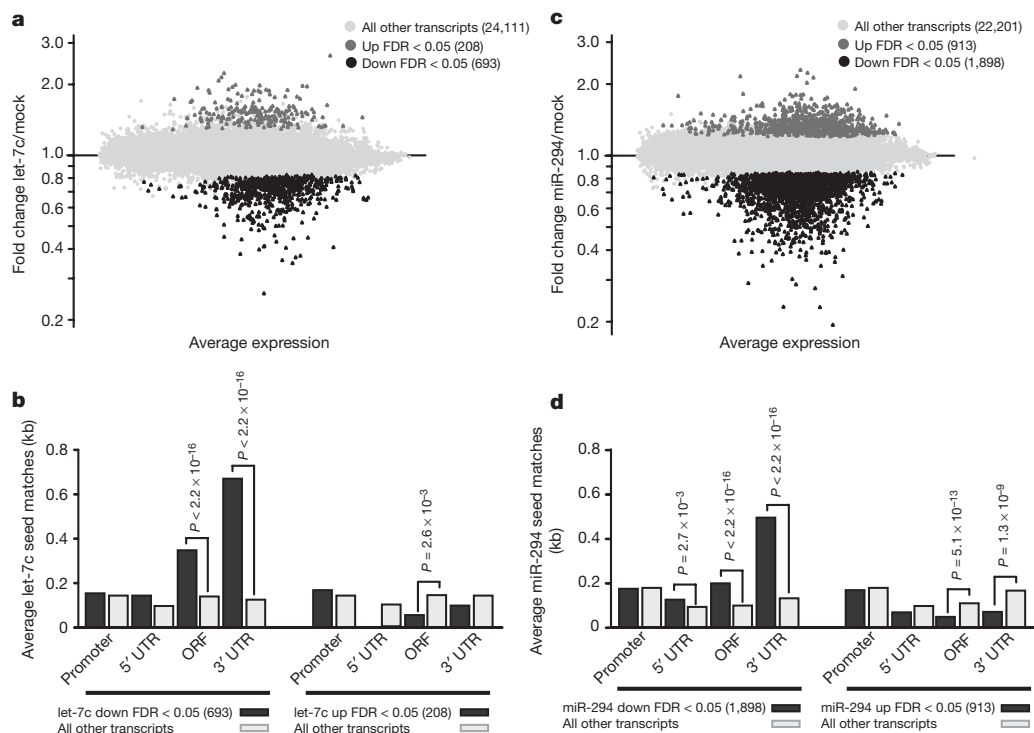
Figure 2 | The let-7 and ESCC miRNAs suppress hundreds of transcripts by binding their ORF and/or 3′UTR. a, Microarray analysis after introduction of let-7c alone. Upregulated transcripts are shown in dark grey, downregulated transcripts in black (FDR < 0.05). **b**, Analysis of seed matches in the promoter, 5′UTR, ORF and 3′UTR of let-7c-downregulated and -upregulated transcripts. Presented are the mean number of seeds matches per kilobase (kb) of sequence for the listed groups of altered genes described in **a**. *P*-values calculated by the Wilcoxon's rank sum test and Bonferroni-corrected are shown for *P* < 0.01. **c**, Microarray analysis after introduction of miR-294 alone. Colour labelling is as in **a**. **d**, Seed analysis as in **b** for miR-294 up- and downregulated transcripts.

Chen *et al.*[19] (a group of targets bound by Oct4, Sox2, Nanog, Smad1 and STAT3). The latter results agree with recent reports showing that genes bound by multiple pluripotency transcription factors are more likely to be transcriptionally activated[19,21]. There was no enrichment in the overlap between the miR-294-regulated transcripts and the Oct4-, Sox2-, Nanog- and Tcf3-bound genes. These data indicate that let-7c inhibits downstream targets of these pluripotency factors whereas miR-294 has no obvious effects on either the transcription factors themselves or on their downstream targets.

Opposing regulation of Myc, Lin28 and Sall4

Having discovered that Myc activity was alternatively downregulated and upregulated by let-7c and miR-294, we sought to identify other factors that might be similarly regulated by these miRNAs. Indeed, Gene Ontology analysis showed enrichment for ESC-enriched genes among the let-7c-downregulated and miR-294-upregulated transcript sets (Supplementary Fig. 7 and Supplementary Table 2). Eighty-eight transcripts were regulated in opposing directions by let-7c and miR-294, of which 44 contained a let-7c seed match (Supplementary Fig. 8 and Supplementary Table 3). Notably, this set of transcripts included the well-known pluripotency factors *Lin28* and *Sall4*. *Lin28* encodes an RNA-binding protein that inhibits let-7 processing[10–13,22], but not transfected let-7 mimic (Supplementary Fig. 9). *Sall4* encodes a transcription factor that promotes ESC self-renewal[23–25]. These findings show that the let-7 and ESCC families antagonistically regulate several genes with described roles in ESC self-renewal.

To verify our genomic analysis, we performed quantitative PCR following reverse transcription (qRT–PCR), western blotting analysis, and reporter assays for a subset of the genes. qRT–PCR confirmed the opposing effects of let-7c and miR-294 on *Lin28*, *Sall4*, *N-myc* and *c-myc* mRNA levels with a combination of the two miRNAs showing intermediate levels (Fig. 4a). Western blotting analysis showed similar results (Fig. 4b and Supplementary Fig. 10). Of note, c-Myc protein was markedly reduced in *Dgcr8*[−/−] versus wild-type ESCs, and was

brought back to wild-type levels by the introduction of miR-294. miR-294 had little effect on N-Myc levels. In contrast, let-7c had little effect on c-Myc, yet notably reduced N-Myc levels. Therefore, the cumulative effect of the miRNAs on total Myc (c-Myc plus N-Myc) protein levels followed a strong pattern of opposing regulation. Similarly, the miRNAs showed significant opposing effects on Lin28 and Sall4 protein levels. *Lin28* and *c-myc* are known targets of let-7 (refs 10, 20), and luciferase assays confirmed that *N-myc* and *Sall4* are also direct targets (Fig. 4c).

Considering that c-Myc was markedly reduced in *Dgcr8*[−/−] cells and then increased with miR-294, we considered the possibility that the loss of c-Myc alone could largely explain the sensitivity of *Dgcr8*[−/−] cells to let-7-induced silencing of ESC self-renewal. To test this possibility, we generated and evaluated *c-myc*[−/−] ESCs (Supplementary Fig. 11). The loss of c-Myc led to decreased expression of *Oct4* relative to the parental cell line (Supplementary Fig. 12a). Introduction of let-7c into the *c-myc*[−/−] cells decreased the expression levels of *Sox2* and *Nanog* (Supplementary Fig. 12b, c). However, levels were not reduced to the same degree as seen with the introduction of let-7c into *Dgcr8*[−/−] cells. These results indicate that the decrease of c-Myc in *Dgcr8*[−/−] cells alone cannot explain the sensitivity of these cells to let-7-induced silencing of ESC self-renewal.

Inhibition of let-7 promotes de-differentiation

Having identified a pro-differentiation function of the let-7 family of miRNAs, we hypothesized that inhibition of this miRNA family would enhance reprogramming of somatic cells to iPS cells. Indeed, Lin28, among other activities[26–29], inhibits let-7 biogenesis[10–13,22] and promotes de-differentiation of human somatic cells to iPS cells[30]. Reprogramming to iPS cells is typically achieved by the introduction of virally expressed Oct4, Sox2 and Klf4 with or without Myc into somatic cells such as mouse embryonic fibroblasts (MEFs). Although Myc markedly increases the efficiency of reprogramming, it is not essential[31,32]. To test the effect of the let-7 family on reprogramming, we used a let-7
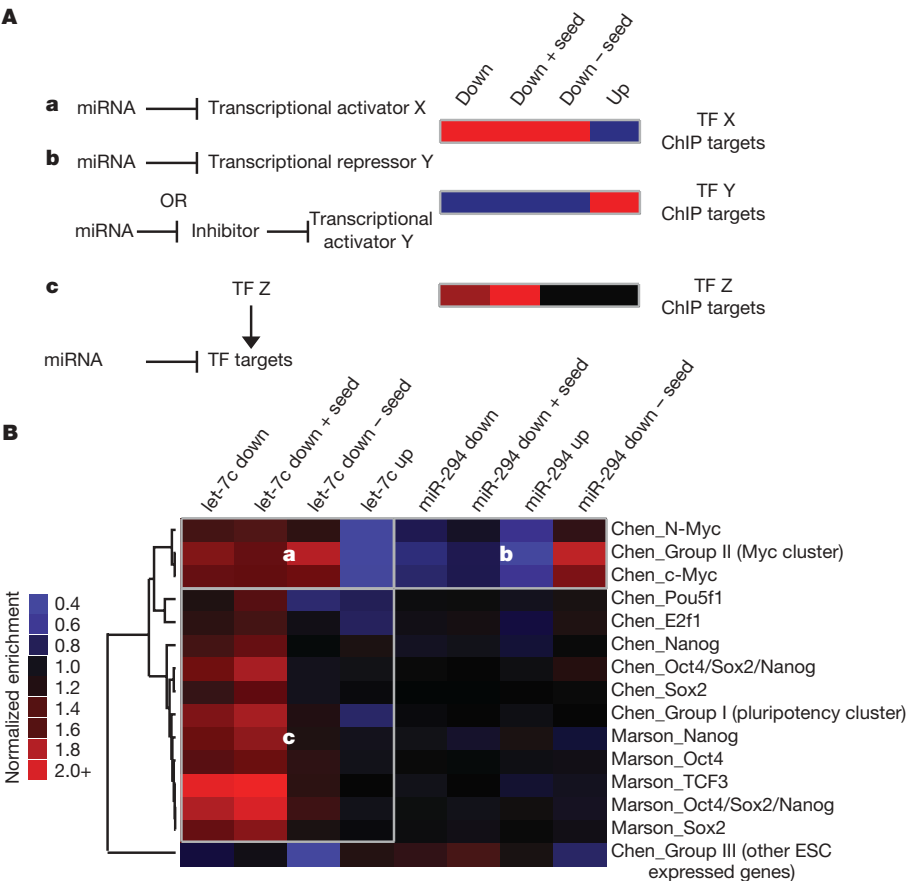
Figure 3 | Enrichment/depletion of transcription-factor-bound genes among miRNA-regulated transcripts. A, A schematic of hypothetical miRNA regulation of a transcription factor (TF) or its targets. Corresponding expected enrichment/depletion of the transcription factor ChIP targets in miRNA-induced up- or downregulated transcript sets are shown in a heat map. A key of colour coding representing relative enrichment is given in **B**. **B**, A heat map showing enrichment of the ChIP targets among the different sets of miRNA-regulated transcripts on the horizontal axis. The vertical axis represents the different ChIP data sets with first author and factor that was immunoprecipitated.
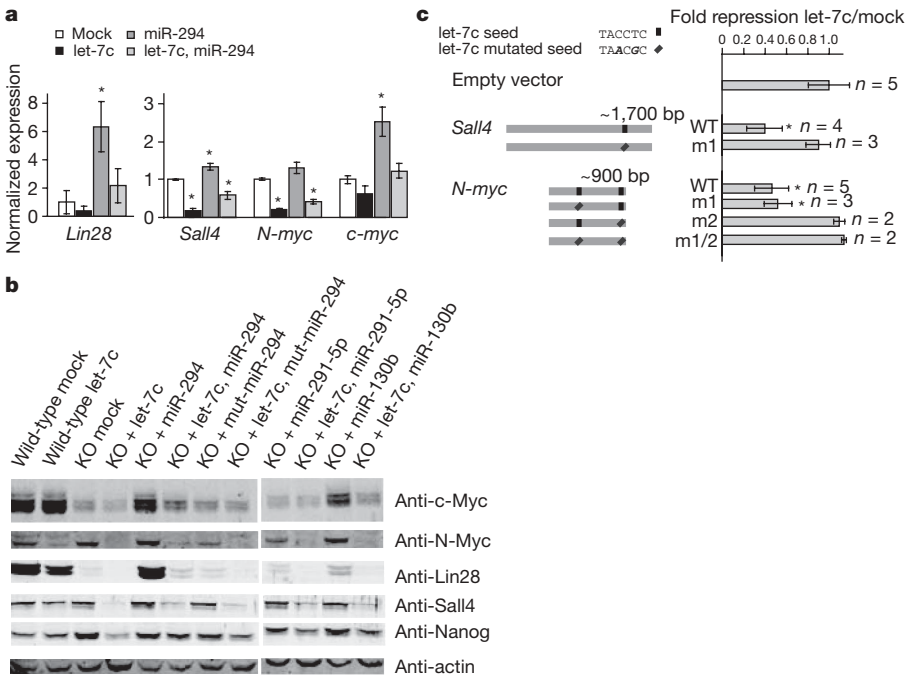
Figure 4 | Let-7c and miR-294 regulate Lin28, Sall4, c-Myc and N-Myc. a, qRT–PCR for *Lin28*, *Sall4*, *N-myc* and *c-myc* 12 h after transfection with let-7c, miR-294 or a combination of the two. $n = 3$. **b**, Representative western blot analysis 48 h after transfection with miRNAs. KO, knockout. Quantification is shown in Supplementary Fig. 10. $n = 3$. **c**, Luciferase analysis of *Sall4* and *N-myc* 3′UTRs. Seed matches for let-7c in the 3′UTRs along with different mutant constructs are represented diagrammatically in the left panel. Luciferase results after co-transfection with let-7c mimic relative to mock transfected are shown in the right panel. M1 and M2, mutant clones 1 and 2, respectively; WT, wild type. All data are represented as mean $\pm$ s.d. *$P < 0.05$ by Bonferroni-corrected t-test.

antisense inhibitor. This inhibitor was able to suppress several let-7 family members simultaneously (Supplementary Fig. 13).

MEFs express high levels of mature let-7 (ref. 15) and, therefore, these cells should be responsive to any pro-reprogramming effects of let-7 downregulation. We used *Oct4-GFP* (green fluorescent protein) transgenic MEFs to quantify changes in reprogramming efficiencies, as *Oct4-GFP* is activated late in the reprogramming process[33,34]. MEFs were transduced with retroviral vectors expressing Oct4, Sox2, Klf4, with or without c-Myc on day 0, as well as being transfected with let-7 or a control inhibitor on days 0 and 6. When three transcription factors were used (minus c-Myc), let-7 inhibition increased the number of GFP-positive colonies on day 10 by 4.3-fold compared to mock whereas a control inhibitor had no significant effect (Fig. 5a, left). In the presence of all four transcription factors, let-7 inhibition resulted in a 1.75-fold increase (Fig. 5a, right). Immunofluorescence confirmed expression of Nanog in reprogrammed cells (Supplementary Fig. 14). Furthermore, the resulting iPS cells expressed endogenous pluripotency markers at levels similar to wild-type ESCs and did not express the exogenously introduced factors (Supplementary Figs 15 and 16), as expected for fully reprogrammed cells[35]. The effect of the let-7 inhibitor is not due to enhanced proliferation of the MEFs as there was actually a subtle decrease in proliferation after transfection of either the let-7 or control inhibitor (Supplementary Fig. 17). These findings show that inhibition

of the let-7 family of miRNAs enhances the reprogramming of somatic cells. The finding that the enhancement was greater in the absence of Myc is consistent with Myc activity being one, but not the only important downstream target of let-7 in stabilizing the somatic cell fate.

Discussion

Our findings show that the let-7 and ESCC miRNA families have opposing effects on ESC self-renewal. We propose that they act in self-reinforcing loops to maintain the ESC self-renewing versus differentiated cell fates (Fig. 5b). In the self-renewing state, ESCC miRNAs indirectly increase expression of Lin28 and c-Myc. Lin28 functions to block the maturation of let-7 (refs 10–13). Therefore, the ESCC miRNAs prevent co-expression of let-7 miRNAs. Furthermore, ESCC-induced upregulation of c-Myc forms a positive feedback loop in which c-Myc and N-Myc, along with Oct4, Sox2 and Nanog, bind and activate expression of the ESCC miRNAs in the miR-290 miRNA cluster[7,15]. As ESCs differentiate, Oct4, Sox2 and Nanog are downregulated, resulting in the loss of ESCC and Lin28 expression. With the loss of Lin28, mature let-7 rapidly increases. This increase in let-7 is enhanced by a positive-feedback loop in which let-7 suppresses its own negative regulator Lin28. In the differentiated state, downregulation of Myc activity by let-7 prevents co-expression of the ESCC miRNAs. Furthermore, let-7 inhibits downstream targets of Oct4, Sox2, Nanog and Tcf3 to stabilize the differentiated state. Sall4, like Myc and Lin28, is positively regulated by the ESCC family and negatively regulated by the let-7 family. Decreases in Myc, Sall4 and Lin28 all promote ESC differentiation[23,25,26,36].

In the model we propose, the function of let-7 in repressing the self-renewing state is restricted to cells that do not express high levels of ESCC miRNAs. In fact, our model suggests that let-7 and ESCC miRNAs are never co-expressed at high levels. For this reason, we propose that the let-7 family does not function to initiate differentiation, but rather the antagonism between the let-7 and ESCC families stabilizes the switch between self-renewal and differentiation. Consistent with this model, the introduction of either ESCC miRNAs[7] or let-7 inhibitors into somatic cells promotes their de-differentiation into iPS cells. Furthermore, the ESCC and let-7 miRNAs make up a preponderance of the miRNAs in self-renewing ESCs and somatic cells, respectively[15], supporting a major role in influencing these alternative cell fates.

Other miRNAs have been reported to target the ESC transcriptional network[37–39]. Unlike the let-7 family, these other miRNAs have a more limited tissue distribution[8,9], suggesting that they may suppress self-renewal during differentiation along specific developmental pathways. Alternatively, these miRNAs may be involved in the early and transient stages of ESC differentiation while the let-7 miRNAs are involved in stabilizing the resulting differentiated cell fate. miRNAs related to the ESCC family (miR-17, miR-20, miR-93 and miR-106) and let-7 miRNAs have analogous roles in cancer, with the ESCC-related miRNAs promoting and the let-7 miRNAs inhibiting cancer growth[40,41]. It will be interesting to determine whether these miRNAs act through similar opposing pathways in cancer as in ESCs.

Figure 5 | Inhibition of let-7 miRNAs promotes reprogramming to induced pluripotency. a, Fold increase in Oct4–GFP-positive colonies in reprogramming with transduction of three (Oct4, Sox2 and Klf4) or four (plus c-Myc) transcription factors (3TF or 4TF, respectively) after mock, let-7 inhibitor, or control inhibitor transfection. *P*-values are indicated for *P* < 0.01 calculated by Bonferroni-corrected *t*-test. *n* = 10 for mock and let-7 inhibitor samples, *n* = 6 for control inhibitor samples. Data are presented as mean ± s.d. **b**, A model of the antagonism between the miR-294 and let-7c in the stabilization of the self-renewing and differentiated states. Bold genes and arrows are active in the indicated state. Mechanisms of ESCC upregulation of Lin28 and c-Myc are unknown and represented by a question mark.

METHODS SUMMARY

Dgcr8[−/−] and wild-type V6.5 ESCs were cultured as previously described[3]. miRNA mimics and inhibitors were obtained from ThermoFisher. mRNA profiling was performed on Affymetrix Mouse Gene 1.0 ST arrays. Bioinformatic analysis was performed using significance analysis of microarrays (SAM), R packages, and custom Python scripts. Reprogramming with *Oct4-GFP* MEFs was performed as previously described[7].

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

1. Babiarz, J. E. & Blelloch, R. Small RNAs — their biogenesis, regulation and function in embryonic stem cells. *StemBook* doi/10.3824/stembook.1.47.1. (2009).
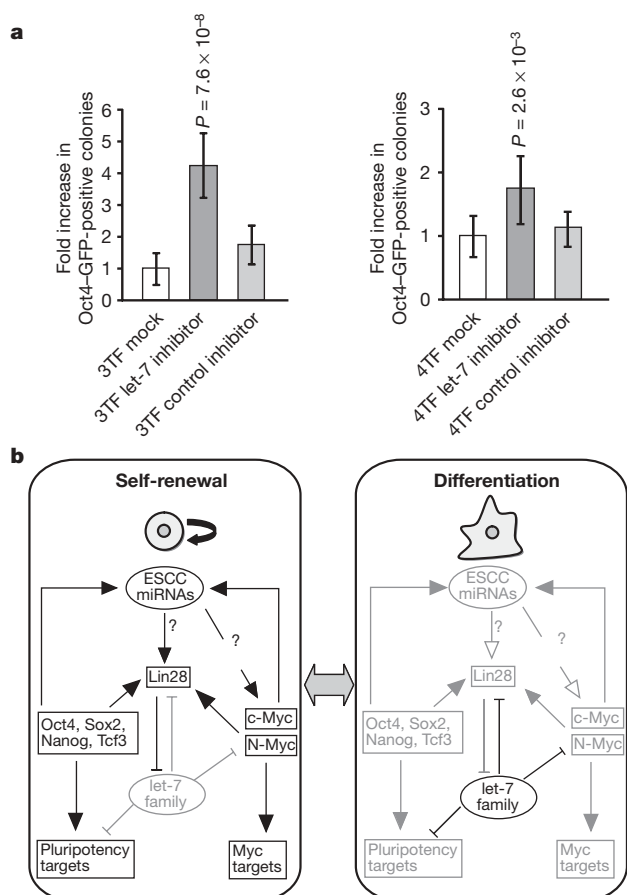
2. Hornstein, E. & Shomron, N. Canalization of development by microRNAs. *Nature Genet.* **38**, S20–S24 (2006).

3. Wang, Y., Medvid, R., Melton, C., Jaenisch, R. & Blelloch, R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nature Genet.* **39**, 380–385 (2007).

4. Kanellopoulou, C. *et al.* Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19**, 489–501 (2005).

5. 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).

6. Wang, Y. *et al.* Embryonic stem cell-specific microRNAs regulate the G1-S transition and promote rapid proliferation. *Nature Genet.* **40**, 1478–1483 (2008).

7. Judson, R., Babiarz, J. E., Venere, M. & Blelloch, R. Embryonic stem cell specific microRNAs promote induced pluripotency. *Nature Biotech.* **27**, 459–461 (2009).

8. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).

9. Chen, C. *et al.* Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets. *Mamm. Genome* **18**, 316–327 (2007).

10. Rybak, A. *et al.* A feedback loop comprising *lin-28* and *let-7* controls pre-*let-7* maturation during neural stem-cell commitment. *Nature Cell Biol.* **10**, 987–993 (2008).

11. Viswanathan, S. R., Daley, G. Q. & Gregory, R. I. Selective blockade of microRNA processing by Lin28. *Science* **320**, 97–100 (2008).

12. Heo, I. *et al.* Lin28 mediates the terminal uridylation of let-7 precursor microRNA. *Mol. Cell* **32**, 276–284 (2008).

13. Newman, M. A., Thomson, J. M. & Hammond, S. M. Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. *RNA* **14**, 1539–1549 (2008).

14. Thomson, J. M. *et al.* Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev.* **20**, 2202–2207 (2006).

15. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).

16. Calabrese, J. M., Seila, A. C., Yeo, G. W. & Sharp, P. A. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **104**, 18097–18102 (2007).

17. Benetti, R. *et al.* A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. *Nature Struct. Mol. Biol.* **15**, 268–279 (2008).

18. Sinkkonen, L. *et al.* MicroRNAs control *de novo* DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nature Struct. Mol. Biol.* **15**, 259–267 (2008).

19. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).

20. Kumar, M. S., Lu, J., Mercer, K. L., Golub, T. R. & Jacks, T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nature Genet.* **39**, 673–677 (2007).

21. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–1061 (2008).

22. Piskounova, E. *et al.* Determinants of microRNA processing inhibition by the developmentally regulated RNA-binding protein Lin28. *J. Biol. Chem.* **283**, 21310–21314 (2008).

23. Lim, C. Y. *et al.* Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages. *Cell Stem Cell* **3**, 543–554 (2008).

24. Wu, Q. *et al.* Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. *J. Biol. Chem.* **281**, 24090–24094 (2006).

25. Zhang, J. *et al.* Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Oct4. *Nature Cell Biol.* **8**, 1114–1123 (2006).

26. Heo, I. *et al.* TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell* **138**, 696–708 (2009).

27. Xu, B., Zhang, K. & Huang, Y. Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells. *RNA* **15**, 357–361 (2009).

28. Jones, M. R. *et al.* Zcchc11-dependent uridylation of microRNA directs cytokine expression. *Nature Cell Biol.* **11**, 1157–1163 (2009).

29. Polesskaya, A. *et al.* Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. *Genes Dev.* **21**, 1125–1138 (2007).

30. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).

31. Nakagawa, M. *et al.* Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnol.* **26**, 101–106 (2008).

32. Wernig, M., Meissner, A., Cassady, J. P. & Jaenisch, R. c-Myc is dispensable for direct reprogramming of mouse fibroblasts. *Cell Stem Cell* **2**, 10–12 (2008).

33. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2**, 230–240 (2008).

34. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2**, 151–159 (2008).

35. Hochedlinger, K. & Plath, K. Epigenetic reprogramming and induced pluripotency. *Development* **136**, 509–523 (2009).

36. Cartwright, P. *et al.* LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* **132**, 885–896 (2005).

37. Tay, Y. M. *et al.* MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1. *Stem Cells* **26**, 17–29 (2008).

38. Tay, Y., Zhang, J., Thomson, A. M., Lim, B. & Rigoutsos, I. MicroRNAs to *Nanog*, *Oct4* and *Sox2* coding regions modulate embryonic stem cell differentiation. *Nature* **455**, 1124–1128 (2008).

39. Xu, N., Papagiannakopoulos, T., Pan, G., Thomson, J. A. & Kosik, K. S. MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* **137**, 647–658 (2009).

40. Mendell, J. T. miRiad roles for the miR-17-92 cluster in development and Disease. *Cell* **133**, 217–222 (2008).

41. Büssing, I., Slack, F. J. & Grobhans, H. *let-7* microRNAs in development, stem cells and cancer. *Trends Mol. Med.* **14**, 400–409 (2008).

Author Contributions C.M. contributed to Figs 1–4, 5b and Supplementary Figs 1–13 and 18. R.L.J. contributed to Fig. 5a and Supplementary Figs 13–17. C.M., R.L.J. and R.B. conceived the experiments, analysed the data, and wrote the manuscript.

Author Information Our microarray profiling data can be found at the Gene Expression Omnibus (GEO) database, under accession GSE18840. 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 R.B. (blellochr@stemcell.ucsf.edu).

METHODS

Tissue culture, transfection and alkaline phosphatase staining. ESC lines and culture conditions were previously described[3]. ESCs were weaned off MEFs and maintained in MEF conditioned media. For ESC differentiation assays, 40,000 *Dgcr8*[−/−] or 12,000 wild-type ESCs were plated in gelatinized 12-well plates (or half the number of cells were plated on 24-well plates) on day 0 in LIF media. On day 1, miRIDIAN miRNA mimics (Dharmacon, ThermoFisher) were transfected at a concentration of 50 nM using Dharmafect1 (Dharmacon, ThermoFisher) following the manufacturer's protocol. Media was changed daily. On the third day after transfection, cells were either lysed in Trizol (Invitrogen) for qRT–PCR analysis or fixed in 4% paraformaldehyde (PFA) for alkaline phosphatase staining. Alkaline phosphatase staining was performed per the manufacturer's instructions (Vector Labs). iPS cell lines were maintained in ESC media plus 15% knockout serum on irradiated MEF feeders. Colony reformation assays were performed as previously described[3]. In brief, cells were exposed to miRNA mimics for 3 days then trypsinized and counted. A defined number of cells were replated on MEFs to form colonies for 5–7 days. The efficiency of colony reformation was determined by dividing the number of alkaline-phosphatase-positive colonies by the number of cells plated. Neural progenitor cells used in Supplementary Fig. 3 were generated by *in vitro* differentiation of ESCs as described previously[42].

Animal use. All animal experiments described in this article were approved by the Institutional Animal Care and Use Committee of the University of California San Francisco.

ESC derivation. Timed matings were set up for *c-myc* f/f mice[43]. ESCs were derived from embryos isolated at embryonic day (E)3.5. These embryos were cultured on an irradiated MEF feeder layer in ESC media supplemented with 50 μM PD98059 (ref. 44) and disassociated onto fresh feeders. ESCs were PCR-genotyped as previously described[43]. A flox/flox line was grown out, infected with Ad5 Cre-IRES-GFP virus, sorted by FACS, and plated back onto MEF feeders. *c-myc*[−/−] colonies were grown out and verified by PCR genotyping and western blotting.

mRNA arrays. qRT–PCR showed that mRNA levels of a known let-7 target, *Lin28* (ref. 10), was maximally reduced 12 h after transfection before a large decline in *Oct4* and *Nanog* (Supplementary Fig. 18). Therefore, we chose 12 h for all microarray analysis to minimize secondary effects of let-7c-induced differentiation. On day 0, 150,000 cells were plated in a 3.5-cm dish. miRIDIAN miRNA mimics (Dharmacon, ThermoFisher) were transfected at a concentration of 50 nM in media in the absence of LIF. At 12 h after transfection cells were lysed in Trizol (Invitrogen) and RNA was prepared according to the manufacturer's protocol. Affymetrix Mouse Gene 1.0 ST arrays were probed by the Gladstone Genomics Core (http://www.gladstone.ucsf.edu/gladstone/site/genomicscore/). Three biological samples were assayed for each treatment. Data were analysed by Affymetrix Expression Console software. The robust multichip analysis (RMA) algorithm was used to normalize the array signal across chips. SAM (http://www-stat.stanford.edu/~tibs/SAM/) was used to determine FDR cutoffs for significantly altered genes.

qRT–PCR analysis. RNA for all qRT–PCR analyses was prepared using Trizol (Invitrogen) and quantified on a Nanodrop Spectrophotometer (ThermoFisher). Five-hundred nanograms of RNA was DNase-treated using DNaseI amplification grade (Invitrogen). For qRT–PCR of mRNAs, DNase-treated samples were reverse-transcribed using the Superscript III first-strand synthesis system for RT–PCR (Invitrogen). qPCR reactions on resulting cDNAs were performed on either an ABI Prism 7100 or an ABI 7900HT (Applied Biosystems). Primers are listed in Supplementary Table 4. For miRNAs, qRT–PCR was performed either by using TaqMan miRNA assays (Applied Biosystems) or by polyadenylating the miRNAs and then using a modified oligodT reverse transcription primer as described previously[45].

Lin28 and GFP expression in 293T cells. *Lin28* was cloned into an expression vector under the EF1α promoter and upstream of IRES Pac (puromycin resistance). A similarly constructed GFP expression construct was previously generated[6]. 293T cells were transfected with 5 μg of each construct and selected with 0.6 μg ml⁻¹ puromycin for 12 days.

Luciferase reporter assays. Constructs were produced as follows. The *N-myc* and *Sall4* 3′UTRs were amplified from ESC cDNA and cloned into the NotI and XhoI sites in psiCheck-2 vector (Promega). Mutant UTRs were generated by a two-step PCR strategy with overlapping mutated PCR primers. Products of two PCRs with mutations were used in a second PCR reaction to generate full-length mutated inserts that were cut and ligated into a cut empty vector. For transfections, 8,000 *Dgcr8*[−/−] ESCs were plated in ESC media in a 96-well plate pretreated with 0.2% gelatin. The next day, miRIDIAN miRNA mimics (Dharmacon, ThermoFisher) were transfected with Dharmafect1 (Dharmacon, ThermoFisher) following the manufacturer's protocol at a concentration of 100 nM. Simultaneously, luciferase

constructs were transfected into ESCs at a concentration of 200 ng per well using FUGENE 6 (Roche) transfection reagent following the manufacturers protocol. The next day, 14–18 h later, cells were lysed and luciferase assays were performed using a Dual-Luciferase Reporter Assay System (Promega) on a single automatic injection Mithras (Berthold technologies) luminometer following the manufacturer's protocol. Transfection of each construct was performed in triplicate in each assay. Ratios of *Renilla* luciferase readings to firefly luciferase readings were averaged for each experiment. Replicates performed on separate days were mean-centred with the common readings from the individual runs.

Seed match analysis. Promoter (1,000 base pairs from the transcriptional start), 5′UTR, ORF and 3′UTRs for Ensembl Transcripts (mm9) and known genes (mm8) were downloaded separately from the UCSC Genome Browser Table Browser. Seed match analysis was performed on these transcripts using a custom Python script. 7-nucleotide seeds were defined as either 7mer-1A or 7mer-m8 (ref. 46). Seed match results were mapped to Affymetrix IDs. A Python script was then implemented to eliminate redundant transcripts as transcripts often mapped >1:1 with Affymetrix IDs. The transcript with the most 7-nucleotide seed matches was chosen to produce a 1:1 transcript to Affymetrix ID mapping. This mapping was done separately for the promoters, 5′UTRs, ORFs and 3′UTRs. In rare cases, duplicate Affymetrix IDs exist for the same gene. These were retained in our analyses. Microsoft Access (Microsoft) was used to generate list overlaps for analyses. *P*-values were calculated in Fig. 2b, d with the number of seed matches per kb of transcript using Wilcoxon's rank sum test in R. *P*-values were calculated in Supplementary Fig. 6a, b using a binary 0 for no seed matches or 1 for a seed match using the hypergeometric distribution function in R.

ChIP target overlap analysis. ChIP targets were downloaded from the supplementary tables[15,19]. Scripts were written to convert provided transcript IDs to a non-redundant list of Affymetrix IDs. Microsoft Access (Microsoft) and custom Python scripts were used to perform comparisons between gene lists and ChIP gene target lists. ChIP data described previously[19] was downloaded as an association score between any particular gene and the transcription factor of interest. These scores were used directly for enrichment. For the Oct4-, Sox2- Nanog-bound group from Chen *et al.*[19] any score above 0 was counted as bound. For all data, enrichment for ChIP gene target sets in miRNA-regulated gene sets was performed relative to all genes analysed to produce the miRNA-regulated gene sets (that is, all genes with Affymetrix IDs mapping to coding transcripts). The enrichments for any given ChIP target set were median normalized with all the miRNA-regulated genes sets in Fig. 3B. We performed this normalization because both the ChIP targets of the transcription factors and the miRNA-regulated gene sets in our analysis are enriched for more highly expressed genes[19]. We get a similar pattern of results without this normalization, although all comparisons appear more highly enriched owing to the expression levels (data not shown). Un-normalized enrichment is defined as: (genes in overlap of miRNA-altered group and ChIP group/all genes in miRNA-altered group)/(all genes in ChIP group/all genes used in analysis to generate miRNA altered groups).

Our enrichment analysis could yield several possible outcomes depending on whether the miRNA targeted the transcription factor directly versus targeted transcripts downstream of the transcription factor. The following outcomes are presented in Fig. 3A. (1) If an miRNA directly targets a specific transcriptional activator, this activator will be downregulated, and thus its ChIP target genes will likewise tend to be downregulated. This will result in an enrichment of ChIP target genes within the miRNA's downregulated gene set independent of there being a seed match in these targets. Similarly, the ChIP target genes should be depleted in the miRNA's upregulated gene set (Fig. 3A, a). (2) If an miRNA directly targets a transcriptional repressor, there would be the inverse outcome; that is, the ChIP target genes should be enriched in the miRNA's upregulated gene set and depleted in the miRNA's downregulated gene set regardless of seed match (Fig. 3A, b). (3) If an miRNA targets an activating transcription factor's downstream targets, but not the transcription factor itself, ChIP target genes would be enriched in the downregulated gene set with a seed match but not without a seed match. Furthermore, there should not be enrichment in the upregulated transcripts (Fig. 3A, c).

Gene Ontology. Stem-cell-associated genes (genes upregulated in ESCs relative to brain and bone marrow) were generated from data described previously[47] and were downloaded as a list from MySigDB (http://www.broad.mit.edu/gsea/msigdb). Enrichment of these stem cell associated genes in miRNA altered gene sets was performed, and *P*-values were calculated by Fischer's exact test.

Immunohistochemistry. Cells were fixed with 4% PFA and washed twice in PBS with 0.1% Triton X-100 (PBT). PBT with 2% bovine serum albumin (BSA) and 1% goat-serum was used to block for 1 h before the addition of primary antibody against Oct4 (Santa Cruz, rabbit polyclonal, product sc-9081) or Nanog (Calbiochem, rabbit polyclonal, product sc-1000), which was incubated overnight

at 4 °C or at room temperature for approximately 2 h. Cells were washed with PBT, blocked with PBT plus 2% BSA and 10% goat-serum for 1 h before addition of secondary antibodies (Alexa Fluor 488 goat anti-rabbit IgG, Invitrogen).

Western blots. On day 0, approximately 200,000 *Dgcr8*$^{-/-}$ or 50,000 wild-type ESCs were plated in a 6-well plate. The next day miRIDIAN miRNA mimics (Dharmacon, ThermoFisher) were transfected at a concentration of 50 nM. Lysates were collected 2 days after transfection in EBC buffer (50 mM Tris-HCl, pH 8.0, 120 mM NaCl, 0.5% Nonidet P-40, 1 mM EDTA) containing 1× protease inhibitor cocktail (Roche). Lysates were incubated at 4 °C for 45 min rocking then spun at 4 °C and approximately 20,000*g* in a table-top centrifuge. Protein was quantified using a Bio-Rad protein assay (Bio-Rad). Thirty micrograms of protein was resolved on an 8% SDS–PAGE gel. Proteins were transferred to Immobilon-FL (Millipore) and processed for immunodetection. Blots were scanned on a Licor Odyssey Scanner (Licor). The actin antibody was used at a 1:1,000 dilution (Sigma, mouse monoclonal clone AC-40, A4700), the c-Myc antibody at 1:500 (Epitomics, N-term rabbit monoclonal, 1472-1), the N-Myc antibody at 1:500 (Calbiochem, mouse monoclonal, OP13), the Nanog antibody at 1:1,000 (Abcam, rabbit polyclonal, ab21603), the Sall4 antibody at 1:500 (Abcam, rabbit polyclonal, ab29112), and the Lin28 antibody at 1:1,000 (Abcam, rabbit polyclonal, ab46020). Secondary infrared-dye antibodies from Licor were used at 1:10,000. Data were exported from the Licor Odyssey as jpg and quantified using ImageJ software (NIH).

MEF isolation. E13.5 embryos from *Oct4-GFP/Rosa-26-β*-galactosidase transgenic crosses were isolated by Caesarean section and washed in HBSS. Heads and visceral tissues were removed. Remaining tissue was washed in fresh HBSS, briefly rinsed with 70% ethanol, then submerged in 0.05 mM trypsin/1 mM EDTA HBSS solution and incubated at 37 °C for 10 min. Tissue was pipetted repeatedly to aid in tissue dissociation, then added to MEF media containing 10% FBS and plated (passage 0).

Retrovirus infection. The retroviral packaging vector pCL-ECO was transfected into 293T cells simultaneously with pMXs vectors containing *Oct4*, *Sox2*, *Klf4* or *c-myc* cDNA (Addgene) using Fugene 6 (Roche)[48]. At 24 h, the media was changed, and at 48 h, the media was collected, filtered (0.45 μM), and frozen in aliquots at −80 °C. Retrovirus was never thawed more than once. To induce reprogramming, passage 3 *Oct4-GFP*, *Rosa26-Glb1/neo* MEFs[49] were plated on gelatin-coated 12-well plates at 12,000 cells per well. Retrovirus-containing media was added 24 h later (day 0). Cells were transfected with 16 nM microRNA inhibitors (Dharmacon, ThermoFisher, I-310106-04 for let-7 inhibitor, IN-001000-01-05 for control inhibitor). Cells in reprogramming assays were transfected on days 0 and 6 after retroviral infection. Media was changed daily. Media was replaced with ESC media + 15% FBS + LIF on day 2, and ESC media + 15% knockout serum replacement (Invitrogen) + LIF on day 6. GFP$^+$ colonies were counted on day 10. Individual iPS cell colonies were picked and expanded for analysis between days 10 and 15.

42. Okabe, S., Forsberg-Nilsson, K., Spiro, A. C., Segal, M. & McKay, R. D. Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. *Mech. Dev.* **59**, 89–102 (1996).
43. Trumpp, A. *et al.* c-Myc regulates mammalian body size by controlling cell number but not cell size. *Nature* **414**, 768–773 (2001).
44. Buehr, M. & Smith, A. Genesis of embryonic stem cells. *Phil. Trans. R. Soc. Lond. B* **358**, 1397–1402 (2003).
45. Shi, R. & Chiang, V. L. Facile means for quantifying microRNA expression by real-time PCR. *Biotechniques* **39**, 519–525 (2005).
46. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
47. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C. & Melton, D. A. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600 (2002).
48. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
49. Blelloch, R., Venere, M., Yen, J. & Ramalho-Santos, M. Generation of induced pluripotent stem cells in the absence of drug selection. *Cell Stem Cell* **1**, 245–247 (2007).

ARTICLES

An aspartyl protease directs malaria effector proteins to the host cell

Justin A. Boddey[1], Anthony N. Hodder[1], Svenja Günther[1], Paul R. Gilson[2], Heather Patsiouras[3], Eugene A. Kapp[3], J. Andrew Pearce[1], Tania F. de Koning-Ward[4], Richard J. Simpson[1,3], Brendan S. Crabb[2] & Alan F. Cowman[1]

Plasmodium falciparum* causes the virulent form of malaria and disease manifestations are linked to growth inside infected erythrocytes. To survive and evade host responses the parasite remodels the erythrocyte by exporting several hundred effector proteins beyond the surrounding parasitophorous vacuole membrane. A feature of exported proteins is a pentameric motif (RxLxE/Q/D) that is a substrate for an unknown protease. Here we show that the protein responsible for cleavage of this motif is plasmepsin V (PMV), an aspartic acid protease located in the endoplasmic reticulum. PMV cleavage reveals the export signal (xE/Q/D) at the amino terminus of cargo proteins. Expression of an identical mature protein with xQ at the N terminus generated by signal peptidase was not exported, demonstrating that PMV activity is essential and linked with other key export events. Identification of the protease responsible for export into erythrocytes provides a novel target for therapeutic intervention against this devastating disease.

Malaria is a disease of global importance causing significant morbidity and death[1]. *Plasmodium falciparum* is responsible for the lethal form of malaria and morphological changes occur in infected erythrocytes altering physical properties and impairing circulation[2]. Parasitized erythrocytes become rigid and adhere to different cell types and this has a pivotal role in complications of *P. falciparum* malaria[3,4]. As the parasite develops inside the erythrocyte over 300 effector proteins are exported to the host[5,6], some of which associate with the cytoskeleton[7] or are inserted into the host membrane[8–10] and these alter the mechanical and adhesive properties of infected erythrocytes[11]. Trafficking of parasite proteins to erythrocytes is a multi-step process involving entry into the endoplasmic reticulum (ER) and trafficking to the cell surface followed by translocation across the parasitophorous vacuole membrane (PVM) into the erythrocyte or assembly into membranous structures (Maurer's clefts) to reach their destinations[12] (Supplementary Fig. 1). The *P. falciparum* translocon of exported proteins (PTEX, *Plasmodium* translocon of exported proteins) complex is associated with the PVM and appears to be responsible for export across this barrier[13].

For export beyond the PVM of infected erythrocytes, or hepatocytes by liver stage parasites, *P. falciparum* employs a signal sequence for ER entry followed by a motif termed PEXEL (*Plasmodium* export element) or vacuolar translocation signal, which is conserved across *Plasmodium* spp.[5,6]. The motif allowed identification of exported proteins in *Plasmodium* spp., revealing that 5–8% of parasite proteins are destined for export[14,15]. The PEXEL consists of the pentameric sequence RxLxE/Q/D that is processed after the conserved leucine, and N-acetylated in the ER of the parasite[16,17]. The parasite protease responsible for cleavage of the PEXEL remains unknown, although signal peptidase has been suggested as a candidate[16]. Because successful export to the host relies on the PEXEL motif[5,6,17,18], the protease responsible for processing would be a crucial part of the export machinery and its identification provides an important target for development of novel antimalarials.

PMV is an ER protease that cleaves PEXEL

To identify candidate proteases responsible for PEXEL processing we used features of the cleavage site to predict the type of enzyme involved. We noted that human immunodeficiency virus (HIV)-1 aspartic acid protease cleaved substrate proteins within a motif similar to PEXEL[19,20], indicating that the *Plasmodium* enzyme shares similarity with it. To test this we incubated inhibitors of HIV-1 protease in culture with *P. falciparum*-infected erythrocytes and monitored PEXEL cleavage. PEXEL processing was reduced in the presence of lopinavir, nelfinavir, ritonavir and saquinavir for PfEMP3– and knob associated histidine-rich (KAHRP)–green fluorescent protein conjugated (GFP) proteins (Fig. 1a–f). Although the inhibition was inefficient it indicated that PEXEL processing might involve a protease similar to HIV-1 aspartic acid protease. Sensitivity of signal peptidase was assessed using the inhibitors with acyl carrier protein signal sequence (ACPs)–GFP[21]. This transgenic protein lacks a PEXEL and is secreted into the parasitophorous vacuole[21]. In contrast to the PEXEL-containing proteins, we observed no detectable reduction in processing of the signal sequence from the ACPs–GFP chimaera with inhibitors (Fig. 1g–i). We conclude that the PEXEL protease is not the parasite signal peptidase but involves a unique protease belonging to the aspartic acid family.

A list of candidate aspartic acid proteases was generated from the *P. falciparum* genome on the basis of first, expression in blood and liver stages of the parasite[22,23], second, the presence of a signal sequence for ER entry, and third, specificity to *Plasmodia* spp. because the PEXEL is absent from closely related apicomplexans[14]. PMV (PF13_0133), a putative aspartic acid protease localizing to the ER that is functionally divergent to other members of the plasmepsin family, satisfied our criteria[24]. To investigate the function of PMV, we introduced a haemagglutinin (HA) tag at the carboxy terminus (3D7-PMVHA) (Fig. 2a). HA-tagged PMV localized to the ER[25], as has been shown previously for the endogenous protein[24] (Fig. 2b).

To determine if PMV could cleave PEXEL substrates we immuno-precipitated the HA-tagged protease (ipPMVHA) and incubated KAHRP peptides containing either a wild-type PEXEL or mutations that interfere with cleavage (Fig. 2e)[17]. The resulting peptides were separated by reversed-phase high-pressure liquid chromatography (RP-HPLC) and analysed by nano liquid chromatography tandem mass spectrometry (LC-MS/MS) (Fig. 2e, h, i). Incubation of the

---

[1]The Walter and Eliza Hall Institute of Medical Research, Melbourne 3052, Australia. [2]Macfarlane Burnet Institute for Medical Research and Public Health, Melbourne 3004, Australia. [3]Joint Proteomics Facility, Ludwig Institute for Cancer Research, Melbourne 3050, Australia. [4]Deakin University, Waurn Ponds 3217, Australia.
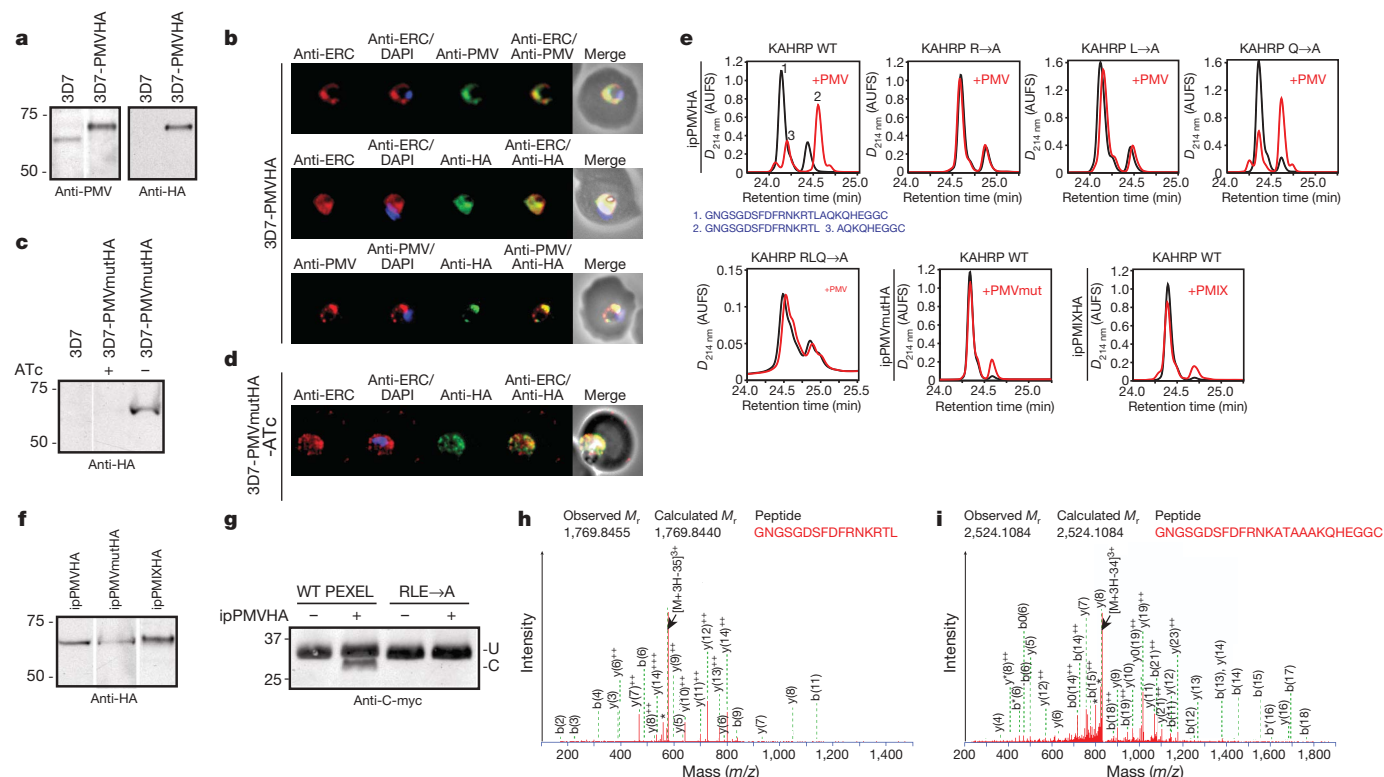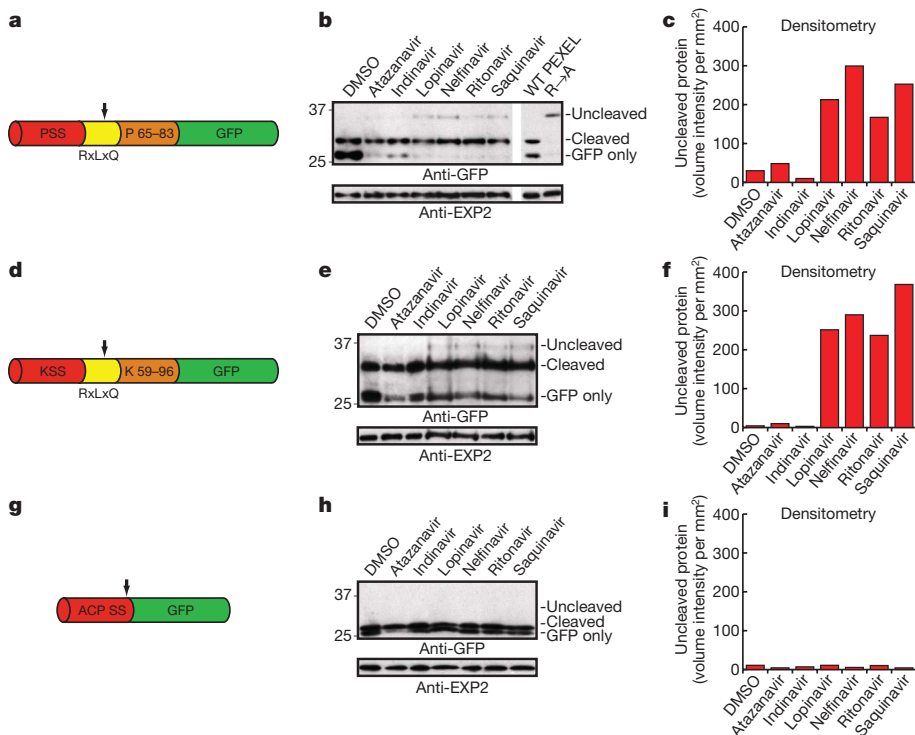
Figure 1 | PEXEL processing is sensitive to HIV protease inhibitors. a, Structure of PfEMP3–GFP. PSS, signal sequence; RxLxQ, PEXEL; P 65–83, residues 65–83 of PfEMP3. Arrow, PEXEL cleavage. **b**, PfEMP3–GFP processing with HIV-1 inhibitors. **c**, Densitometry of uncleaved bands in **b**. **d**, Structure of KAHRP–GFP. KSS, signal sequence; RxLxQ, PEXEL; K 59–96, residues 59–96 of KAHRP. Arrow, PEXEL cleavage. **e**, KAHRP–GFP PEXEL processing with inhibitors. **f**, Densitometry of uncleaved bands in **e**. **g**, Structure of ACPs–GFP. ACP SS, signal sequence. Arrow, signal sequence cleavage. **h**, ACPs–GFP signal sequence processing is not inhibited. Cleaved, signal sequence cleavage. **i**, Densitometry of the uncleaved region of immunoblot. (See Supplementary Information for details.) Molecular masses listed on the left of gels are in kDa, in all figures.

Figure 2 | PMV cleaves PEXEL. a, PMV HA-tagged (3D7-PMVHA). **b**, PMV localizes to endoplasmic reticulum. **c**, Inactive HA-tagged PMV (3D7-PMVmutHA). **d**, Mutant PMV (3D7-PMVmutHA) localizes to endoplasmic reticulum. **e**, HA-tagged PMV (ipPMVHA) with KAHRP peptides; unmutated, R→A, L→A, Q→A, RLQ→A. ipPMVmutHA (inactive) or ipPMIXHA (plasmepsin IX) with unmutated peptide. WT, wild type. The ultraviolet detector was operated at a wavelength of 214 nm and set at a sensitivity of two absorbance units full scale (AUFS). D, attenuance. **f**, Recovered enzyme. **g**, Recombinant GBP130 cleavage by PMV. Last lanes are GBP130 mutant (RLE→A). U, Uncleaved; C, cleaved; WT, wild type. **h**, Identification of cleaved peptide GNGSGDSFDFRNKRTL. **i**, Identification of uncleaved RLQ→A mutant peptide. *, Multiple H_2O and NH_3 losses. (See Supplementary Information for details.)

KAHRP peptide with ipPMVHA produced two major peptide species (Fig. 2e; first panel) derived from cleavage within the PEXEL after leucine (Fig. 2e, h). In contrast, mutation of the conserved amino acids R→A, L→A or RLQ→A resulted in no cleavage (Fig. 2e, i), consistent with the properties of the PEXEL cleaving protease *in vivo*[17]. Furthermore, mutation of the conserved fifth amino acid (Q→A) did not inhibit cleavage (Fig. 2e), corroborating our previous data that this residue is not required for PEXEL processing[17]. To verify that the PEXEL cleavage was due to PMV, and not other coprecipitated proteases, we constructed *P. falciparum* transgenic parasites that either conditionally or constitutively expressed an inactive form of PMV tagged with HA (PMVmutHA and PMVmutHA2, respectively). We confirmed expression and ER localization of PMVmutHA (Fig. 2c, d) and the immunoprecipitated proteins (ipPMVmutHA or ipPMVmutHA2) were incubated with KAHRP peptides and no processing occurred (Fig. 2e and Supplementary Fig. 2). Transgenic parasites were generated that expressed HA-tagged plasmepsin IX (PF14_0281) (3D7-PMIXHA), another member of the aspartic acid protease family in *P. falciparum*. Immunoprecipitated plasmepsin IX (ipPMIXHA) could not process the KAHRP peptide (Fig. 2e, f). ipPMVHA also cleaved a recombinant GBP130 protein to the expected size (Fig. 2g) as well as a quenched fluorescent substrate of nine amino acids (Fig. 3a), both of which contained a PEXEL. Mutation of conserved amino acids in the PEXEL blocked cleavage of both substrates (Figs 2g and 3a). The HIV protease inhibitor lopinavir reduced PMV activity by 19.1% ($P = 0.0145$; Fig. 3b), consistent with *in vivo* inhibition of PEXEL cleavage of PfEMP3– and KAHRP–GFP (Fig. 1). Taken together, these results show that PMV specifically cleaves the PEXEL and has the functional specificity expected of the PEXEL protease.

To provide further evidence that PMV was responsible for PEXEL cleavage we expressed amino acids 37 to 521 in *Escherichia coli* with an N-terminal hexa-His fusion tag. The recombinant PMV was extracted from inclusion bodies and purified on NiNTA under denaturing conditions and further fractionated using RP-HPLC (Fig. 3c). Individual fractions were analysed using SDS–PAGE (Fig. 3d) or transferred to nitrocellulose and probed with anti-PMV antibodies (Fig. 3e). A major band was observed at approximately 55 kDa (fraction number 3, 4 and 5) and smaller processed forms of PMV were also detected (fractions 2–3) indicating N-terminal processing (Fig. 3e, f) occurred either by the protease itself or other *E. coli* proteases present in the purified inclusion bodies. Processed forms of ipPMVHA are also detectable in immunoblots, suggesting that one may be the activated form (Supplementary Fig. 3). Recombinant protein fractions

was assayed for PEXEL cleavage using KAHRP peptides and PEXEL-specific activity was observed with fractions 2 and 3, which contained the processed form of PMV (Fig. 3c–g and Supplementary Fig. 4). This data suggests that a recombinant, N-terminally truncated form of PMV can cleave the PEXEL motif.

PMV directs export of effector proteins at the ER

The PEXEL motif is required for export to infected erythrocytes[5,6,17,18], so we next determined the importance of PMV for export *in vivo*. First, we investigated the importance of signal peptidase in export by mutating the PEXEL R→A in the normally exported protein, PfEMP3–GFP, to assess whether the signal sequence was processed. The mutant was retained in the ER as a full-length integral membrane protein by its signal peptide, which was not cleaved by signal peptidase (Fig. 4a), as we have shown previously for KAHRP[17], indicating that signal peptidase is not essential for trafficking. This highlights the important role PMV has in freeing cargo proteins from the ER membrane by processing the PEXEL (Fig. 4a). Additionally, we previously showed that the non-exported protein MSP3 is exported if the signal sequence is fused to a PEXEL[26], indicating that PMV is the dominant protease over signal peptidase in determining the final destination of proteins. Second, we determined if the *Plasmepsin V* orthologous gene in *Plasmodium berghei* (*PbPMV*) could be disrupted, as the transfection methodology for this rodent malaria is more efficient than in *P. falciparum*, providing a more robust test of essentiality. Three attempts to disrupt the *PbPMV* gene in *P. berghei* were unsuccessful, indicating that it is essential (Supplementary Fig. 5). Third, we determined that PMV directly binds exported proteins *in vivo* by immunoprecipitation of the HA-tagged enzyme and proteomic analysis of interacting proteins. This showed that PMV was specifically immunoprecipitated and, importantly, that PEXEL-containing proteins (PF10_0159, MAL7P1.172 and PFE0060w) specifically coprecipitated with the protease (Fig. 4b and Supplementary Fig. 6). Fourth, we constructed transgenic parasites expressing a protein that would functionally bypass PMV. A control PfEMP3–GFP transgenic line was constructed expressing a protein that would be cleaved within the PEXEL by PMV to reveal the acetylated export signal ^{Ac}xQ (Fig. 4c). The second construct was based on PfEMP3–GFP but had a modified signal peptidase cleavage site and no PEXEL RxL, so it would be cleaved by signal peptidase to reveal the same mature protein (Fig. 4d). We confirmed that both processed proteins had ^{Ac}xQ at the N terminus and were identical at the C terminus (Fig. 4e–h). Additionally, mass spectrometry of trypsin-cleaved peptides across regions of each protein was consistent with their being identical (Supplementary Fig. 7). The
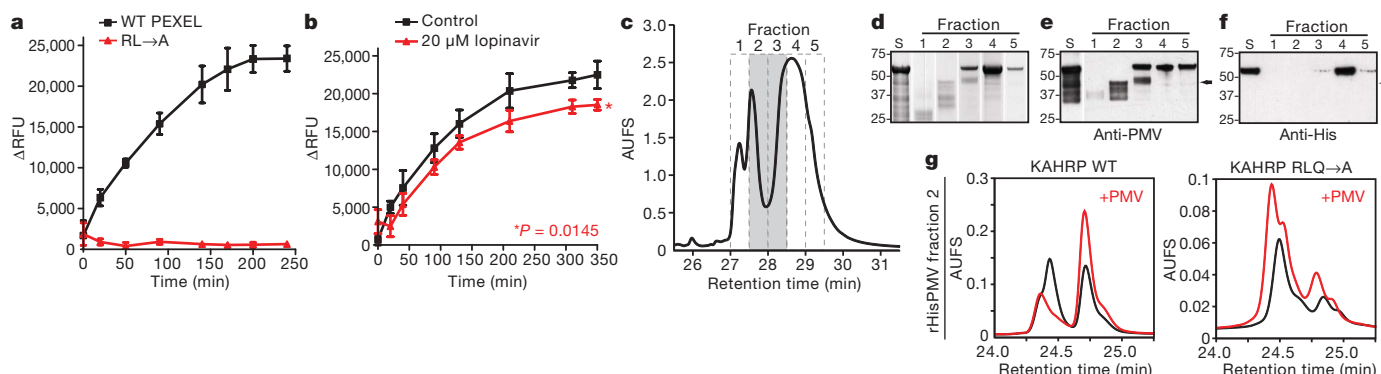

Figure 3 | Activity of PMV and cleavage of PEXEL. a, PMV cleavage of fluorescent substrate with PEXEL (black line) and PEXEL RL→A mutant (red line). Data shown is one representative experiment done in triplicate. Data shown is the mean ± standard deviation. RFU, relative fluorescence units. **b,** Lopinavir reduces PMV activity ($P = 0.0145$). Data shown is one representative experiment done in triplicate. Data shown is the mean ± standard deviation. **c,** N-terminal hexa-His fusion of PMV

(rHisPMV) from *E. coli* separated by HPLC. Elutions with activity indicated in grey. **d,** Starting material (S) and refolded fractions were Coomassie stained. **e,** Starting material (S) and fractions probed with anti-PMV antibodies. **f,** Starting material (S) and fractions probed with anti-hexa-His antibodies. **g,** HPLC of recombinant PMV activity (rHisPMV) with KAHRP peptides. (See Supplementary Information for details.)
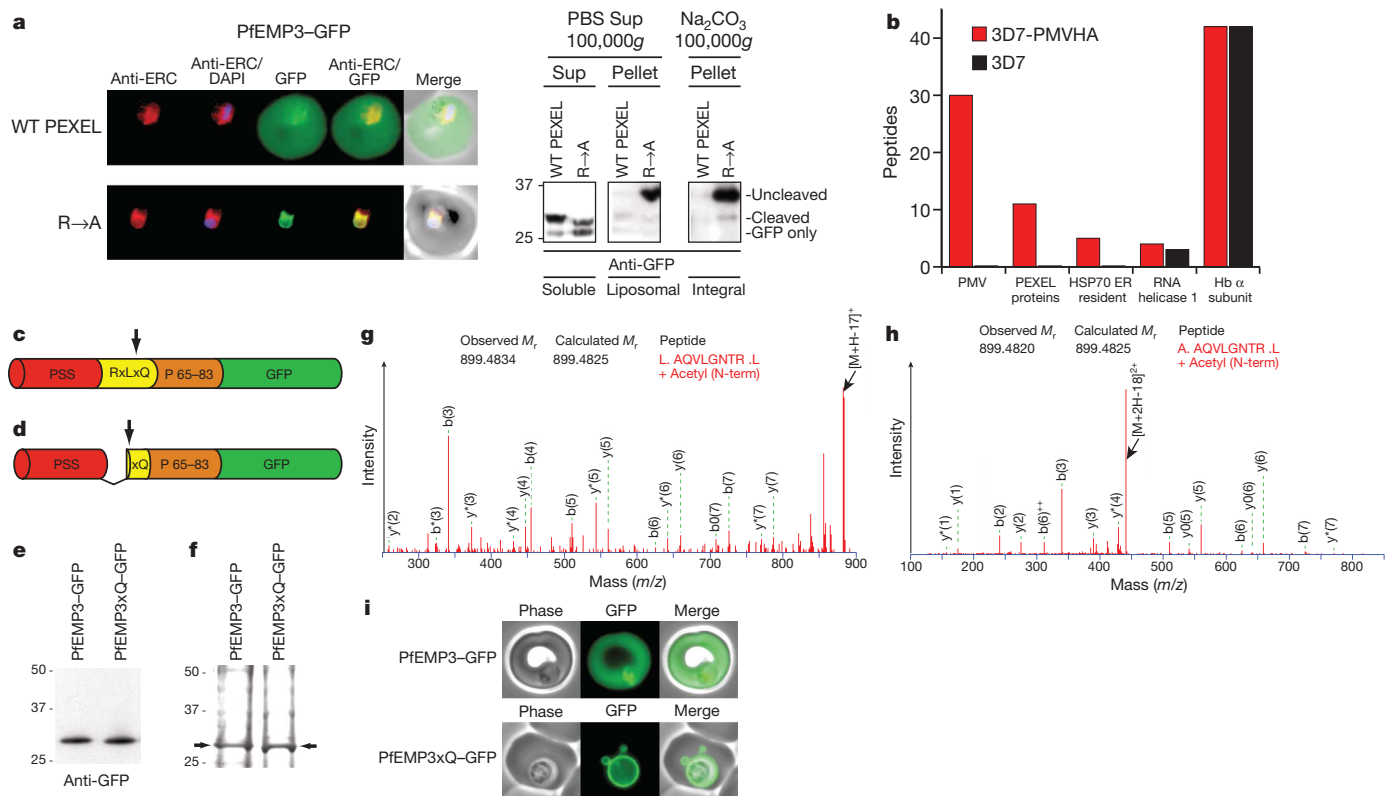
Figure 4 | Cleavage by PMV is essential for export. a, PfEMP3–GFP is exported (top) whereas R→A mutant accumulated in the endoplasmic reticulum (bottom). Right, fractionation showing full-length R→A mutant accumulates in membrane. Sup, Supernatant. **b**, Proteomic analysis of PMV immunoprecipitated proteins. Hb α, Haemoglobin α. **c**, Structure of PfEMP3–GFP. PSS, signal sequence; RxLxQ, PEXEL; residues 65–83 of PfEMP3. Arrow, PEXEL cleavage. **d**, Structure of PfEMP3xQ–GFP. Arrow, signal peptidase cleavage. **e**, N-terminal processing confirmed with anti-GFP. **f**, Immunopurified PfEMP3–GFP and PfEMP3xQ–GFP analysed by Coomassie and LC-MS/MS. **g**, Identification of N-terminal ^{Ac}AQVLGNTR of PfEMP3–GFP after PMV processing. **h**, Identification of ^{Ac}AQVLGNTR at N terminus of PfEMP3xQ–GFP after signal peptidase processing. **i**, Imaging of exported PfEMP3–GFP (top) and non-exported PfEMP3xQ–GFP (bottom). (See Supplementary Information for details.)

export profile of the two mature proteins was dramatically different: PMV-processed PfEMP3–GFP was efficiently exported to the erythrocyte whereas signal peptidase-cleaved PfEMP3xQ–GFP accumulated in the parasitophorous vacuole (Fig. 4i). Therefore, revealing the N-terminal signal for export (^{Ac}xE/Q/D) alone is not sufficient to reach the host cell; PMV must cleave the PEXEL and it then directs the next step in export to the host cell.

Role of PMV in export of *P. falciparum* proteins

How might PMV direct export of proteins into *P. falciparum*-infected erythrocytes? PMV has an N-terminal signal sequence for ER entry and a C-terminal transmembrane sequence such that it would be inserted into the ER membrane with the protease domain protruding into the lumen[24] (Supplementary Fig. 1). The signal sequences of exported proteins would direct cotranslational insertion into the ER membrane but are not efficiently cleaved[17] (Fig. 4a), indicating that the PEXEL is recognized, bound and cleaved before signal peptidase identifies its cleavage site. That ^{Ac}xE/Q/D alone is insufficient for export strongly indicates that PMV functions in a complex with another protein(s) that has exclusive access to PMV-matured cargo. We propose that this access may assist in sorting mature cargo to a subset of vesicles for trafficking to discrete domains of the parasitophorous vacuole in a manner distinct from general secretion but dependent on xE/Q/D[13,17] (Supplementary Fig. 1). Consistent with this is localization of the PTEX translocon on the parasitophorous vacuole membrane in discrete domains[13].

PMV is conserved in *Plasmodium* spp., including *Plasmodium vivax*[27,28], but is absent from higher eukaryotes; its identification as a critical enzyme for protein export provides an important target for development of novel antimalarials. HIV-1 protease inhibitors have been very successful treatments in the combat against HIV and so

these inhibitors may provide a platform for the design of novel anti-malarial compounds intended to impede the function of this important malarial protease.

METHODS SUMMARY

Transgenic parasites expressed PfEMP3–GFP or PfEMP3$_{R→A}$–GFP chimaeras from plasmids pPfEMP3$_{WT}$Glux.1 or PfEMP3$_{R→A}$Glux.1, respectively. Transgenic parasites expressing KAHRP–GFP contained the plasmid pKAHRPGlux.1, comprising DNA encoding residues 1–96 of KAHRP (PFB0100c). Parasites expressing ACPs–GFP were described previously[21]. Parasites expressing PfEMP3xQ–GFP contained the plasmid pPfEMP3xQGlux.1, encoding residues 1–36 of PfEMP3, followed by an alanine (to allow for signal peptidase cleavage at IYSEA-) and residues 63–82 of PfEMP3 (which include residues ^{63}aQ^{64} of the PEXEL) cloned into pGlux.1. The plasmids pPMVHA1.2 and pPMIXHA1.2 were integrated into either the PMV locus (PF13_0133) or plasmepsin IX locus (PF14_0281), respectively, to append HA tags. The transgenic parasites 3D7-PMVmutHA episomally expressed plasmid pTETPMVmutHA, which contains DNA encoding PMV with D118A, D365A and F370A substitutions fused to the same tags as in 3D7-PMVHA, cloned into the anhydrotetracycline (ATc) regulatable plasmid pTGFP[29]. The same DNA encoding PMV with D118A, D365A and F370A substitutions fused to HA, but with alternate cloning sites, was cloned into pGlux.1, removing GFP, to create pPMVmutHA2. This plasmid contains the constitutively expressed *crt* promoter and is non-inducible. Recombinant GBP130 or GBP130 RLE→A was expressed in *E. coli* harbouring plasmids pGBP130-3Cmyc18His or pGBP130-3A-3Cmyc18His, respectively, which encodes residues 66–196 of GBP130 (containing the wild-type or mutant PEXEL). Recombinant PMV was expressed in *E. coli* harbouring plasmid pHisPMVtrunc, which contains DNA encoding residues 37–521 of PMV cloned into pProExHTb.

PMV activity was assayed either with anti-HA beads containing ipPMVHA or ipPMVmutHA or 3–5 μg of HPLC-eluted protein added to the reaction that contained either PEXEL peptide substrate, recombinant PEXEL substrate or fluorogenic PEXEL peptide substrate. Digest reactions were analysed by HPLC and LC-MS/MS, immunoblot and fluorescence intensity, respectively.

1.   Snow, R. W. *et al.* The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434**, 214–217 (2005).
2.   Miller, L. H., Good, M. F. & Milon, G. Malaria pathogenesis. *Science* **264**, 1878–1883 (1994).
3.   Raventos-Suarez, C., Kaul, D. K., Macaluso, F. & Nagel, R. L. Membrane knobs are required for the microcirculatory obstruction induced by *Plasmodium falciparum*-infected erythrocytes. *Proc. Natl Acad. Sci. USA* **82**, 3829–3833 (1985).
4.   Barnwell, J. W. *et al.* A human 88-kD membrane glycoprotein (CD36) functions *in vitro* as a receptor for a cytoadherence ligand on *Plasmodium falciparum*-infected erythrocytes. *J. Clin. Invest.* **84**, 765–772 (1989).
5.   Marti, M. *et al.* Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930–1933 (2004).
6.   Hiller, N. L. *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934–1937 (2004).
7.   Pei, X. *et al.* Structural and functional studies of interaction between *Plasmodium falciparum* knob-associated histidine-rich protein (KAHRP) and erythrocyte spectrin. *J. Biol. Chem.* **280**, 31166–31171 (2005).
8.   Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77–87 (1995).
9.   Su, X. Z. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**, 89–100 (1995).
10. Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101–110 (1995).
11. Maier, A. G. *et al.* Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes. *Cell* **134**, 48–61 (2008).
12. Maier, A. G., Cooke, B. M., Cowman, A. F. & Tilley, L. Malaria parasite proteins that remodel the host erythrocyte. *Nature Rev. Microbiol.* **7**, 341–354 (2009).
13. de Koning-Ward, T. F. *et al.* A newly discovered protein export machine in malaria parasites. *Nature* **459**, 945–949 (2009).
14. Sargeant, T. J. *et al.* Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol.* **7**, R12 (2006).
15. van Ooij, C. *et al.* The malaria secretome: from algorithms to essential function in blood stage infection. *PLoS Pathog.* **4**, e1000084 (2008).
16. Chang, H. H. *et al.* N-terminal processing of proteins exported by malaria parasites. *Mol. Biochem. Parasitol.* **160**, 107–115 (2008).
17. Boddey, J. A., Moritz, R. L., Simpson, R. J. & Cowman, A. F. Role of the *Plasmodium* export element in trafficking parasite proteins to the infected erythrocyte. *Traffic* **10**, 285–299 (2009).
18. Przyborski, J. M. *et al.* Trafficking of STEVOR to the Maurer's clefts in *Plasmodium falciparum*-infected erythrocytes. *EMBO J.* **24**, 2306–2317 (2005).
19. Weber, I. T. *et al.* Molecular modeling of the HIV-1 protease and its substrate binding site. *Science* **243**, 928–931 (1989).
20. Miller, M. *et al.* Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. *Science* **246**, 1149–1152 (1989).
21. Waller, R. F., Reed, M. B., Cowman, A. F. & McFadden, G. I. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *EMBO J.* **19**, 1794–1802 (2000).
22. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, e5 (2003).
23. Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503–1508 (2003).
24. Klemba, M. & Goldberg, D. E. Characterization of plasmepsin V, a membrane-bound aspartic protease homolog in the endoplasmic reticulum of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **143**, 183–191 (2005).
25. La Greca, N. *et al.* Identification of an endoplasmic reticulum-resident calcium-binding protein with multiple EF-hand motifs in asexual stages of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **89**, 283–293 (1997).
26. Knuepfer, E., Rug, M. & Cowman, A. F. Function of the *Plasmodium* export element can be blocked by green fluorescent protein. *Mol. Biochem. Parasitol.* **142**, 258–262 (2005).
27. Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757–763 (2008).
28. Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455**, 799–803 (2008).
29. Meissner, M. *et al.* Tetracycline analogue-regulated transgene expression in *Plasmodium falciparum* blood stages using *Toxoplasma gondii* transactivators. *Proc. Natl Acad. Sci. USA* **102**, 2980–2985 (2005).

METHODS

Parasites, plasmid construction and transfection. *P. falciparum* strain 3D7 and transgenic parasites were cultured in human O+ erythrocytes at 4% haematocrit as described previously[17]. Transgenic parasites expressed PfEMP3–GFP or PfEMP3$_{R\rightarrow A}$-GFP chimaeras from plasmids pPfEMP3$_{WT}$Glux.1 or PfEMP3$_{R\rightarrow A}$Glux.1, respectively. These constructs contained DNA encoding residues 1–38 of PfEMP3 (PFB0095c), double HA followed by PfEMP3 residues 56–82, which contained either a native PEXEL or PEXEL R→A point mutation, respectively (DNA synthesized by Epoch Biolabs) cloned into pGlux.1 (ref. 17) using XhoI and XmaI sites to generate GFPmut2 fusions. Transgenic parasites expressing KAHRP–GFP contained the plasmid pKAHRPGlux.1, comprising DNA encoding residues 1–96 of KAHRP (PFB0100c; DNA synthesized by Epoch Biolabs) cloned into pGlux.1 with XhoI and XmaI to fuse with GFPmut2. Parasites expressing ACPs–GFP have been described previously[21]. Parasites expressing PfEMP3xQ–GFP contained the plasmid pPfEMP3xQGlux.1, which contained DNA encoding residues 1–36 of PfEMP3 followed by an alanine (to allow for signal peptidase cleavage at IYSEA-) and residues 63–82 of PfEMP3 (which include residues ^{63}aQ^{64} of the PEXEL; DNA synthesized by Epoch Biolabs) cloned into pGlux.1 using XhoI and XmaI to fuse with GFPmut2. Transgenic parasites 3D7-PMVHA and 3D7-PMIXHA contained the plasmids pPMVHA1.2 or pPMIXHA1.2, respectively, which integrated into either the PMV locus (PF13_0133) or plasmepsin IX locus (PF14_0281), respectively. The pPMVHA1.2 insert was generated by amplifying the 3′ end of PF13_0133 using the oligonucleotides JB116 5′-AT CAGATCTGGGCTGTCATATGCATGAAG-3′ and JB117 5′-ATCCTGCAGG TGTTGATTCCTGTATGGGAG-3′ and the pPMIXHA1.2 insert was generated by amplifying the 3′ end of PF14_0281 using the oligonucleotides PF14_0281-HA-fwd 5′-AGATCTGTGATGATAGTAACATTGATCAG-3′ and PF14_0281-HA-rev 5′-CTGCAGCTAAATTATTTATTTTATTATGTAAGG-3′. Inserts were individually cloned into p1.2-SHA (ref. 30) using BglII and PstI (underlined) to generate pPMVHA1.2 or pPMIXHA1.2, respectively. The transgenic parasites 3D7-PMVmutHA episomally expressed plasmid pTETPMVmutHA, which contains DNA encoding PMV with D118A, D365A and F370A substitutions fused to exactly the same tags as in 3D7-PMVHA (DNA synthesized by Epoch Biolabs), cloned into the anhydrotetracycline (ATc) regulatable plasmid pTGFP (ref. 31) via PstI and SpeI, which removed GFP. The transgenic parasites 3D7-PMVmutHA2 episomally expressed plasmid pPMVmutHA2, which contains DNA encoding PMV with D118A, D365A and F370A substitutions fused to exactly the same tags as in 3D7-PMVHA (DNA synthesized by Epoch Biolabs), cloned into the non-inducible *crt* promoter-containing plasmid, pGlux.1, via XhoI and PacI, which removed GFP. Recombinant GBP130 was expressed in *E. coli* harbouring plasmid pGBP130-3Cmyc18His, which contains codon optimized DNA encoding residues 66 to 196 (containing the PEXEL) of GBP130 fused to triple C-myc tags and triple hexa-His tags (DNA synthesized by Epoch Biolabs) cloned into pET11a using NdeI and BamHI. The PEXEL mutant GBP130 RLE→A protein was expressed as above but contained R84A, L86A and E88A mutations. Recombinant PMV was expressed in *E. coli* harbouring plasmid pHisPMVtrunc, which contains DNA encoding residues E37 to N521 of PMV (DNA synthesized by GENEART AG) cloned into pProExHTb (Invitrogen), using BamHI and XhoI to create an N-terminal hexa-His fusion protein.

All plasmids (100 µg) were transfected into *P. falciparum* 3D7 as described previously[32] and selected with WR99210 and, where appropriate, ATc.

HIV protease inhibitor assays. *In vivo* assays: tightly synchronous, 20-h-old transgenic parasites (expressing chimaeras from the *crt* promoter) at 15% parasitaemia were cultured with 50 µM HIV protease inhibitors solubilized in DMSO for 7 h. Saponin-treated parasite pellets were washed twice in media containing the inhibitor and solubilized in 4× reducing SDS–PAGE sample buffer. *In vitro* cleavage assays: ipPMVHA activity was assessed with fluorogenic substrates (see below) in the presence of lopinavir at various concentrations solubilized in 100% ethanol, or ethanol alone (maximum 2% final ethanol concentration).

Indirect immunofluorescence analysis (IFA). Erythrocytes infected with transgenic parasites were either viewed live or fixed in 100% ice-cold methanol for 3 min and air dried or fixed in 4% paraformaldehyde, 0.01% glutaraldehyde as described previously[17]. Cells were incubated with rabbit anti-Pf ERC (1:500)[25], 3F10 rat anti-HA (1:50) or mouse anti-PMV (1:5) for 2 h followed by Alexa Fluor 488- or 594-conjugated secondary IgG antibodies (1:1,000; Molecular Probes) for 1 h and nuclei were labelled with DAPI nuclear stain (Roche) at 0.2 µg ml^{-1} in Vectashield (Vector Labs). Cells were viewed with a Carl Zeiss Axioskop 2 microscope (Thornwood) and images collected using a PCO SensiCam (Motion Engineering) and Axiovision 3 software (Carl Zeiss). Images were assembled with Photoshop CS4 (Adobe).

Immunoprecipitations. PMVHA, PMVmutHA2 and PMIXHA was affinity purified from trophozoite parasites at >5% parasitaemia by solubilizing saponin pellets in 1% (v/v) Triton X-100 in PBS by gentle sonication and rotation at 4 °C

for 1 h. The supernatant was mixed with goat anti-HA antibodies coupled to agarose beads (Sapphire Bioscience) for 1 h at 4 °C and the beads washed six times with PBS, and resuspended in the same buffer. Beads could be stored with active enzyme for 5 days (longer was not tested). PMVmutHA was affinity purified as above from late stage parasites (expressed from *MSP2* promoter in pTETPMVmutHA) using magnet separation through CS columns (Miltenyi Biotech) and whole cells solubilized in 1% (v/v) Triton X-100 in PBS.

GFP-tagged chimaeras were affinity purified from saponin supernatants for mass spectrometry using one litre of each parasite line and anti-GFP agarose (MBL) exactly as described previously[17].

Recombinant PMV expression and purification. PMV was produced and purified in a manner similar to that described previously[33]. Briefly, the protein was synthesized in *E. coli* strain BL21(DE3) for 3 h at 37 °C and deposited into insoluble inclusion bodies. The cells were lysed by sonication and the insoluble inclusion bodies solubilized by the addition of 6 M guanidine-HCl, pH 8.0. The solubilized protein was isolated by metal-chelate chromatography using Ni-NTA agarose resin (Qiagen), and eluted in 8 M urea, 1 M imidazole, pH 8.0. The protein was then HPLC purified as described later; then various fractions containing PMV were concentrated using a vacuum concentrator (Centrivac, Labconco). The acidified PMV fractions were then added directly to digest solutions and incubated for various periods at 37 °C.

PMV cleavage assays. Cleavage assays with synthetic PEXEL peptide substrate: PMV activity was measured in 100 µl total digest volume consisting of digest buffer (25 mM Tris hydrochloride + 25 mM 4-morpholine ethane sulphonic acid monohydrate, pH 6.4), 2 nmol PEXEL peptide substrate (synthesized by GenScript USA) and either 5 or 20 µl of anti-HA beads containing ipPMVHA or ipPMVmutHA, respectively (<400 µg). Digests were carried out overnight at 37 °C, with shaking. The following day, beads with bound enzyme were removed by passing the reaction through Micro Bio-Spin Chromatography Columns (Bio-Rad) and collecting the flow through, which was analysed using an Agilent 1100 modular HPLC (see below). For activity with the recombinant PMV, approximately 3–5 µg of each fraction (1–5) of HPLC-eluted protein was added directly to the reaction, as above. Cleavage assays with recombinant PEXEL substrate: approximately 1 µg of recombinant GBP130, or the PEXEL mutant protein GBP130 RLE→A, was incubated with ipPMVHA in digest buffer (30 µl total volume) for 1 h at 37 °C with shaking. Beads with bound enzyme were removed, as above, and the flow through was resuspended in SDS loading buffer and analysed by SDS–PAGE and immunoblot with anti-C-myc antibodies. Cleavage assays with fluorogenic PEXEL peptide substrate: 5 µM FRET substrate (DABCYL-G-NKRTLAQKQ-G-EDANS or PEXEL mutant DABCYL-G-NKATAAQKQ-G-EDANS; GenScript) in water was incubated with 0.5 µl of anti-HA beads containing ipPMVHA or ipPMVmutHA2 (<100 µg) in digest buffer (100 µl total volume). Reactions were prepared on ice and the fluorescence intensity (excitation 340 nm, emission 492 nm) was measured using a GENios fluorescence plate reader (Tecan Group) heated to 37 °C with orbital shaking between measurements.

After all cleavage assays, enzyme was eluted from the beads retained in the column in non-reducing sample buffer, reduced in 0.1 M dithiothreitol and analysed by SDS–PAGE and immunoblotted with anti-HA antibodies.

Reversed-phase HPLC analysis of the refolding reaction. Reversed-phase HPLC was performed using an Agilent 1100 modular HPLC consisting of an on-line degasser, piston pump, autosampler, column oven, diode-array detector and fraction collector. Instrument control, data acquisition and evaluation were performed using Hewlett-Packard ChemStation software for LC and LC/MS systems. Buffer A was 0.05% (v/v) trifluoroacetic acid (HPLC/Spectro grade, Pierce) in Milli-Q grade water (Millipore), whereas buffer B was 0.05% (v/v) trifluoroacetic acid in acetonitrile (ChromAR HPLC grade, Malinckrodt Baker). Protein/peptide samples were centrifuged at 10,000*g* for 15 min at room temperature (25 °C) before loading onto a C8 column (2.1 mm internal diameter × 100 mm; Brownlee columns, Perkin-Elmer Instruments) in the presence of buffer A. Bound proteins/peptides were eluted using a linear gradient of 0–100% buffer B over 12 min at a flow-rate of 0.5 ml min^{-1} at 37 °C.

Subcellular fractionation, SDS–PAGE and immunoblots. To assess subcellular localization of PfEMP3–GFP and PfEMP3 R→A–GFP, infected erythrocytes were lysed in PBS by gentle sonication and centrifuged at 10,000*g* for 15 min to obtain PBS supernatant and PBS pellet fractions. The PBS supernatant was centrifuged at 100,000*g* for 1 h to separate soluble proteins (supernatant) and liposomal membranes (pellet). The PBS pellet was resuspended in 0.2 M Na$_2$CO$_3$ by sonication and incubated on ice for 1 h and centrifuged at 100,000*g* for 1 h to obtain integral membrane proteins (pellet). Fractions were solubilized in SDS–PAGE loading buffer. For SDS–PAGE, samples were boiled and electrophoresed in 4–12% or 10% Bis/Tris precast polyacrylamide gels (Invitrogen) using MOPS or MES running buffer and transferred to nitrocellulose membrane using an iBlot (Invitrogen). Membranes were blocked for 1 h in 5% skim milk (Diploma)

in PBS containing 0.1% Tween-20 (Sigma). Membranes were probed with either mouse anti-GFP antibodies (Roche; 1:1,000), mouse anti-PMV antibodies (1:300), 3F10 rat anti-HA antibodies (Roche; 1:1,000), mouse anti-cMyc antibodies (Sigma; 1:1,000) or rabbit anti-His probe antibodies (H-15: sc-803; Santa Cruz Biotechnology; 1:1,000) followed by horseradish peroxidase-conjugated secondary antibodies (Silenius; 1:2,000) and visualized using enhanced chemiluminescence (ECL; Amersham). Bio-Rad prestained precision markers were used throughout.

Gel excision, digestion and nano-LC-MS/MS. For protein samples in gels, bands (~2 mm) were excised from the one-dimension gel lane and subjected to automated in-gel reduction, alkylation and tryptic digestion using the MassPREP Station (Micromass) as described previously[17]. Digested peptides were subjected to manual in-gel digestion. The gel bands were reduced with 10 mM DTT (Calbiochem) for 30 min, alkylated for 30 min with 50 mM iodoacetic acid (Fluka) and digested with 375 ng trypsin (Worthington) for 16 h at 37 °C. The extracted peptide solutions (0.1% formic acid) were then concentrated to approximately 10 μl by centrifugal lyophilization using a SpeedVac AES 1010 (Savant). Digested peptides were subjected to nano liquid chromatography in conjunction with collisional tandem mass spectrometry (nano-LC-MS/MS) as described previously[17]. Briefly, extracted peptides were injected and fractionated by nanoflow reversed-phase liquid chromatography on a nano-LC system (1200 series, Agilent) using a nanoAcquity C18 150 mm × 0.15 mm internal diameter column (Waters) developed with a linear 60-min gradient with a flow rate of 0.5 μl min^{-1} at 45 °C from 100% solvent A (0.1% formic acid in Milli-Q water) to 100% solvent B (0.1% formic acid, 60% acetonitrile, (Mallinckrodt Baker) 40% Milli-Q water). The nano-HPLC was coupled on-line to an LTQ-Orbitrap mass spectrometer equipped with a nanoelectrospray ion source (Thermo Fisher Scientific) for automated MS/MS. Up to five of the most intense ions per cycle were fragmented and analysed in the linear trap, with target ions already selected for MS/MS being dynamically excluded for 3 min.

Mass spectra, protein identification, bioinformatic and statistical analysis. Mass spectra peak lists were extracted using extract-msn as part of Bioworks 3.3.1 (Thermo Fisher Scientific) linked into Mascot Daemon (Matrix Science). The parameters used to generate the peak lists for the Orbitrap were as follows: minimum mass 400; maximum mass 5,000; grouping tolerance 0.01 Da;

intermediate scans 1; minimum group count 1; 10 peaks minimum and total ion current of 100. Peak lists for each nano-LC-MS/MS run were used to search MASCOT v2.2.04 search algorithm (Matrix Science) provided by the Australian Proteomics Computational Facility (www.apcf.edu.au). Automatic charge state recognition was used because of the high-resolution survey scan (30,000). LC-MS/MS files were searched against a PEXEL sequence database (552,306 entries) comprising sequences from the latest version of Swiss-Prot (all species), Trembl (*Plasmodium* entries), plasmoDB (all entries), synthetic peptides and the PEXEL constructs. The search parameters consisted of carboxymethylation of cysteine as a fixed modification (+58 Da) with variable modifications set for NH_2-terminal acetylation (+42 Da) and oxidation of methionine (+16 Da). A peptide mass tolerance of ± 20 p.p.m., ^{13}C defined as 1, fragment ion mass tolerance of ±0.8 Da, and an allowance for up to three missed cleavages for tryptic searches where trypsin was used as the digest reagent. No-enzyme (unrestricted) searches were also carried out on all Orbitrap-LTQ data and manual validation of all peptide spectral matches was done irrespective of peptide scores or expectation values (*E*-values) to ensure that all major fragment ions were annotated in accordance with known rules of peptide fragmentation.

The signal sequence-processing site of PfEMP3–GFP and PfEMP3xQ–GFP was predicted by SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) with both neural networks and hidden Markov models trained on eukaryotes[34]. Statistical analyses were performed with GraphPad Prism 5 for Windows.

30. Baum, J. *et al.* A malaria parasite formin regulates actin polymerization and localizes to the parasite-erythrocyte moving junction during invasion. *Cell Host Microbe* **3**, 188–198 (2008).

31. Gilson, P. R. *et al.* MSP1$_{19}$ miniproteins can serve as targets for invasion inhibitory antibodies in *Plasmodium falciparum* provided they contain the correct domains for cell surface trafficking. *Mol. Microbiol.* **68**, 124–138 (2008).

32. Duraisingh, M. T., Triglia, T. & Cowman, A. F. Negative selection of *Plasmodium falciparum* reveals targeted gene deletion by double crossover recombination. *Int. J. Parasitol.* **32**, 81–89 (2002).

33. Hodder, A. N. *et al.* Structural insights into the protease-like antigen *Plasmodium falciparum* SERA5 and its noncanonical active-site serine. *J. Mol. Biol.* **392**, 154–165 (2009).

34. Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).

ARTICLES

Plasmepsin V licenses *Plasmodium* proteins for export into the host erythrocyte

Ilaria Russo[1]†, Shalon Babbitt[1]*, Vasant Muralidharan[1]*, Tamira Butler[1], Anna Oksman[1] & Daniel E. Goldberg[1]

During their intraerythrocytic development, malaria parasites export hundreds of proteins to remodel their host cell. Nutrient acquisition, cytoadherence and antigenic variation are among the key virulence functions effected by this erythrocyte takeover. Proteins destined for export are synthesized in the endoplasmic reticulum (ER) and cleaved at a conserved (PEXEL) motif, which allows translocation into the host cell via an ATP-driven translocon called the PTEX complex. We report that plasmepsin V, an ER aspartic protease with distant homology to the mammalian processing enzyme BACE, recognizes the PEXEL motif and cleaves it at the correct site. This enzyme is essential for parasite viability and ER residence is essential for its function. We propose that plasmepsin V is the PEXEL protease and is an attractive enzyme for antimalarial drug development.

The human malaria parasite *Plasmodium falciparum* exports an estimated 200–300 proteins into the host erythrocyte[1,2]. In doing so, the parasite remodels the cytoskeleton and plasma membrane to create cytoadherence knobs, nutrient permeation pathways and altered erythrocyte mechanical stability[3,4]. Export of these effectors is dependent on a *Plasmodium* export element or PEXEL sequence, RxLxE/Q/D[5,6]. Proteins destined for export are cleaved after the conserved PEXEL leucine in the ER and mutation of the R or L residues attenuates cleavage and export[7,8]. Plasmepsin V (PMV) is an aspartic protease that has distant homology to mammalian BACE or beta-secretase[9], an enzyme involved in the processing of amyloid precursor protein[10]. Both have a carboxy-terminal extension that contains a hydrophobic membrane anchor sequence. An amino-terminal aspartic protease 'pro-domain' remains unprocessed in PMV[9]. PMV is expressed in intraerythrocytic *P. falciparum* parasites and has orthologues in other *Plasmodium* species. *Phytophthora infestans*, the potato blight pathogen that has a similar export system[11], has a homologous sequence in the database (PITG_02623.1). PMV has been localized to the ER[9] and is therefore a candidate to be the PEXEL processing protease.

Role of the transmembrane domain in PMV localization

PMV lacks a classical ER retention signal. To identify the element responsible for localization and to assess the importance of ER residence for PMV function, we made sequential C-terminal truncation mutants (Fig. 1a). Single crossover homologous recombination into the endogenous locus was performed, introducing a green florescence protein (GFP) tag after a full-length C terminus or in place of C-terminal sequence (Fig. 1b). Full-length PMV–GFP integrants (clone DC6) and integrants with deletion of the C terminus downstream of the membrane-spanning segment (clone EF2) had no phenotype and retained ER targeting (Fig. 1c–e and Supplementary Fig. 2), whereas deletions involving the membrane anchor were lethal (Fig. 1b). Fusion of the transmembrane region but not other portions of PMV was sufficient to target a reporter to the ER (Fig. 1f). Thus the transmembrane sequence is important for ER localization and probably for cellular activity on its substrates as well.
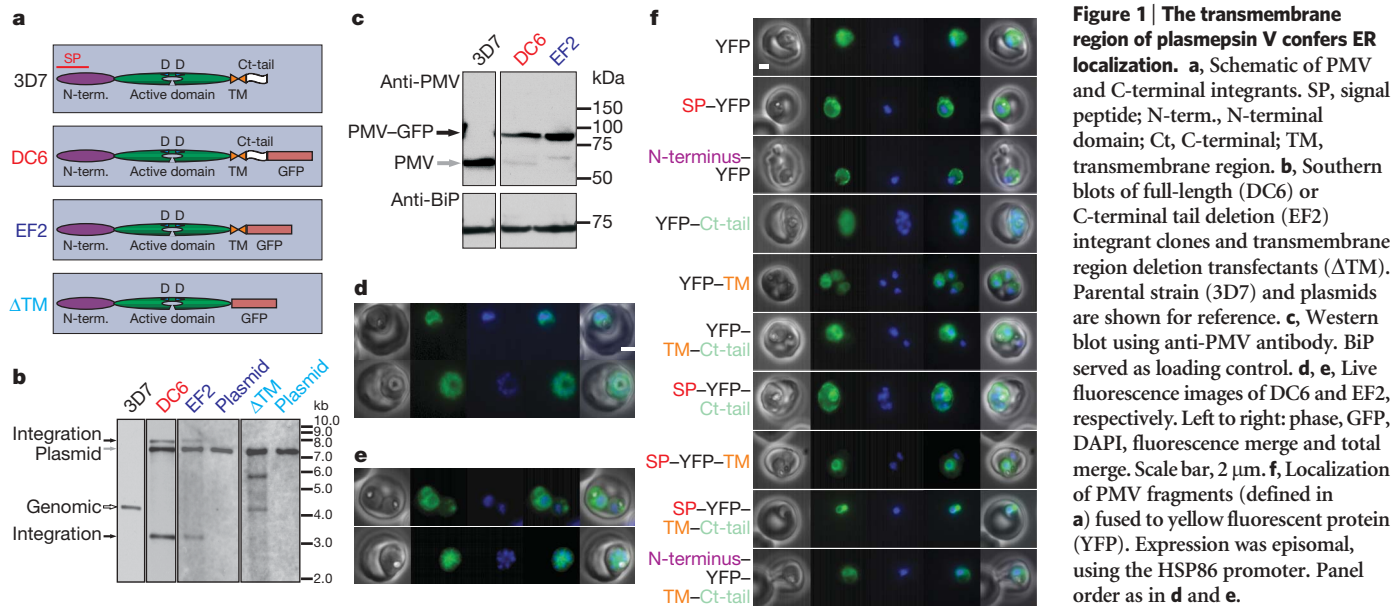
PMV essentiality

We further assessed essentiality of the PMV gene for intraerythrocytic parasites by using an allelic replacement approach[12]. An integration vector in which the first catalytic aspartate was modified by synonymous or non-synonymous mutation was transfected into parasites (Fig. 2a) and recombinants were obtained (Fig. 2b, c). Crossover into the endogenous PMV gene proved possible only when the aspartate codon was preserved (replacement via the synonymous mutation vector or downstream crossover bypassing the mutation with the non-synonymous mutation vector) (Fig. 2d). Non-synonymous alteration of the active site codon could not be achieved in four separate transfection experiments. These results support the notion that PMV has an essential function in the cell.

Dominant-negative PMV phenotype

To investigate PMV function, we episomally expressed two GFP-tagged versions of PMV, one wild type and the other containing a D to A mutation in the active site aspartate 108 to render expressed protein catalytically dead. Both versions localized to the ER but the mutant had threefold reduced signal by immunofluorescence (Fig. 3a–c) and by western blot (Fig. 3d). Mutant enzyme-expressing parasites were frequently seen encased in erythrocyte ghosts (Fig. 3e), indicating impaired host cell homeostasis. Indeed the mutant PMV-expressing culture grew more slowly than the wild-type PMV-expressing culture (Fig. 3f). Occasional cells expressing the mutant construct had intense signal similar to wild type; in such cases two parasites could be seen in a single erythrocyte, one of which was not fluorescent and had presumably lost or downregulated the plasmid (Fig. 3g). Our interpretation of this result is that the plasmid-free parasite can export proteins normally into the shared red blood cell, overcoming the effect that the mutant PMV has on the neighbouring parasite. To explore this further, we assessed processing of the exported histidine-rich protein II (HRPII) in the PMV-transfected cells (Fig. 3h, i). Unprocessed HRPII was barely detectable in wild-type PMV-expressing parasites. In contrast, unprocessed HRPII accumulated in mutated PMV-expressing parasites. Plasmepsin II and DPAP1, non-PEXEL containing proteins that use the secretory pathway but are then internalized instead of exported[13,14], were

[1]Howard Hughes Medical Institute, Washington University School of Medicine, Departments of Molecular Microbiology and Medicine, St Louis, Missouri 63110, USA. †Present address: Department of Experimental Medicine and Biochemical Science, University of Perugia, 06126 Perugia, Italy.
*These authors contributed equally to this work.

Figure 1 | The transmembrane region of plasmepsin V confers ER localization. a, Schematic of PMV and C-terminal integrants. SP, signal peptide; N-term., N-terminal domain; Ct, C-terminal; TM, transmembrane region. **b,** Southern blots of full-length (DC6) or C-terminal tail deletion (EF2) integrant clones and transmembrane region deletion transfectants (ΔTM). Parental strain (3D7) and plasmids are shown for reference. **c,** Western blot using anti-PMV antibody. BiP served as loading control. **d, e,** Live fluorescence images of DC6 and EF2, respectively. Left to right: phase, GFP, DAPI, fluorescence merge and total merge. Scale bar, 2 μm. **f,** Localization of PMV fragments (defined in **a**) fused to yellow fluorescent protein (YFP). Expression was episomal, using the HSP86 promoter. Panel order as in **d** and **e**.

processed normally (Fig. 3i). Export was assessed by immunofluorescence (Fig. 3j, k). The levels of host erythrocyte HRPII and another exported protein, RESA (ring-infected erythrocyte surface antigen), were diminished in the mutated PMV-expressing parasitized erythrocytes by 30–50%. These data indicate that episomal expression of catalytically dead PMV has a dominant-negative effect on parasite growth and on protein export. A survey of PEXEL gene essentiality estimated that about one-fourth are required for intraerythrocytic parasite growth[15]. Thus, perhaps 50–75 exported proteins are essential. Some will be required at near wild-type levels for optimal growth, whereas others will tolerate more drastic reduction without consequence. The 30–50% reduction in protein export is therefore about what would be expected, given the growth phenotype seen (Fig. 3f).

Enzyme activity and specificity

To assess enzyme activity, PMV–GFP was detergent-solubilized from recombinant clone DC6 (see Fig. 1a, c) and enzyme was isolated using anti-GFP antibody. The enzyme was able to cleave a fluorogenic decapeptide based on the PEXEL motif from the exported HRPII (Fig. 4a, b). Pull-downs from the parental strain (3D7) with untagged PMV had no activity. When anti-PMV antibody was used for enzyme isolation, both tagged (DC6) and untagged (3D7) enzyme could be isolated and were active. A second PEXEL peptide based on PfEMP2 was also cleaved by isolated enzyme (Fig. 4b). Mutation of P1 Leu or P3 Arg abolishes export of PEXEL proteins[10]. Peptides with either of these residues changed to Ala were refractory to cleavage (Fig. 4b), confirming specificity of the enzyme for the
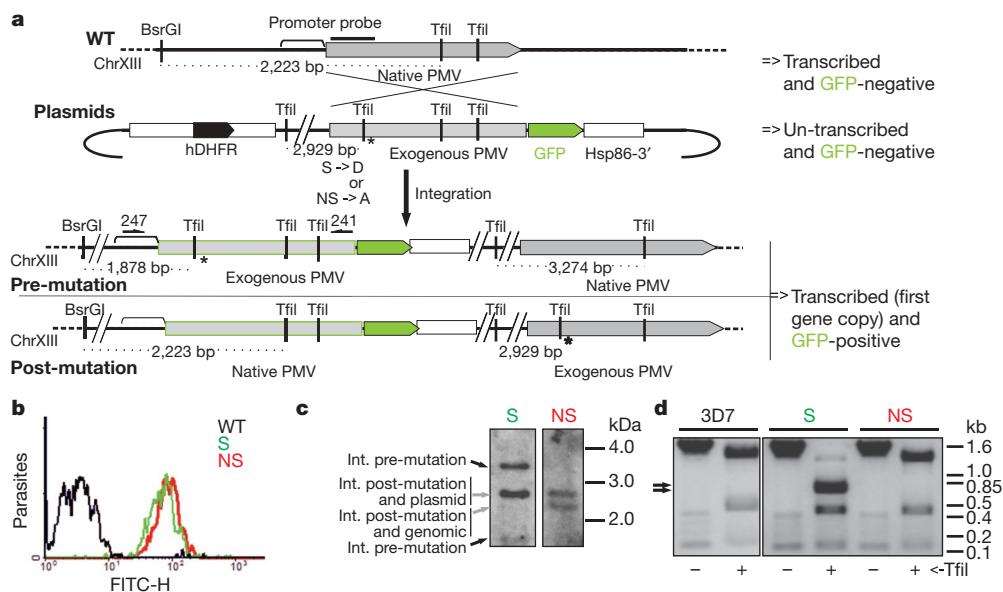
Figure 2 | Plasmepsin V is essential for intraerythrocytic parasite viability. a, Active site allelic replacement scheme. Integration vectors possessing a synonymous (S) or non-synonymous (NS) mutation in the Asp 108 codon and a new TfiI restriction site just upstream were transfected into parental strain 3D7. Possible outcomes for upstream crossover (pre-mutation) and downstream crossover (post-mutation) are shown. Integrants were selected and assessed by PCR. Primers used for amplification (241 and 247) are marked. **b,** Flow cytometry of parasites. Parasites were assessed for GFP expression, an indication of integration at the endogenous locus. Parental strain 3D7 (WT) is shown as control. **c,** Southern blot of transfected parasite pools. Int., integration. At left are expected positions of possible products of the BsrGI/TfiI digest mapped in **c. d,** Screening of integrants by PCR and restriction digest. Arrows indicate a doublet (predicted 787 and 748 bp) resulting from restriction fragmentation of product from recombinants that have crossed over upstream of the active site codon.
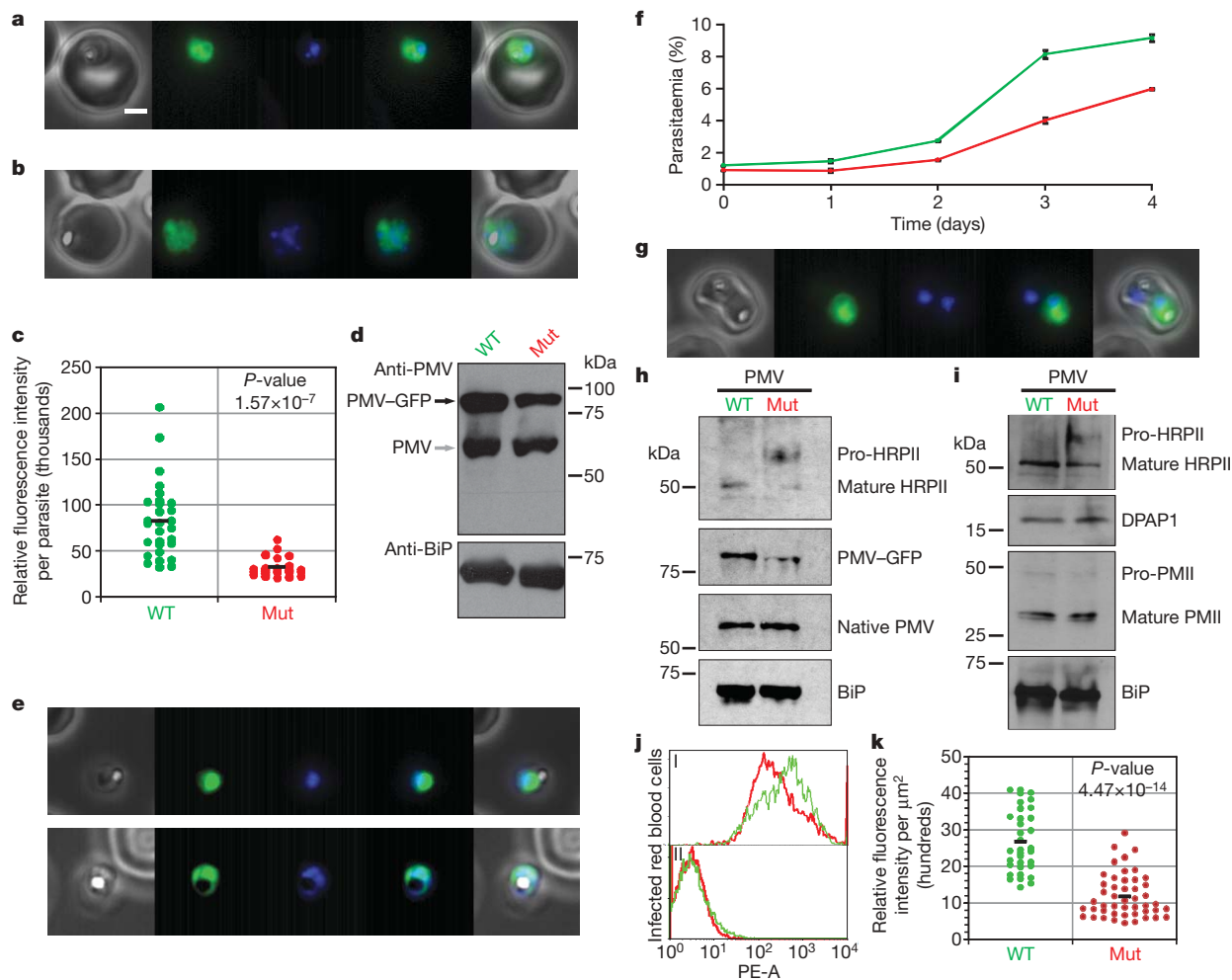
Figure 3 | Episomal expression of wild-type and catalytic mutant PMV.
a, **b**, Immunofluorescence of wild-type and 108 D to A (mutant) PMV
expressing lines, respectively. Left to right: phase, GFP, DAPI, merge. Bar,
2 μm. **c**, Fluorescence intensity of 52 transfected early trophozoites was
measured. Mean relative fluorescence units were 82,559 and 32,059 for wild-
type and mutant-transfected parasites, respectively. **d**, Western blot. BiP
serves as a loading control. **e**, Mutant PMV-expressing parasites encased in
erythrocyte ghosts. **f**, Growth curves. Asynchronous cultures episomally
expressing wild-type (green) or mutant (red) PMV were monitored by flow
cytometry for four days in triplicate (error bars indicate standard deviation).
Mutant growth rate is reduced by 49.2 ± 1.6%. **g**, Double-infected
erythrocyte containing a parasite expressing high levels of mutant PMV.
h, **i**, Western blots of HRPII processing in parasites episomally expressing
mutant or wild-type PMV. **h**: top, anti-HRPII; middle panels, anti-PMV;

bottom, anti-BiP. Contrast and brightness enhanced slightly over the entire
panel to bring out features seen on the original film. **i**: top, anti-HRPII;
middle panels, antibodies against non-exported secretory proteins
dipeptidyl peptidase I (DPAP1) and plasmepsin II (PMII); bottom, anti-BiP.
j, Flow cytometry. Wild-type (green) and mutant (red) PMV-expressing
parasites were fixed, treated with tetanolysin (5 min, 20 units ml^{-1}) to
selectively permeabilize the erythrocyte compartment and exported RESA
was quantified by flow cytometry with specific antibody. Control without
primary antibody is shown below. Results are representative of four
experiments. **k**, HRPII staining was inconsistent by the fixation procedure
used in **j**, so HRPII erythrocyte fluorescence intensity (80 cells) was
quantified. Mean relative fluorescence intensity per μm^2 was 2,671 and 1,179
for wild-type and mutant, respectively.

PEXEL motif. Activity was seen between pH 5 and 7 (Fig. 4c). The pH
of the mammalian ER has been measured to be 7.1 (ref. 16). Although
we do not know the pH of the *Plasmodium* ER and cannot reproduce
the cellular ionic environment in our *in vitro* assays, the measured
activity is consistent with an ER function. Activity of the PMV active
site mutant enzyme was undetectable (Fig. 4d). This result shows that
PMV itself is the active protease, not an associated protein. Others
have obtained active recombinant enzyme from *E. coli*[17], another
indication that PMV is the protease in question. PMV, but not control
preparations, cleaved a full-length PEXEL-containing proprotein
(Fig. 4e).

To confirm cleavage specificity, the products of the HRPII PEXEL
peptide incubation with isolated wild-type PMV enzyme were frac-
tionated by reverse phase HPLC (Fig. 5a) and analysed by mass
spectrometry (Fig. 5b). The fragments generated corresponded to
proteolysis after the leucine that is the *in vivo* processing site.
Similar results were obtained using the PfEMP2 peptide (Fig. 5c, d).

PMV interactions

We have shown that PMV is an essential ER protease in *Plasmodium
falciparum*. Residence in the ER is necessary for its function, as
deletion of the C-terminal tail had no effect on location or parasite
viability whereas deletion of the transmembrane (TM) region
rendered parasites non-viable (Fig. 1). Our data suggest that PMV is
the enzyme that processes PEXEL-containing proteins to send them
on their way for export into the host cell. It is not clear how PEXEL-
containing proteins are recognized by the translocon in the parasito-
phorous vacuole[18] when most of the PEXEL has been cleaved off by
PMV in the ER. It is conceivable that the propeptide stays associated
with the mature polypeptide during transport, but we favour a model
(Supplementary Fig. 1) in which chaperones associate with PMV in
the ER. On PEXEL cleavage, these chaperones receive the protein
destined for export, usher it through the secretory pathway and then
thread it through the translocon channel. In support of this, PMV
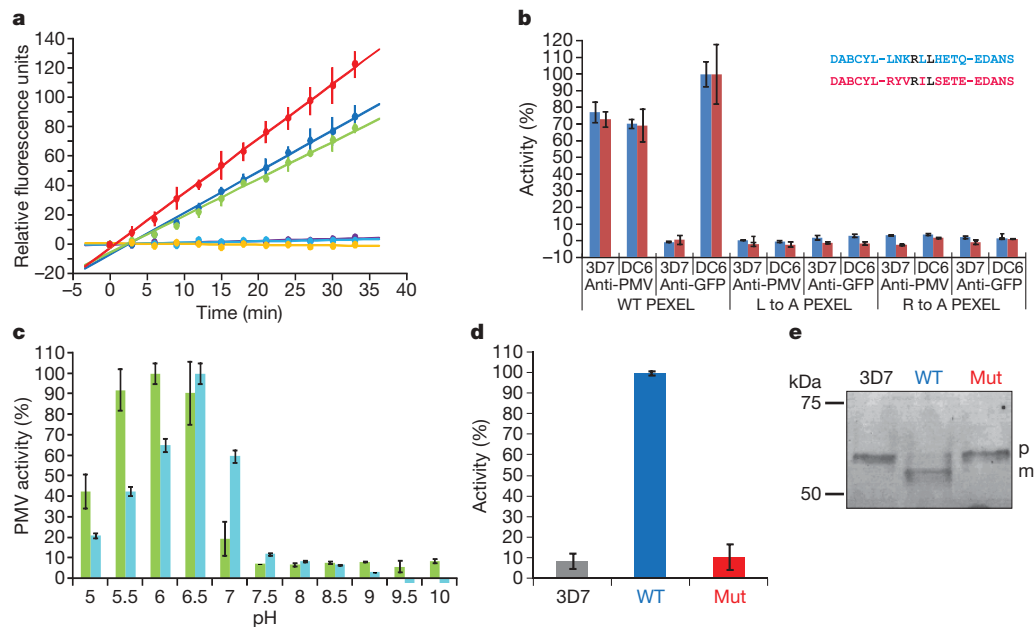pull-downs (Fig. 6) consistently identified an ER-resident HSP70 and

Figure 4 | Plasmepsin V activity. a, Substrate cleavage. PMV was isolated from parental (3D7) and PMV–GFP fusion (DC6) clones using anti-GFP (yellow and red, respectively) or anti-PMV (blue and green, respectively). A similar isolation without antibody (protein A) served as a control for non-specific binding (purple and cyan, respectively). Purified enzyme was incubated at pH 6.5 for 35 min with fluorogenic peptide (AnaSpec) corresponding to the PEXEL motif for HRPII (DABCYL-LNKRLLHETQ-EDANS). Error bars indicate standard deviation of three activity curves. **b**, PMV specificity for PEXEL motif. Processing of fluorogenic peptides containing the PEXEL motifs of HRPII (blue) or PfEMP2 (red, DABCYL-

RYVRILSETE-EDANS) are shown. Left, wild-type peptide (WT); middle, L to A mutant peptide; right, R to A mutant peptide. Inset: peptide sequences with mutated residues in black. Error bars as in **a**. **c**, pH dependence in Tris-malate buffer. Two separate experiments are shown in different colours. Error bars as in **a**. **d**, Activity of active-site mutant PMV compared to wild type. PMV–GFP was isolated from 3D7 (untransfected mock isolation), wild-type- or mutant-PMV–GFP-transfected parasites using anti-GFP for purification. Activity on HRPII–PEXEL peptide is normalized to PMV protein content. Error bars as in **a**. **e**, Cleavage of pro-HRPII by isolated enzyme from **d**. p, pro-form; m, mature protein.

HSP101 (a key translocon component[18]) as associated proteins. Much remains to be done to define the PMV-chaperone relationship but it is certainly plausible that chaperones could act in a complex or relay to shepherd proteins from ER to translocon for export.

Enzyme inhibition

In vitro enzyme activity was partially inhibited by high micromolar concentrations of HIV protease inhibitors or pepstatin A (Supplementary Fig. 3a) but not by other classes of inhibitors. We tested a panel of

protease inhibitors for ability to block processing of the PEXEL-containing exported protein HRPII but have not yet found a good inhibitor. BACE inhibitors had minimal effect, perhaps not surprising given the evolutionary distance between the two orthologues. Only HIV protease inhibitors had any effect and the blockade was partial (Supplementary Fig. 3b, c). Action on PMV is unlikely to be their primary effect because they kill cultured parasites in the single digit micromolar range[19,20], whereas effects on protein export and on isolated PMV were observed at 50–200 micromolar concentrations.
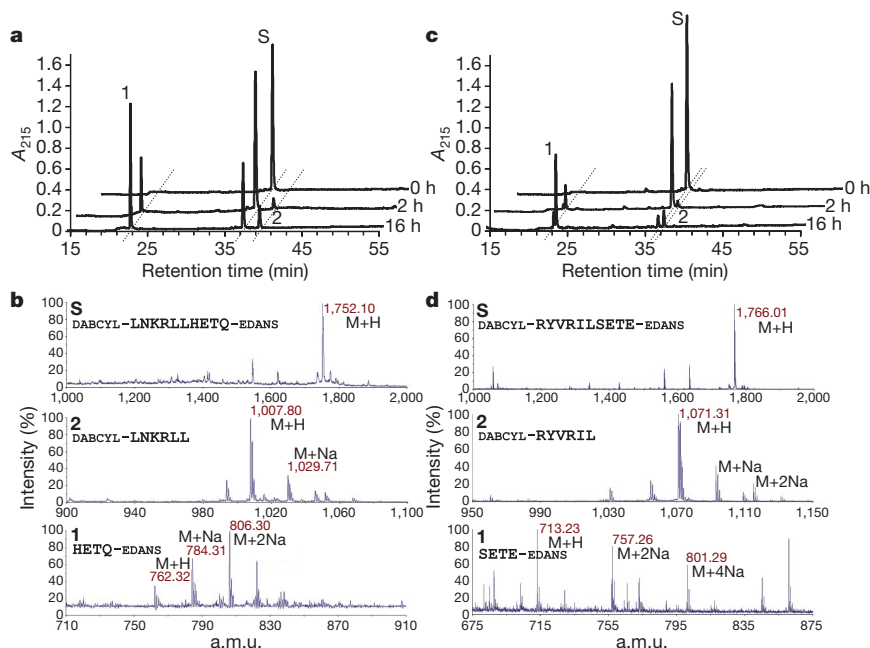
Figure 5 | Analysis of substrate cleavage.
a, c, PMV was incubated with HRPII (**a**) or PfEMP2 (**c**) PEXEL peptides (DABCYL-LNKRLLHETQ-EDANS and DABCYL-RYVRILSETE-EDANS, respectively). Cleavage products were separated on a C18 column by reverse-phase HPLC. Back to front: incubation for 0, 2 and 16 h. S, substrate peak. **b, d**, Isolated products and substrates from **a** and **c** were analysed by MALDI-TOF mass spectrometry. Ion peaks and sodium adducts are labelled. HRPII peptide product masses: calculated 762.51 and 1,008.25; detected 762.32 and 1,007.80. PfEMP2 peptide product masses: calculated 1,071.30 and 713.72; detected 1,071.31 and 713.23. Ions corresponding to alternative peptide bond cleavage were not detected. a.m.u., atomic mass unit.
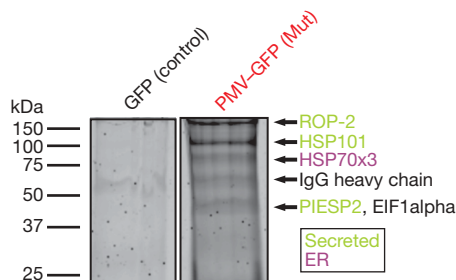
Figure 6 | Proteomic analysis of PMV-associated proteins. Coomassie-stained gel of an anti-GFP pull-down for the episomal GFP-expressing control strain (GFP) and the mutant PMV–GFP-expressing strain (Mut). Bands were excised, trypsinized and analysed by MS–MS. The same analysis was also performed on the wild-type PMV–GFP strain, and on the parental 3D7 strain but using anti-PMV antibody. Proteins for which at least four peptides were identified from the anti-PMV pull-down and three or more peptides from at least one of the two anti-GFP pull-downs of episomal PMV-expressing parasites are shown. All proteins identified are tabulated in Supplementary Table 1. None of the proteins were detected in the GFP control pull-down.

We propose that plasmepsin V is the PEXEL protease. This enzyme recognizes a simple RxL motif on secretory proteins destined for export into the host erythrocyte. Because PMV cleaves the PEXEL sequence away from the mature protein, the simplest conclusion is that PMV is primarily responsible for the specificity of export. An xE/Q/D dipeptide at the N terminus of mature exported proteins is also important for export though not for the cleavage itself[8]. Perhaps this polar residue comprises a secondary recognition element that interacts with the chaperone that will bring the protein to the translocon for export. It is very likely that the physical association of an escort system with PMV is needed to transfer the license for export. PMV seems then to be the gatekeeper for protein export. If potent inhibitors can be found, blocking the entire parasite virulence and intracellular survival program with one stroke will be a promising new strategy for combating this organism.

METHODS SUMMARY

Techniques for parasite culture, 3′-end integrations and truncations and their analysis, allelic replacement, site-directed mutagenesis, fluorescence imaging, parasite extraction and western blotting, as well as flow cytometry growth monitoring, have been described previously[12,21]. Parasite fluorescence intensity was measured blinded on random fields using Velocity 4 software (Improvision). For enzyme isolation, 50 ml of parasite culture at 2% haematocrit, 10% parasitaemia was harvested and parasites freed by saponin treatment as described[12]. Cells were solubilized for 30 min in 0.5% Triton X-100 in PBS buffer and incubated with anti-GFP (3E6; Invitrogen) or anti-PMV[9] antibodies for 1 h at 4 °C. Immune complexes were collected using protein A-Sepharose, washed extensively with PBS and isolated enzyme released in the activity buffer (see below) with 2 mM DTT. PMV enzyme activity assays were performed by incubation at 37 °C in 50 mM Tris-malate, pH 6.5, 50 mM NaCl, 0.05% Triton X-100 and reaction progress monitored on a Bio-Rad fluorimeter. Activity against pro-HRPII (ref. 22) was assessed after 16 h incubation by SDS–PAGE/anti-HRPII western blot. C18 reverse-phase liquid chromatography/MALDI mass spectrometry was performed as reported[23]. For pull-downs, enzyme was isolated as earlier except that solubilization was performed in RIPA buffer[21]. Immunoprecipitates were fractionated by SDS–PAGE and gel slices were analysed by MS–MS after trypsinization[24]. Primers for allelic replacement diagnosis were 241: AATTCCTAGGAGAAACTTTTAAGAAGATATTTTCT TTTTCTCTATATTC and 247: AATTCCTAGGAGAAACTTTTAAGAAGATATT TTCTTTTTCTCTATATTC. PMV sequences for 3′ fusion constructs were full-length, nucleotides 280 to 1770; tail deletion, nucleotides 280 to 1721; transmembrane region deletion, nucleotides 280 to 1623. Mutagenesis oligonucleotides for the allelic replacements: S, CGGCAAAGAATTTCTTTGATTCTAGACACAGGTT CATCTTCGTTAAGTTTCCCGTG; NS, CGGCAAAGAATTTCTTTGATTCTAGC GACAGGTTCATCTTCGTTAAGTTTCCCGTG.

1. Sargeant, T. J. *et al.* Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol.* **7**, R12 (2006).
2. van Ooij, C. *et al.* The malaria secretome: from algorithms to essential function in blood stage infection. *PLoS Pathog.* **4**, e1000084 (2008).
3. Haldar, K. & Mohandas, N. Erythrocyte remodeling by malaria parasites. *Curr. Opin. Hematol.* **14**, 203–209 (2007).
4. Maier, A. G., Cooke, B. M., Cowman, A. F. & Tilley, L. Malaria parasite proteins that remodel the host erythrocyte. *Nature Rev. Microbiol.* **7**, 341–354 (2009).
5. Hiller, N. L. *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934–1937 (2004).
6. Marti, M. *et al.* Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930–1933 (2004).
7. Chang, H. H. *et al.* N-terminal processing of proteins exported by malaria parasites. *Mol. Biochem. Parasitol.* **160**, 107–115 (2008).
8. Boddey, J. A., Moritz, R. L., Simpson, R. J. & Cowman, A. F. Role of the *Plasmodium* export element in trafficking parasite proteins to the infected erythrocyte. *Traffic* **10**, 285–299 (2009).
9. Klemba, M. & Goldberg, D. E. Characterization of plasmepsin V, a membrane-bound aspartic protease homolog in the endoplasmic reticulum of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **143**, 183–191 (2005).
10. Tang, J. in *Handbook of Proteolytic Enzymes* 2nd edn, Vol. 1 (eds Barrett, A. J., Rawlings, N. D. & Woessner, J. F.) Ch. 13 (Elsevier, 2004).
11. Bhattacharjee, S. *et al.* The malarial host-targeting signal is conserved in the Irish potato famine pathogen. *PLoS Pathog.* **2**, e50 (2006).
12. Russo, I., Oksman, A., Vaupel, B. & Goldberg, D. E. A calpain unique to alveolates is essential in *Plasmodium falciparum* and its knockdown reveals an involvement in pre-S-phase development. *Proc. Natl Acad. Sci. USA* **106**, 1554–1559 (2009).
13. Klemba, M., Gluzman, I. & Goldberg, D. E. A *Plasmodium falciparum* dipeptidyl aminopeptidase I participates in vacuolar hemoglobin degradation. *J. Biol. Chem.* **279**, 43000–43007 (2004).
14. Klemba, M., Beatty, W., Gluzman, I. & Goldberg, D. E. Trafficking of plasmepsin II to the food vacuole of the malaria parasite *Plasmodium falciparum*. *J. Cell Biol.* **164**, 47–56 (2004).
15. Maier, A. G. *et al.* Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes. *Cell* **134**, 48–61 (2008).
16. Kim, J. H. *et al.* Noninvasive measurement of the pH of the endoplasmic reticulum at rest and during calcium release. *Proc. Natl Acad. Sci. USA* **95**, 2997–3002 (1998).
17. Boddey, J. A. *et al.* An aspartyl protease directs malaria effector proteins to the host cell. *Nature* doi:10.1038/nature08728 (this issue).
18. de Koning-Ward, T. F. *et al.* A newly discovered protein export machine in malaria parasites. *Nature* **459**, 945–949 (2009).
19. Skinner-Adams, T. S. *et al.* Antiretrovirals as antimalarial agents. *J. Infect. Dis.* **190**, 1998–2000 (2004).
20. Parikh, S. *et al.* Antimalarial activity of human immunodeficiency virus type 1 protease inhibitors. *Antimicrob. Agents Chemother.* **49**, 2983–2985 (2005).
21. Russo, I., Oksman, A. & Goldberg, D. E. Fatty acid acylation regulates trafficking of the unusual *Plasmodium falciparum* calpain to the nucleolus. *Mol. Microbiol.* **72**, 229–245 (2009).
22. Sullivan, D. J., Gluzman, I. Y. & Goldberg, D. E. *Plasmodium* hemozoin formation mediated by histidine-rich proteins. *Science* **271**, 219–222 (1996).
23. Drew, M. E. *et al. Plasmodium* food vacuole plasmepsins are activated by falcipains. *J. Biol. Chem.* **283**, 12870–12876 (2008).
24. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).

Author Contributions I.R. designed and executed most of the experiments and wrote the paper; S.B., V.M., T.B. and A.O. designed and executed experiments; D.E.G. designed experiments and 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 D.E.G. (Goldberg@borcim.wustl.edu).

nature

LETTERS

A ground-based near-infrared emission spectrum of the exoplanet HD 189733b

Mark R. Swain[1], Pieter Deroo[1], Caitlin A. Griffith[2], Giovanna Tinetti[3], Azam Thatte[5], Gautam Vasisht[1], Pin Chen[1], Jeroen Bouwman[6], Ian J. Crossfield[4], Daniel Angerhausen[7], Cristina Afonso[6] & Thomas Henning[6]

Detection of molecules using infrared spectroscopy probes the conditions and compositions of exoplanet atmospheres. Water (H_2O), methane (CH_4), carbon dioxide (CO_2), and carbon monoxide (CO) have been detected[1–5] in two hot Jupiters. These previous results relied on space-based telescopes that do not provide spectroscopic capability in the 2.4–5.2 μm spectral region. Here we report ground-based observations of the dayside emission spectrum for HD 189733b between 2.0–2.4 μm and 3.1–4.1 μm, where we find a bright emission feature. Where overlap with space-based instruments exists, our results are in excellent agreement with previous measurements[2,6]. A feature at ~3.25 μm is unexpected and difficult to explain with models that assume local thermodynamic equilibrium (LTE) conditions at the 1 bar to 1×10^{-6} bar pressures typically sampled by infrared measurements. The most likely explanation for this feature is that it arises from non-LTE emission from CH_4, similar to what is seen in the atmospheres of planets in our own Solar System[7–9]. These results suggest that non-LTE effects may need to be considered when interpreting measurements of strongly irradiated exoplanets.

Spectroscopic detection of molecules in exoplanet atmospheres is a relatively recent development. High-precision spectroscopy data in combination with spectral retrieval techniques enables the determination of the atmospheric temperature profile and composition; this, in turn, provides some degree of knowledge about the atmospheric chemistry via molecular abundance ratios. Combinations of the molecules H_2O, CH_4, CO_2, and CO have been detected in the hot-Jupiter-type exoplanets HD 189733b and HD 209458b during primary eclipse[1,2] and/or secondary eclipse[3–5] events (when the planet passes either in front of or behind the stellar primary). In the case of HD 189733b, the primary eclipse spectrum probes the terminator region and the secondary eclipse spectrum probes the dayside region of the planet's atmosphere, thus permitting atmospheric conditions in these two regions to be compared. Temporal variability has also been detected in the mid-infrared emission spectrum[3] and may indicate that dynamical processes have a significant role in establishing the atmospheric properties[10,11]. Until now, the detection of molecules in exoplanet atmospheres has required the stability of the Hubble and Spitzer space telescopes. Observing exoplanet emission spectra from the ground is complicated by both the intrinsic variability of the Earth's atmosphere and instrument instability caused, for example, by changes in the instrument orientation as the telescope tracks.

The spectrum we present here is based on data obtained on 11 August 2007, using data from the SpeX instrument[12] on the NASA Infrared Telescope Facility (IRTF). The observations of the HD 189733 system were timed to observe the secondary eclipse light

curve, beginning approximately one hour before the onset of ingress and ending approximately one hour after the termination of egress. HD 189733 was observed at two positions (A and B) on the spectrograph entrance slit using an AB BA sequence. Spectral calibration was done using the internal instrument calibration arc lamps. To minimize the effects of seeing-based modulation of the measured flux density, a 1.6-arcsecond slit width was used. The SpeX instrument was configured to observe between 1.9 and 4.2 μm at an average spectral resolution of 470. The slit length for these observations was 15 arcseconds, and the sky background and source were measured simultaneously.

We developed a novel calibration method (described in the Supplementary Information) for systematic error removal and subsequent detection of the eclipse. This method is based on an iterative approach to removing systematic errors while the secondary eclipse is extracted by computing the self-coherent spectrum of groups of channels[13]. Although we have not reached the theoretical noise level and further improvements are possible, this approach is highly successful in detecting the secondary eclipse and, significantly, does not require the use of priors (for example, system ephemeris). We validated the calibration/spectral extraction method by comparing the ground-based results to previous space-based measurements of HD 189733b. Between 2.0 and 2.4 μm, the ground-based results clearly show the CO_2 absorption feature present in the exoplanet atmosphere and agree well with previous Hubble measurements[2] (see Fig. 1). In the 3.1–4.1 μm region, we compared the ground-based results with Spitzer photometry measurements[6], by integrating the ground-based spectrum over the Spitzer 3.6-μm filter pass band, and we find agreement at the 1σ level. Between 3.1 and 4.1 μm, the ground-based observations provide a unique capability, and it is here that we find emission features far exceeding those characteristic of other wavelengths in the HD 189733b dayside spectrum (see Fig. 2).

We investigated the plausibility of purely thermal emission by calculating the brightness-temperature spectrum; these values indicate the atmospheric temperatures probed by the light emitted at each wavelength when LTE conditions (collision-dominated) apply. The 3.25-μm flux density corresponds to a brightness temperature of 2,700 K, which markedly exceeds the typical brightness temperatures in HD 189733's spectrum (see Fig. 2 inset). Prior studies[3,6,11,14] and our own modelling successfully explained most of the infrared spectrum as thermal emission modulated by molecular opacity; these studies imply a temperature profile that decreases from roughly 1,600 K at 1 bar to roughly 800 K in the $\sim 10^{-3}$–10^{-5} bar range. However, these LTE models do not explain the flux density of 3.1–4.1 μm; modification of these models to include a tropopause and an appropriate absorber could explain the weak emission around

[1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109-8099, USA. [2]University of Arizona, Lunar and Planetary Laboratory, Space Science Building Room 525, 1629 East University Boulevard, Tucson, Arizona 85721, USA. [3]Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT, London, UK. [4]Astronomy Department UCLA, 475 Portola Plaza, Los Angeles, California 90034, USA. [5]Woodruff School of Mechanical Engineering, MRDC Building, Room 4111, Georgia Institute of Technology, Atlanta, Georgia 30332-0405, USA. [6]Max-Planck Institute for Astronomy, Koenigstuhl 17, D-69117, Heidelberg, Germany. [7]German SOFIA Institute, Institute for Space Systems, Pfaffenwaldring 3170569, Stuttgart, Germany.
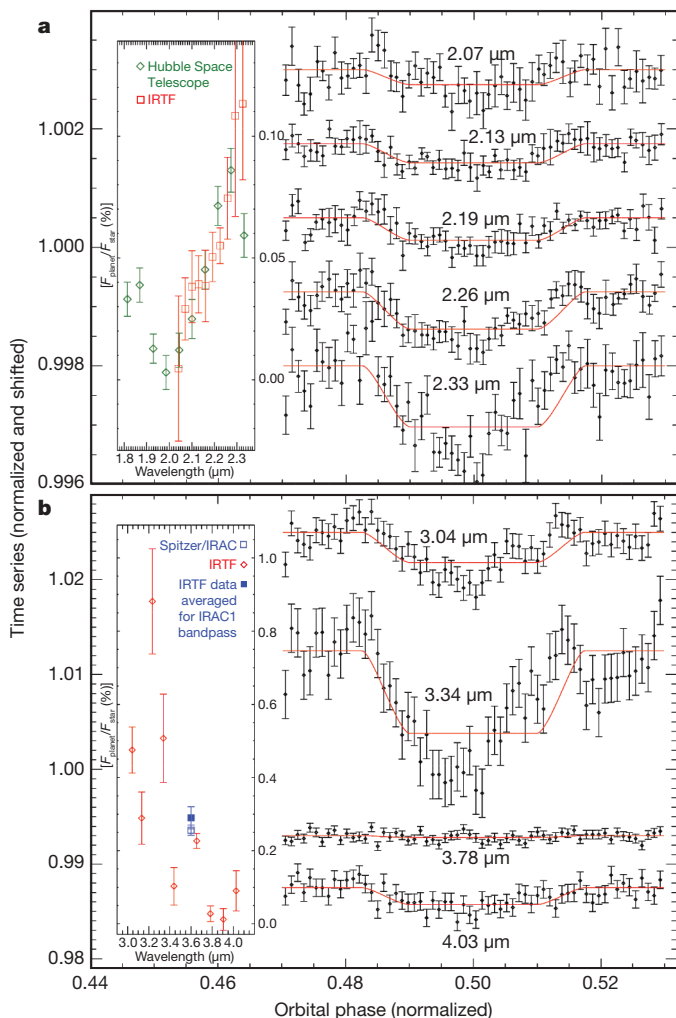
Figure 1 | Dayside spectra and secondary eclipse light curves. The ground-based spectra are in excellent agreement with space-based measurements. **a**, Comparison of the IRTF measurements (red) with previous measurements using the Hubble Space Telescope (green). **b**, Comparison of the IRTF spectrum (red) and IRTF data averaged to the Spitzer 3.6-μm pass band (blue solid square) with the Spitzer Space Telescope 3.6-μm photometry measurement (blue open square). IRAC, the InfraRed Array Camera. Each calibrated light curve includes the averaged measurements (black diamonds), and the best-fit eclipse model (red line). In the comparison between ground-based and space-based measurements, the secondary eclipse depth is shown as the planet/star contrast ratio. All data are shown ±1σ errors.

3.8 μm. Applying LTE models to explain the 3.25-μm emission requires a temperature structure and CH_4 abundance that is incompatible with all other portions of the dayside spectrum; the inclusion of a 'hot layer' of CH_4 makes it difficult to explain the spectrum without causing the brightness temperature to increase around 2.2 and 7.8 μm (the Octad and ν_4 band respectively). Thus, we conclude that the dayside emission at between 3.1 and 3.4 μm is probably caused by a non-LTE emission process.

In our own Solar System, CH_4 fluorescence has been detected in Jupiter[7,8], Saturn[7,8], and Titan[9]. The 3.1-μm and 3.25-μm features in the HD 189733b dayside spectrum resemble those of Titan, where the fluorescence of the ν_3 bands of CH_4 and HCN at pressures below 10^{-4} bar cause emission far exceeding that possible in LTE conditions. The absence of a 3.9-μm feature suggests that H_3^+ emission is not significant in HD 189733b. When taken together, (1) the approximate alignment between the HD 189733b spectrum and the peaks in the observed fluorescence spectrum of Titan, (2) the difficulty of explaining the 3.1–3.4-μm band emission in HD 189733b using LTE
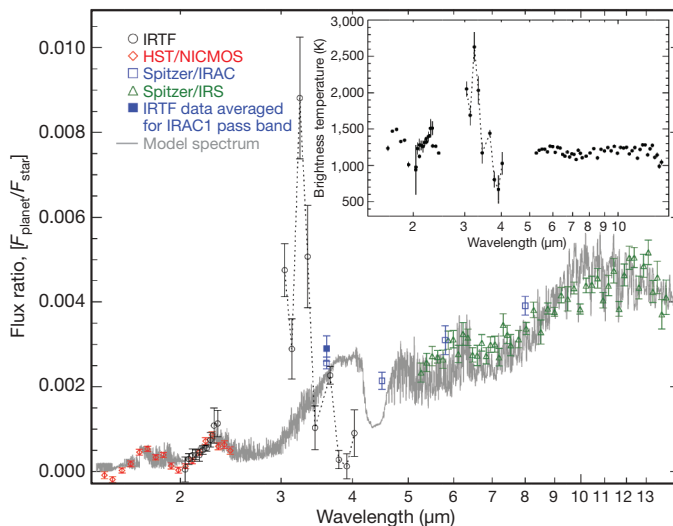
Figure 2 | Unexpectedly strong 3.25-μm emission present in the dayside spectrum. The brightness temperature of the 3.25-μm emission feature indicates the likely presence of a non-LTE emission mechanism. The dayside emission spectrum is based on the new measurements reported in this paper (black), together with previous results from Hubble spectroscopy (red), Spitzer spectroscopy (green), and Spitzer photometry (blue); all data are shown ±1σ errors. A radiative transfer model (grey) assuming LTE conditions and consistent with the measurements made with the Spitzer and Hubble space telescopes fails to describe the emission structure at 3.1–4.1 μm, and we find no plausible combination of atmospheric parameters that provides a good model of the observations under LTE conditions. The inset plot displays the brightness temperature at each wavelength and shows the large temperature change needed to produce the 3.25-μm emission if LTE conditions hold. IRS, the Infrared Spectrograph; NICMOS, the Near Infrared Camera and Multi-Object Spectrometer. The dotted lines are a visual aid to help draw the eye to the new IRTF data.

models, and (3) the prevalence of fluorescence in our own Solar System, all indicate that non-LTE CH_4 emission is probably present in the atmosphere of HD 189733b. A puzzling aspect of these data is the absence of strong emission from the CH_4 ν_3 band P branch; a similar effect has been detected in the upper atmosphere of Titan (see Supplementary Information) and remains unexplained.

The 3.1–4.1-μm spectrum shows that the strong emission from the CH_4 ν_3 band dominates the 3.6-μm Spitzer photometry measurement of this planet (see Fig. 1). This is significant because previous models, based on the assumption of LTE, systematically underpredict the 3.6-μm Spitzer flux density measurement[3,6,11,14]. The underprediction is now easily explained by the presence of strong non-LTE emission in the filter pass band. Non-LTE emission may also explain another puzzling aspect of HD 189733b in which the onset of secondary eclipse in the Spitzer 3.6-μm measurements is delayed by 5.6 ± 0.8 min relative to other Spitzer photometric bands[6]; the authors of ref. 6 speculated that this delay "may arise from strong brightness variations across the visible hemisphere of the planet". Given that the CH_4 ν_3 band lies within the Spitzer 3.6-μm filter band pass and that emission at other wavelengths in the filter band pass is weak (see Fig. 1), a significant non-uniformity of the non-LTE emission in HD 189733b could produce the observed delay in the onset of secondary eclipse. Possible sources for spatially inhomogeneous non-LTE emission include localized vertical transport of CH_4 from lower levels of the atmosphere or the presence of clouds.

Of equal significance to the discovery of non-LTE emission in an exoplanet atmosphere is the ground-based detection of molecules in an exoplanet atmosphere. We believe the calibration method used here can be applied directly to many existing instruments. The IRTF is not considered to be a 'large' telescope, and the SpeX instrument has no specific optimization for high-dynamic-range spectroscopy; this suggests that many facilities could make similar measurements

and that improvements could be realized by purpose-built instrumentation on large telescopes. Given the availability of numerous large telescopes equipped with infrared spectrometers, the result in this paper foreshadows a large quantity of 'molecular-abundance-grade' exoplanet spectra. Although telluric opacity limits ground-based observations to certain spectral regions, these results, together with recent optical detections of atomic features[15,16], decisively show that ground-based spectroscopy will have a powerful and lasting influence in the emerging field of exoplanet characterization.

Hot-Jupiter-type planets experience powerful radiation forcing caused by close proximity to the stellar primary and, in the cases of HD 189733b and HD 209458b, are thought to be tidally locked as well. No analogue for these planets exists in our own Solar System, and thus this class of object represents an opportunity to study planetary atmospheres in a completely new regime. Detection of molecules via infrared spectroscopy is currently probing the conditions, composition and chemistry of HD 189733b and HD 209458b and, by extension, other members of the hot-Jupiter exoplanet class. Previous observations have shown both relatively high CO_2 and CH_4 abundances[4,5] and temporal variability[3]; the former indicates the possible role of non-equilibrium chemistry, while the latter indicates the possible presence of dynamical effects. The results in this paper show that interpreting hot-Jupiter spectra by assuming only LTE conditions is questionable. Thus, the simple picture of a time-stationary exoplanet atmosphere governed by LTE conditions and dominated by equilibrium chemistry is being challenged. The observations reported here are the latest in a series of results supporting the view that the highly irradiated gas giant planets exhibit a degree of complexity, together with a rich variety of physical and chemical processes, that we have only begun to understand.

1. Tinetti, G. *et al.* Water vapour in the atmosphere of a transiting extrasolar planet. *Nature* **448,** 169–171 (2007).
2. Swain, M. R., Vasisht, G. & Tinetti, G. The presence of methane in the atmosphere of an extrasolar planet. *Nature* **452,** 329–331 (2008).
3. Grillmair, C. J. *et al.* Strong water absorption in the dayside emission spectrum of the planet HD 189733b. *Nature* **456,** 767–768 (2008).
4. Swain, M. R. *et al.* Molecular signatures in the near-infrared dayside spectrum of HD 189733b. *Astrophys. J.* **690,** L114–L117 (2009).
5. Swain, M. R. *et al.* Water, methane, and carbon dioxide present in the dayside spectrum of the exoplanet HD 209458b. *Astrophys. J.* **704,** 1616–1221 (2009).
6. Charbonneau, D. *et al.* The broadband infrared emission spectrum of the exoplanet HD 189733b. *Astrophys. J.* **686,** 1341–1348 (2008).
7. Drossart, P. *et al.* Fluorescence in the 3 microns bands of methane on Jupiter and Saturn from ISO/SWS observations. *ESA SP* **427,** 169–172 (1999).
8. Brown, R. H. *et al.* Observations with the Visual and Infrared Mapping Spectrometer (VIMS) during Cassini's flyby of Jupiter. *Icarus* **164,** 461–470 (2003).
9. Kim, S. J., Geballe, T. R. & Noll, K. S. Three-micrometer CH_4 line emission from Titan's high-altitude atmosphere. *Icarus* **147,** 588–591 (2000).
10. Rauscher, E. & Menou, K. Atmospheric circulation in hot Jupiters: a shallow three-dimensional model. *Astrophys. J.* **700,** 887–897 (2009).
11. Showman, A. P., Cooper, C. S., Fortney, J. J. & Marley, M. S. Atmospheric circulation of hot Jupiters: three-dimensional circulation models of HD 209458b and HD 189733b with simplified forcing. *Astrophys. J.* **682,** 559–576 (2008).
12. Rayner, J. T. *et al.* SpeX: a medium-resolution 0.8–5.5 micron spectrograph and imager for the NASA infrared telescope. *Publ. Astron. Soc. Pacif.* **115,** 362–382 (2003).
13. Pagiatakis, S. D., Yin, H. & Abd El-Gelil, M. Least-squares self-coherency analysis of superconducting gravimeter records in search for the Slichter triplet. *Phys. Earth Planet. Inter.* **160,** 108–123 (2007).
14. Barman, T. B. On the presence of water and global circulation in the transiting planet HD 189733b. *Astrophys. J.* **676,** L61–L64 (2008).
15. Redfield, S., Endl, M., Cochran, W. D. & Koesterke, L. Sodium absorption from the exoplanetary atmosphere of HD 189733b detected in the optical transmission spectrum. *Astrophys. J.* **673,** L87–L90 (2008).
16. Snellen, I. A. G., Albrecht, S., de Mooij, E. J. W. & Lee Poole, R. S. Ground-based detection of sodium in the transmission spectrum of exoplanet HD 209458b. *Astron. Astrophys.* **487,** 357–362 (2008).

Author Contributions M.R.S. wrote the paper, participated in developing the calibration methods, and wrote the telescope proposals. P.D. worked on developing and testing the calibration method. C.A.G. and G.T. worked on the interpretation of the results. A.T. worked on the calibration method. G.V. and P.C. worked on calibration validation. J.B. worked on the calibration of the Spitzer data. I.J.C. worked on an early version of the calibration method. D.A. contributed to the atmospheric calibration approach. C.A. and T.H. were co-authors on the original telescope proposal and provided comments on the paper text.

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 M.R.S. (mark.r.swain@jpl.nasa.gov).

LETTERS

Directional water collection on wetted spider silk

Yongmei Zheng[1]*, Hao Bai[2]*, Zhongbing Huang[3], Xuelin Tian[3], Fu-Qiang Nie[3], Yong Zhao[3], Jin Zhai[1] & Lei Jiang[3]

Many biological surfaces in both the plant and animal kingdom possess unusual structural features at the micro- and nanometre-scale that control their interaction with water and hence wettability[1–5]. An intriguing example is provided by desert beetles, which use micrometre-sized patterns of hydrophobic and hydrophilic regions on their backs to capture water from humid air[6]. As anyone who has admired spider webs adorned with dew drops will appreciate, spider silk is also capable of efficiently collecting water from air. Here we show that the water-collecting ability of the capture silk of the cribellate spider _Uloborus walckenaerius_ is the result of a unique fibre structure that forms after wetting, with the 'wet-rebuilt' fibres characterized by periodic spindle-knots made of random nanofibrils and separated by joints made of aligned nanofibrils. These structural features result in a surface energy gradient between the spindle-knots and the joints and also in a difference in Laplace pressure, with both factors acting together to achieve continuous condensation and directional collection of water drops around spindle-knots. Submillimetre-sized liquid drops have been driven by surface energy gradients[7–9] or a difference in Laplace pressure[10], but until now neither force on its own has been used to overcome the larger hysteresis effects that make the movement of micrometre-sized drops more difficult. By tapping into both driving forces, spider silk achieves this task. Inspired by this finding, we designed artificial fibres that mimic the structural features of silk and exhibit its directional water-collecting ability.

Spider silk, which is composed of humidity-sensitive[11–16] hydrophilic flagelliform proteins[17,18], enjoys a high reputation as a fibre with excellent mechanical properties[11,13–15,17–20]. Another intriguing but less studied feature is its ability to collect water from humid air (see Supplementary Fig. 1). This is seen in webs built by cribellate spiders, which use a cribellum—a comb-like device—to separate silk fibres drawn from their spinnerets into many exceedingly fine fibres. The environmental scanning electronic microscopy (SEM) images of the spider silk in Fig. 1 illustrate its structure. Puffs composed of nanofibrils are spaced along two main-axis fibres with a periodicity of $85.6 \pm 5.1\,\mu\mathrm{m}$ (Fig. 1a)[20]. The puffs' diameter is $130.8 \pm 11.1\,\mu\mathrm{m}$, and they are separated by joints with a diameter of $41.6 \pm 8.3\,\mu\mathrm{m}$. The zoomed-in image in Fig. 1b shows the puffs to be composed of random nanofibrils (20–30 nm in diameter). These highly hydrophilic nanofibrils[17,18] enhance the wettability of spider silk, which is favourable for condensing water drops.

When dry spider silk is placed in fog, its structure changes as water starts to condense and form drops that move along the silk fibre (Fig. 2). At the initial stage, tiny water drops (black dots indicated by arrows in Fig. 2a) condense on the semitransparent puffs. As water condensation continues, the puffs shrink into opaque bumps (Fig. 2b, c) and finally form periodic spindle-knots (Fig. 2d). The impact of fog on the mechanical properties of spider silk has been noted before[15,16], and the present observations indicate that in the case of

cribellate spider silk, the material changes its fibre structure as a result of wetting. After this 'structural wet-rebuilding', directional water collection starts.

To analyse this process, images of the wet-rebuilt spider silk are divided into four regions that each contain one joint (that is, regions I, II, III and IV in Fig. 2d–f). When exposed to mist, small water drops (identified by numbers 1 to 10 in Fig. 2e) randomly condense on joints (drops 1, 3, 4, 5, 8 and 10) and on spindle-knots (drops 2, 6, 7 and 9). As water condensation continues, the drops grow in size and those on joints move directionally to the nearest spindle-knots (white arrows in Fig. 2e) where they coalesce to form larger water drops (Fig. 2f). Drops 1–5, which formed in regions I and II, coalesced into
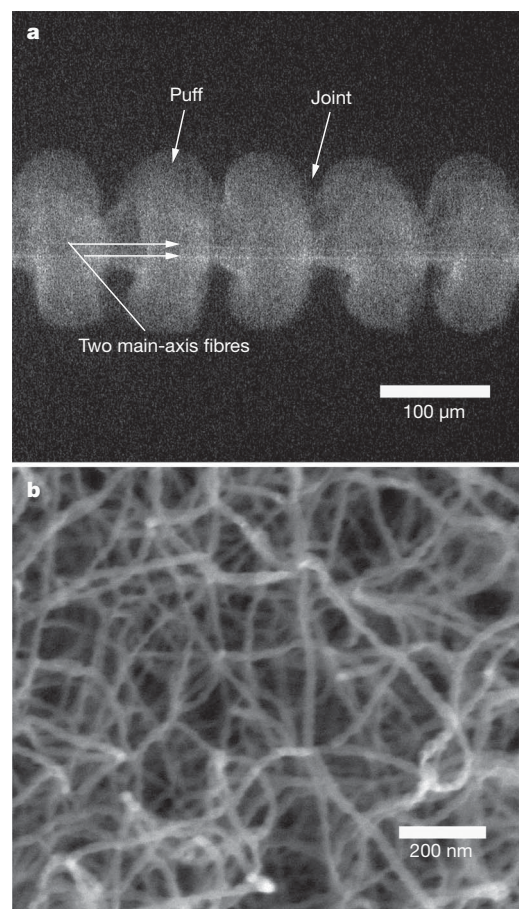
Figure 1 | Structures of dry capture silk of cribellate spider. a, Low-magnification environmental SEM image of periodic puffs and joints surrounding two main-axis fibres. **b,** Magnified image of puff composed of countless nanofibrils.

[1]School of Chemistry and Environment, Beijing University of Aeronautics and Astronautics, Beijing 100191, China. [2]National Center for Nanoscience and Technology, Beijing 100190, China. [3]Beijing National Laboratory for Molecular Sciences (BNLMS), Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, China. *These authors contributed equally to this work.
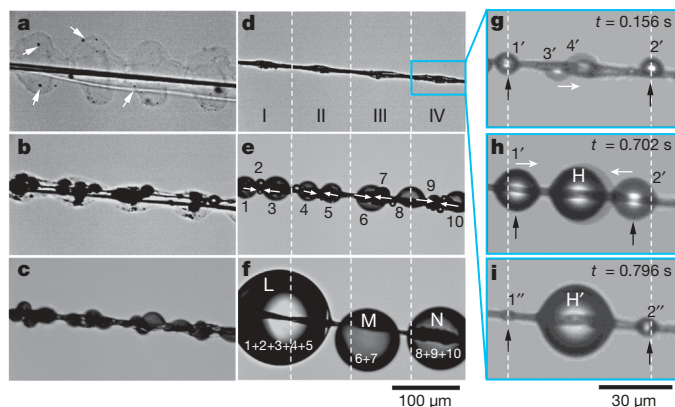
Figure 2 | *In situ* optical microscopic observation of directional water collection on spider silk in mist. a, Dry spider silk is of four semitransparent puffs. In mist, few tiny water drops (indicated by arrows) first condense on puffs. **b–d,** Puffs shrink to opaque bumps with water condensing and form spindle-knots linked by joint. **d–f,** The directional water collection on the wet-rebuilt multistructural spider silk. The images are divided into four regions (I, II, III and IV). **e,** Smaller water drops condense on spider silk (denoted 1−10). Water drops move directionally from joints to spindle-knots (as indicated by arrows) with volume increasing. **f,** The growing water drops 1−5 coalesce to a larger water drop L covering two spindle-knots (I and II), while water drops 6, 7 and 8−10 coalesce to two medium water drops: M covering single spindle-knot III and N covering single spindle-knot IV. **g–i,** More detailed directional drop movement on individual spindle-knot. Smaller water drops (1′, 2′, 3′ and 4′) first randomly condense on the spindle-knot and the joints at 0.156 s (**g**). With volume increasing, droplets 3′ and 4′ coalesce into water drop H on the spindle-knot at 0.702 s, while water droplets 1′ and 2′ spontaneously move from joints to spindle-knot (as indicated by white arrows) (**h**) and coalesce to a larger water drop H′ on the spindle-knot at 0.796 s (**i**). Then the joints refresh and a new directional water collection cycle starts (two tiny water drops 1′′ and 2′′ recondense on the joints) (**i**). The black arrows indicate the condensing sites of joints that are favourable for continuous directional water collecting.

a larger drop L that covers two spindle-knots, while drops 6 and 7 in region III and drops 8–10 in region IV coalesced to form medium-sized drops M and N, each covering a single spindle-knot. Overall, these observations imply that the periodic fibre structure of spindle-knots and joints found in wetted spider silk plays a part in directional water collection.

We further monitored water collection by focusing on an individual spindle-knot of the spider silk fibre (Fig. 2g–i). In the initial stage, small water drops 1′, 2′, 3′ and 4′ randomly condensed on the spindle-knot and the two adjacent joints (Fig. 2g). The growing drop 3′ then moved to the spindle-knot (white arrow in Fig. 2g) where it encountered drop 4′ and coalesced to form drop H (Fig. 2h). Meanwhile, the growing drops 1′ and 2′ located on the joints also moved towards the spindle-knot (white arrows in Fig. 2h) and coalesced with the already-present drop H to form a larger drop H′ (Fig. 2i). A rather large water drop finally formed on the spindle-knot through sequential coalescence of smaller drops originating from the joints.

In this process, the spindle-knot serves in the initial stage as a condensing site and then as a drop collecting site for the coalescence of smaller drops that originate from the joints. In contrast, the joints mainly act as condensing sites (black arrows in Fig. 2g–i), where water condenses into drops that are then transported to the spindle-knot. Importantly, after water drops have left the joints and been collected on the spindle-knot, a new cycle of water condensation and directional drop movement can start on the joints. This is seen in Fig. 2i, in which drops 1′′ and 2′′ condense on the joints after drops 1′ and 2′ have formed and moved to the spindle-knot. Taken together, these images illustrate how the cooperation between joints acting as condensing sites and spindle-knots acting mainly as collecting sites enables uninterrupted directional water

collection (see also Supplementary Fig. 2 and Supplementary Movie 1). We observed such directional water collection behaviour only with wetted silk fibres (that is, wet-rebuilt silk) from the cribellate spider *Uloborus walckenaerius*; in contrast, silkworm silk and nylon fibres with a uniform structure did not exhibit the directional water collection phenomenon (Supplementary Figs 3, 4).

To explore in more detail the role of fibre structure in directional water collection behaviour, we show, in Fig. 3, environmental SEM images of the wet-rebuilt spider silk. The wetted spider silk is composed of alternate spindle-knots and joints (apex angles $2\beta \approx 19°$) with periodicity of $89.3 \pm 13.5\,\mu m$ (Fig. 3a). The diameters of spindle-knots and joints are $21.0 \pm 2.7\,\mu m$ and $5.9 \pm 1.2\,\mu m$, respectively. Magnified images of a spindle-knot (Fig. 3b, c) reveal highly random nanofibrils that give a rough surface topography, while comparable images of a joint (Fig. 3d, e) show that it is composed of nanofibrils that run relatively parallel to the silk fibre axis and form an anisotropic aligned and relatively smooth topography. These structural features can give rise to a surface energy gradient and a difference in Laplace pressure, which can both act as a driving force for the directional movement of water drops as detailed below (see also Fig. 4).

Surface energy gradients can arise from differences in either surface chemical composition[7,8] or surface roughness[21,22], and such gradients will drive water drops towards the more wettable region with a higher surface energy. According to Wenzel's law[23]:

$$\cos \theta_w = r \, \cos \theta \qquad (1)$$

where r is surface roughness, θ_w and θ are the apparent and intrinsic contact angles on rough and smooth surfaces, respectively. As for hydrophilic spider silk[17,18], its chemical composition does not change much along the fibre, but the joint has a smaller axial-parallel roughness and hence larger water contact angle than the spindle-knot
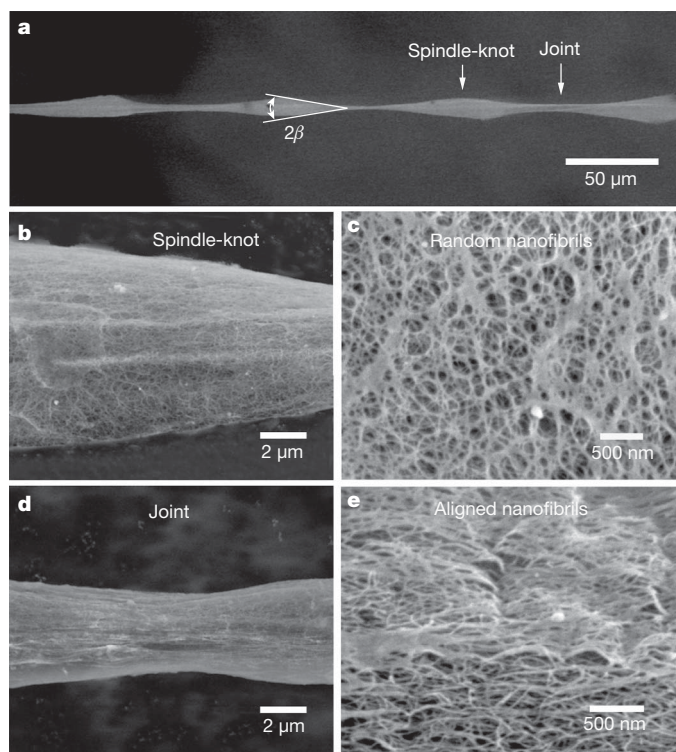

Figure 3 | Structure of wet-rebuilt spider silk. a, Environmental SEM images of periodic spindle-knots linking with slender joints. The apex angle of spindle-knots (2β) is about 19°. Low-magnification (**b**) and zoomed (**c**) images show that the spindle-knot is randomly interweaved by nanofibrils. **d, e,** Low-magnification (**d**) and high-magnification (**e**) images of the joint, which is composed of nanofibrils aligned relatively parallel to the silk axis.
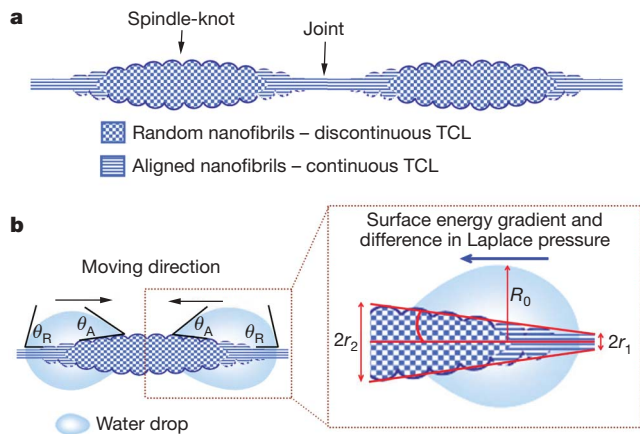
Figure 4 | Mechanism of directional water collection on wet-rebuilt spider silk. a, Spindle-knots are interweaved by highly random nanofibrils, while joints are composed of relatively aligned nanofibrils. Spindle-knots are of larger silk-axial roughness than joints. Meanwhile, random surface topography of spindle-knots forms discontinuous TCL, while aligned surface topography of joints forms continuous TCL for water drops, which is helpful for water drops' movement along joints. **b**, Surface structural anisotropy generates a surface energy gradient so that spindle-knots possess higher apparent surface energy than joints. At the same time, the conical shape of the spindle-knot generates a difference in Laplace pressure from the high-curvature region (joint) to the low-curvature region (spindle-knot), as the inset shows. The cooperation of these two factors drives water drops from joint to spindle-knot. The arrows denote the directions of drop movement.

(Fig. 3c, e). That is, the spindle-knot is more hydrophilic and has a higher apparent surface energy than the joint[24] (see also Fig. 4b). The force generated by a surface energy gradient that arises from a difference in surface roughness is given by[7,8]:

$$F = \int_{L_j}^{L_k} \gamma \left(\cos \theta_A - \cos \theta_R \right) \mathrm{d}l \qquad (2)$$

where γ is the surface tension of water, θ_A and θ_R are the advancing and receding angles of water drop on spider silk ($\theta_A < \theta_R$, according to equation (1)), respectively, and $\mathrm{d}l$ is the integrating variable along the length from the joint (L_j) to the spindle-knot (L_k). The surface energy gradient arising from differences in roughness will thus drive water drops to move from the less hydrophilic region (joint with relative lower surface energy) to the more hydrophilic region (spindle-knot with high surface energy) (Fig. 4b).

The second possible driving force for directional water drop movement arises from the spindle-shaped geometry of the knots, which will generate a difference in Laplace pressure. As illustrated in Fig. 4b, a spindle-knot can be thought of as two oppositely curved and joined conical objects (inset of Fig. 4b). Such a conical shape with a curvature gradient will give rise to a difference in Laplace pressure (ΔP) acting on a water drop[10]:

$$\Delta P = -\int_{r_1}^{r_2} \frac{2\gamma}{(r + R_0)^2} \sin\beta \, \mathrm{d}z \qquad (3)$$

where r is the local radius, R_0 is the drop radius ($R_0 = (3V/4\pi)^{1/3}$, with V the drop volume), β is the half apex-angle of the spindle-knot, and z is the integrating variable along the diameter of the spindle-knot. The Laplace pressure on the high curvature site (the joint with local radius r_1) is larger than that on the low-curvature site (the spindle-knot with local radius r_2) because r_1 is smaller than r_2, and the resultant non-equilibrium Laplace pressure difference within a water drop will propel the drop to move from the joint to the spindle-knot. The overall result is that the surface energy gradient arising from the anisotropic surface structures and the difference in Laplace pressure arising from the conical spindle-knot geometry

act cooperatively to drive condensing and growing water drops from the joint to the spindle-knot (Fig. 4b).

We observe water drops with diameters of 5–20 μm move until they coalesce and are collected on a spindle-knot (Supplementary Fig. 2). We note that the movement of sub-millimetre-sized liquid drops driven by either surface energy gradients[7–9] or by differences in Laplace pressure[10] has been realized before. But either driving force on its own has not yet been shown to move smaller drops with diameters on the micrometre scale, where contact hysteresis effects are increasingly important (see also Supplementary Discussion regarding minimum drop size). This makes it particularly interesting that the unique structural features of wet-rebuilt spider silk enable forces arising from Laplace pressure differences and from surface energy gradients to combine so that hysteresis effects can be overcome and micrometre-sized water drops moved. Once the anisotropic structural features are damaged, spider silk fibres are no longer capable of directional water collection (Supplementary Figs 5–8).

In addition to enabling cooperation between two different driving forces, the structure of wet-rebuilt spider silk also optimizes hysteresis effects so as to favour the directional movement of water drops from joints to spindle-knots. These two regions have the same chemical composition, but different surface topographies that affect the spreading and movement of water drops. More specifically, a water drop will spread or move more readily along the parallel direction of an aligned surface topography than along a surface with a randomly rough topography[25–27]. This is because the vapour–liquid–solid three-phase contact line (TCL) is continuous along the parallel direction of the surface with aligned topography and discontinuous on the randomly rough topography, with a continuous TCL allowing smooth spreading or moving of a liquid whereas a discontinuous TCL gives rise to more pronounced hysteresis effects[1,25–27]. The joints in wet-rebuilt spider silk are comprised of relatively aligned nanofibrils that give rise to a fairly continuous TCL along the silk fibre axis, whereas the spindle-knots are comprised of random nanofibrils that cause the TCL to be discontinuous. Water drops moving along joints therefore experience
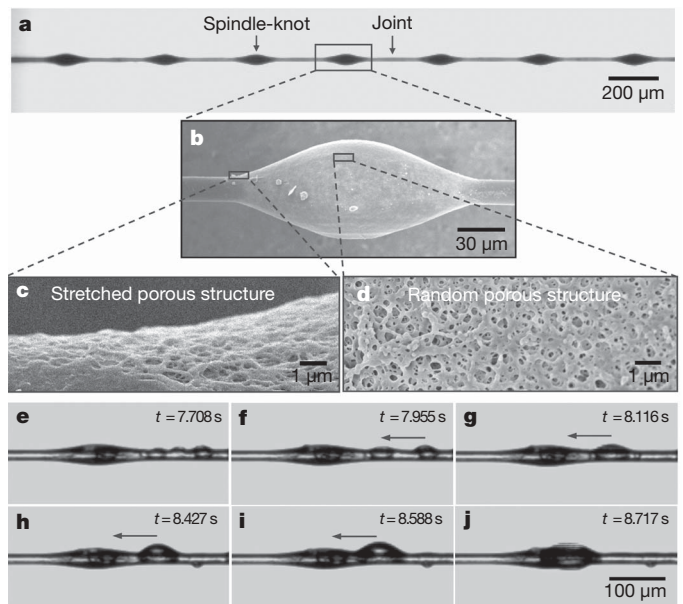

Figure 5 | Artificial spider silk that mimics the structure and water collection capability of natural spider silk. a, Optical image of spindle-knot/joint structure with periodicity of 394.6 ± 16.1 μm. **b–d**, SEM images of a spindle-knot (**b**), a stretched porous structure on the joint (**c**) and a random porous structure on the spindle-knot (**d**). **e–j**, Directional water collection on artificial spider silk. When the artificial spider silk is in mist at 0 s, tiny water drops randomly condense on the artificial spider silk at 7.708 s (**e**), and then directionally move from joint to spindle-knot with the volume increasing from 7.955 s to 8.717 s (**f–j**).

less hysteresis than those moving along spindle-knots, and this difference will further aid the movement of water drops from the joint to the spindle-knot.

Guided by the detailed mechanistic insights into directional water collection on spider silk gained in this study, we designed and fabricated artificial fibres that mimic the structural features of wet-rebuilt spider silk (see Supplementary Information for details). The optical image of this artificial spider silk (Fig. 5a) shows periodic spindle-knots that resemble those of the wetted spider silk and have a periodicity of 394.6 ± 16.1 μm. Spindle-knots and joints have diameters of 43.7 ± 5.4 μm and 13.5 ± 0.7 μm (Fig. 5b), respectively, and their microstructures are similar to those seen in wetted spider silk: the artificial joints have a stretched porous structure mimicking aligned nanofibrils (Fig. 5c), and the artificial spindle-knots exhibit random porous surface structures (Fig. 5d). The functional performance of the material is illustrated by the series of images in Fig. 5e–j. These show that when exposed to mist, small drops start to condense randomly on the artificial spider silk (Fig. 5e); and as the drop volume increases (Fig. 5f), water drops located on joints move towards spindle-knots (Fig. 5g–j). These observations clearly show that our artificial spider silk not only mimics the structure of wet-rebuilt spider silk but also its directional water collection capability. We therefore anticipate that the design principles uncovered and implemented in this study will aid the development of functional fibres for use in water collection and in liquid aerosols filtering in manufacturing processes.

1. Feng, L. *et al.* Super-hydrophobic surfaces: from natural to artificial. *Adv. Mater.* **14,** 1857–1860 (2002).
2. Gao, X. F. & Jiang, L. Water-repellent legs of water striders. *Nature* **432,** 36 (2004).
3. Blossey, R. Self-cleaning surfaces—virtual realities. *Nature Mater.* **2,** 301–306 (2003).
4. Sun, T. L., Feng, L., Gao, X. F. & Jiang, L. Bioinspired surfaces with special wettability. *Acc. Chem. Res.* **38,** 644–652 (2005).
5. Li, X. M., Reinhoudt, D. & Crego-Calama, M. What do we need for a superhydrophobic surface? A review on the recent progress in the preparation of superhydrophobic surfaces. *Chem. Soc. Rev.* **36,** 1350–1368 (2007).
6. Parker, A. R. & Lawrence, C. R. Water capture by a desert beetle. *Nature* **414,** 33–34 (2001).
7. Chaudhury, M. K. & Whitesides, G. M. How to make water run uphill. *Science* **256,** 1539–1541 (1992).
8. Daniel, S., Chaudhury, M. K. & Chen, J. C. Fast drop movements resulting from the phase change on a gradient surface. *Science* **291,** 633–636 (2001).
9. Daniel, S., Sircar, S., Gliem, J. & Chaudhury, M. K. Ratcheting motion of liquid drops on gradient surfaces. *Langmuir* **20,** 4085–4092 (2004).
10. Lorenceau, É. & Quéré, D. Drops on a conical wire. *J. Fluid Mech.* **510,** 29–45 (2004).
11. Vollrath, F. & Porter, D. Spider silk as archetypal protein elastomer. *Soft Matter* **2,** 377–385 (2006).
12. Vollrath, F. *et al.* Compounds in the droplets of the orb spider's viscid spiral. *Nature* **345,** 526–528 (1990).
13. Liu, Y., Shao, Z. & Vollrath, F. Relationships between supercontraction and mechanical properties of spider silk. *Nature Mater.* **4,** 901–905 (2005).
14. Vollrath, F. Strength and structure of spiders' silks. *Rev. Mol. Biotechnol.* **74,** 67–83 (2000).
15. Vollrath, F. & Edmonds, D. T. Modulation of the mechanical properties of spider silk by coating with water. *Nature* **340,** 305–307 (1989).
16. Edmonds, D. T. & Vollrath, F. The contribution of atmospheric water vapour to the formation and efficiency of a spider's capture web. *Proc. R. Soc. Lond. B* **248,** 145–148 (1992).
17. Backer, N. *et al.* Molecular nanosprings in spider capture-silk threads. *Nature Mater.* **2,** 278–283 (2003).
18. Peter, H. M. The spinning apparatus of Uloboridae in relation to the structure and construxtion of capture threads (Arachnida, Araneida). *Zoomorphology* **104,** 96–104 (1984).
19. Porter, D. & Vollrath, F. Nanoscale toughness of spider silk. *Nanotoday* **2,** 6 (2007).
20. Emile, O., Floch, A. L. & Vollrath, F. Biopolymers: shape memory in spider draglines. *Nature* **440,** 621 (2006).
21. Yang, J., Yang, Z., Chen, C. & Yao, D. Conversion of surface energy and manipulation of a single droplet across micropatterned surfaces. *Langmuir* **24,** 9889–9897 (2008).
22. Fang, G., Li, W., Wang, X. & Qiao, G. Droplet motion on designed microtextured superhydrophobic surfaces with tunable wettability. *Langmuir* **24,** 11651–11660 (2008).
23. Wenzel, R. N. Resistance of solid surface to wetting by water. *Ind. Eng. Chem.* **28,** 988–994 (1936).
24. Quéré, D. Wetting and roughness. *Annu. Rev. Mater. Res.* **38,** 16.1–16.29 (2008).
25. Zheng, Y. M., Gao, X. F. & Jiang, L. Directional adhesion of superhydrophobic butterfly wings. *Soft Matter* **3,** 178–182 (2007).
26. Gau, H., Herminghaus, S., Lenz, P. & Lipowsky, R. Liquid morphologies on structured surfaces: from microchannels to microchips. *Science* **283,** 46–49 (1999).
27. Yoshimitsu, Z., Nakajima, A., Watanabe, T. & Hashimoto, K. Effects of surface structure on the hydrophobicity and sliding behavior of water droplets. *Langmuir* **18,** 5818–5822 (2002).

Author Contributions Y. Zheng, H.B., Z.H. and X.T. performed the experiments. Y. Zheng and Z.H. worked on the water collection of natural spider silks, H.B. and X.T. worked on the fabrication and water collection of artificial spider silk. Y. Zheng and H.B. conducted the control experiments. Y. Zheng, H.B., Y. Zhao and L.J. collected and analysed the data and proposed the mechanism of directional water collection on spider silk. Y. Zheng, F.-Q.N., Y. Zhao, J.Z. and L.J. wrote the text. L.J. conceived the project and designed the experiments.

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 L.J. (jianglei@iccas.ac.cn) or Y. Zhao (zhaoyong@iccas.ac.cn).

LETTERS

Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature

Elisabetta Collini[1]*†, Cathy Y. Wong[1]*, Krystyna E. Wilk[2], Paul M. G. Curmi[2], Paul Brumer[1] & Gregory D. Scholes[1]

Photosynthesis makes use of sunlight to convert carbon dioxide into useful biomass and is vital for life on Earth. Crucial components for the photosynthetic process are antenna proteins, which absorb light and transmit the resultant excitation energy between molecules to a reaction centre. The efficiency of these electronic energy transfers has inspired much work on antenna proteins isolated from photosynthetic organisms to uncover the basic mechanisms at play[1–5]. Intriguingly, recent work has documented[6–8] that light-absorbing molecules in some photosynthetic proteins capture and transfer energy according to quantum-mechanical probability laws instead of classical laws[9] at temperatures up to 180 K. This contrasts with the long-held view that long-range quantum coherence between molecules cannot be sustained in complex biological systems, even at low temperatures. Here we present two-dimensional photon echo spectroscopy[10–13] measurements on two evolutionarily related light-harvesting proteins isolated from marine cryptophyte algae, which reveal exceptionally long-lasting excitation oscillations with distinct correlations and anti-correlations even at ambient temperature. These observations provide compelling evidence for quantum-coherent sharing of electronic excitation across the 5-nm-wide proteins under biologically relevant conditions, suggesting that distant molecules within the photosynthetic proteins are 'wired' together by quantum coherence for more efficient light-harvesting in cryptophyte marine algae.

Cryptophytes are eukaryotic algae that live in marine and freshwater environments. They are members of an evolutionary group notable because their photosynthetic apparatus was acquired from red algae by a sequence of endosymbiotic events. As a result, cryptophyte photosynthetic antenna proteins (phycobiliproteins) exhibit exceptional spectral variation between species because they use mainly tunable linear tetrapyroles (bilins) for light-harvesting. Another remarkable feature of cryptophytes is that they can photosynthesize in low-light conditions, which suggests that the absorption of incident sunlight by phycobiliprotein antennae in the intrathylakoid space[14] and the subsequent transfer of that energy among these proteins and eventually to the membrane-bound photosystems is particularly effective[15]. Theory indicates that fast energy transfer is facilitated by small interchromophore separations[2], yet the average nearest-neighbour centre-to-centre separation of chromophores within cryptophyte light-harvesting antenna proteins (Fig. 1) is ~20 Å (ref. 16)—about double that for the major light-harvesting protein in plants. To explore how a light-harvesting antenna can function efficiently with such a counter-intuitive design, we study the antennae of two marine cryptophytes, phycoerythrin PE545 from *Rhodomonas* CS24 and phycocyanin PC645 from *Chroomonas* CCMP270 at ambient temperature (294 K) using two-dimensional photon echo (2DPE) spectroscopy[10–13].

PC645 contains eight light-absorbing bilin molecules covalently bound to a four-subunit protein scaffold[17]. Its structure, determined to 1.4-Å resolution by X-ray crystallography[18] and shown in Fig. 1a, exhibits approximate twofold symmetry. A dihydrobiliverdin (DBV) dimer (green) located in the centre of the protein and two mesobiliverdin (MBV) molecules (blue) located near the protein periphery give rise to the upper half of the complex's absorption spectrum (Fig. 1c), spanned by our laser pulse spectrum. The electronic coupling of ~320 cm^{-1} between the DBV molecules C and D (labelled according to the protein subunit that binds them) leads to delocalization of the excitation and yields the dimer electronic excited states, or so-called molecular excitonic states[19], labelled DBV$_+$ and DBV$_-$. Excitation
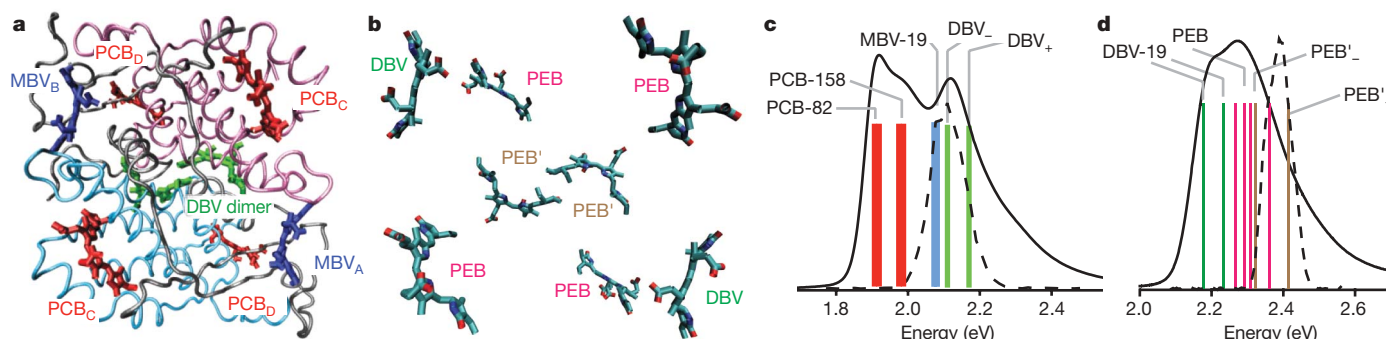

Figure 1 | Structure and spectroscopy of cryptophyte antenna proteins. a, Structural model of PC645. The eight light-harvesting bilin molecules are coloured red (PCB), blue (MBV) and green (DBV). **b,** Chromophores from the structural model for PE545 showing the different chromophore incorporation. **c,** Electronic absorption spectrum of isolated PC645 protein in aqueous buffer (294 K). The approximate absorption energies of the bilin molecules are indicated as coloured bars. **d,** Electronic absorption spectrum of isolated PE545 protein in aqueous buffer (294 K) with approximate absorption band positions indicated by the coloured bars. The spectrum of the ultrafast laser pulse is plotted as a dashed line in **c** and **d**.

[1]Department of Chemistry, Institute for Optical Sciences and Centre for Quantum Information and Quantum Control, University of Toronto, 80 St George Street, Toronto, Ontario, M5S 3H6 Canada. [2]School of Physics and Centre for Applied Medical Research, St Vincent's Hospital, The University of New South Wales, Sydney, New South Wales 2052, Australia. †Present address: Dipartimento di Scienze Chimiche, Università di Padova, via Marzolo 1, 35100, Padova, Italy. *These authors contributed equally to this work.

energy absorbed by the dimer flows to the MBV molecules. which are each 23 Å from the closest DBV, and ultimately to four phycocyano-bilins (PCB, coloured red) that absorb in the lower-energy half of the absorption spectrum.

The structure of PE545 (Fig. 1b) is closely related to that of PC645 except that the bilin types differ[16,20]. The lowest-energy chromo-phores are DBV bilins. The dimer consists of phycoerythrobilin chro-mophores PEB′, with the prime indicating they are doubly covalently bound to the protein. The remaining chromophores are singly bound PEBs. The electronic couplings between the chromophores are reported elsewhere[21]. The approximate absorption spectrum and band positions are shown in Fig. 1d.

For the experiments the proteins were isolated from the algae and suspended at low concentration in aqueous buffer at ambient tem-perature (294 K). The femtosecond laser pulse (25-fs duration) excites a coherent superposition of the antenna protein's electronic-vibrational eigenstates (absorption bands). The initial state of the system is thus prepared in a non-stationary state[22], where electronic excitation is localized to a greater or lesser degree compared to the eigenstates. The time-dependent solution to quantum dynamics for electronically coupled molecules with this initial condition predicts that excitation subsequently oscillates among the molecules under the influence of the system Hamiltonian until the natural eigenstates are restored owing to interactions with the environment. 2DPE provides a means of observing this experimentally, enabling us to explore the significance of quantum coherence.

Representative 2DPE data for PC645 are shown in Fig. 2 with posi-tions on the diagonal assigned to absorption bands. Rich features such as cross-peaks and excited state absorptions are evident. In the 2DPE experiment the two-pulse excitation sequence (sweeping $\tau > 0$) can prepare population density, for example $|DBV_-\rangle\langle DBV_-|$, that evolves during the delay time T and can be probed as a bleach signal on the diagonal part of a rephasing 2DPE spectrum. Alternatively, off-diagonal contributions like $|DBV_+\rangle\langle DBV_-|$ can be excited when the pump pulse sequence interacts coherently with both absorption bands. The resulting signal will be probed as a cross-peak above the diagonal in rephasing spectra that oscillate as a function of T with frequency $\phi = 2\pi(E_{DBV_+} - E_{DBV_-})/h$ because it carries a phase

$\exp(-i\phi T)$. Similarly the complementary coherence $|DBV_-\rangle\langle DBV_+|$ will contribute a cross-peak below the diagonal in rephasing spectra that will carry an opposite phase, $\exp(+i\phi T)$.

These predicted coherent oscillations can be reproducibly seen in our 2DPE spectra by plotting the intensity of rephasing spectra at lower and upper cross-peaks as a function of waiting time T (Fig. 2b and c). The red line indicates the cross-peak above the diagonal, the black line is that below. As mentioned above, the upper and lower cross-peak oscillations should differ by a phase factor determined by the sign of the energy difference between the states in superposition, leading to anti-correlated upper and lower cross-peak beats with a dominant frequency component equal to the eigenvalue energy dif-ference. Such behaviour is indeed clearly seen in the experimental data, with the anti-correlated oscillations providing striking evidence that both DBV dimer and DBV–MBV electronic superposition states persist for more than 400 fs after photo-excitation. It is remarkable that electronic coherence spans from the DBV dimer to the peri-pheral MBV molecules, over a distance of 25 Å.

The PC645 cross-peak beating is complex; multiple frequencies arise for the same reason they do in a simple mechanical system comprised of a mass connected by a weak spring to a pair of masses coupled by a strong spring. And because these data were recorded at room tem-perature, line broadening is significant, so that overlapping bands partly obscure oscillating features. Nevertheless, Fourier transforms of these data (Supplementary Fig. 1) suggest the presence of frequencies in these beating patterns that can be related to the frequency differences between absorption bands. A careful global analysis of the data (see Sup-plementary Information) provides evidence that the oscillations in the 2DPE data can be decomposed into components corresponding to frequency differences between absorption bands and that—most importantly—the cross-peak beats at each frequency are anti-correlated.

For a comparison with the PC645 results, we also undertook experiments on the PE545 antenna protein by exciting the blue side of the absorption. A typical rephasing 2DPE spectrum (that is, scanned so that $\tau > 0$) is shown in Fig. 3a. To show beats across the entire anti-diagonal slice through the PEB/PEB′ cross-peaks we plot the intensity of the 2DPE rephasing spectrum along the anti-diagonal line drawn in Fig. 3a as a function of population time T (Fig. 3b).
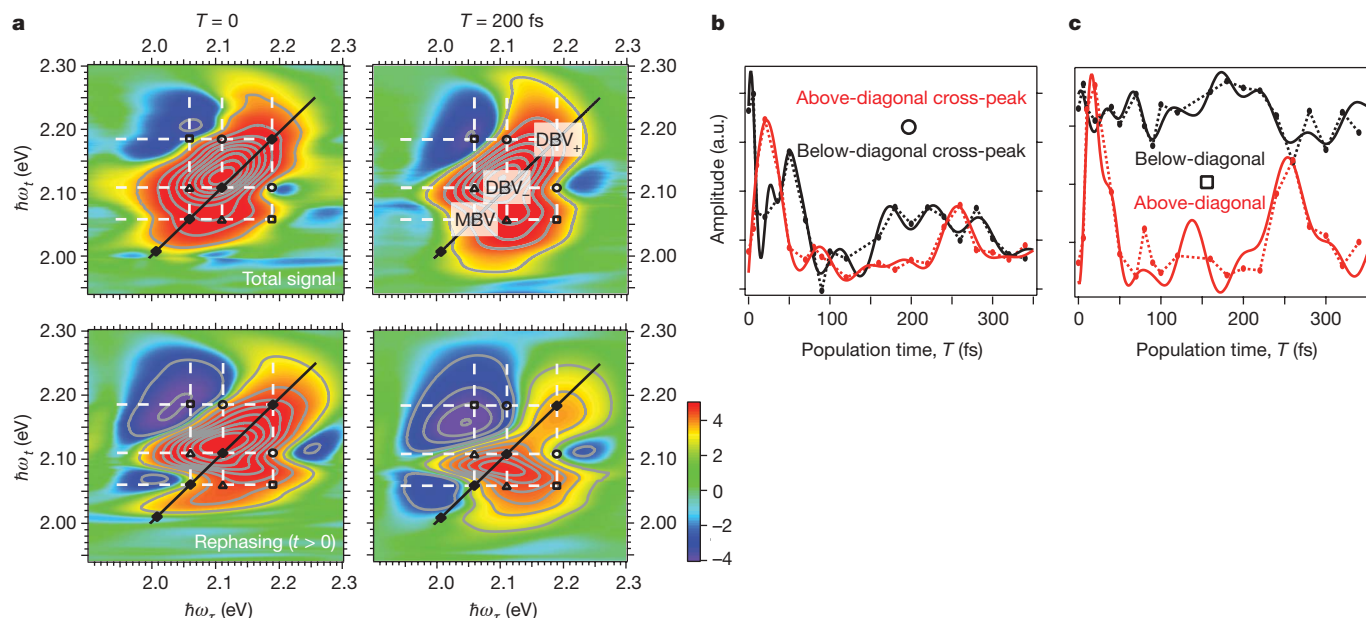

Figure 2 | Two-dimensional photon echo data for PC645. a, The left column shows the total real 2DPE spectrum recorded for PC645 at zero waiting time ($T = 0$), together with the rephasing contribution to this signal. The right column shows the data for $T = 200$ fs. The 2DPE spectra show the signal intensity on an arcsinh scale (colour scale, arbitrary units) plotted as a function of coherence frequency ω_τ and emission frequency ω_t. **b,** Intensity of the DBV dimer cross-peaks (open circle) as a function of time T. **c,** Intensity of the MBV–DBV$_+$ cross-peaks (open square) as a function of time T. The dashed lines interpolate the data points (solid circles). The solid line is a fit to a sum of damped sine functions (Supplementary Information). a.u., arbitrary units.
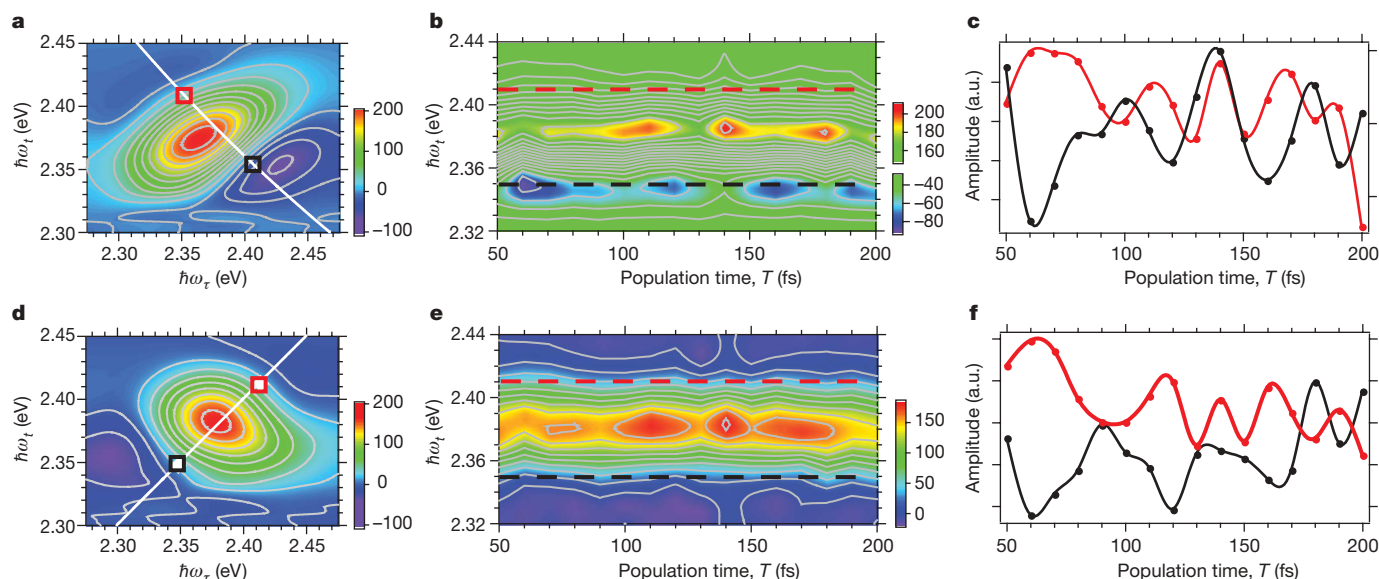
Figure 3 | Two-dimensional photon echo data for PE545. a, 2DPE spectrum (rephasing real signal) for PE545 recorded at $T = 100$ fs. **b,** The intensity of the 2DPE rephasing spectrum along an anti-diagonal slice through the cross-peaks versus population time T. Upper and lower cross-peaks are indicated by red and black dashed lines respectively. **c,** Intensity oscillations in the cross-peaks (red and black squares in **a**). **d,** 2DPE spectrum (non-rephasing real signal) for PE545 recorded at $T = 100$ fs. **e, f,** As for **b** and **c** but for the 2DPE non-rephasing spectrum along the diagonal slice. The 2DPE spectra are plotted on a linear intensity scale.

Oscillations of the main bleach and excited-state absorption peaks are clearly evident. The beats in the centre of this plot are cross-peaks excited because the PEB and PEB' absorption bands overlap owing to spectral line broadening. The cross-peaks, indicated by dashed horizontal lines in Fig. 3b (red is the cross-peak above the diagonal, the black line is that below), are more clearly seen in Fig. 3c, in which the beats are well resolved and markedly anti-correlated—a signature of quantum coherence. These oscillations are directly analogous to those observed for PC645 (compare Fig. 2b).

Figure 3d–f shows plots similar to those in Fig. 3a–c, but for non-rephasing spectra ($\tau < 0$) of PE545. The same electronic coherences giving rise to oscillating cross-peaks in rephasing spectra are predicted to cause oscillations at the diagonal positions of non-rephasing 2DPE spectra[23]. Indeed, we observe a clear phase relationship between beats within the rephasing and non-rephasing spectra and also between the data sets, which is compelling evidence for the presence of long-lived quantum coherence. The first ~130 fs of these spectra reproducibly show clear oscillations with a period of ~60 fs ($\nu \approx 500$ cm^{-1}). After this time the oscillation pattern becomes more complicated, suggesting that the initial coherence may evolve owing to coupling with other molecules in the protein.

In our experiments the light-harvesting process in both PC645 and PE545 antenna proteins involves quantum coherence at ambient temperature, suggesting that coherence may more generally be used by cryptophyte algae. Quantum coherence occurs in an intermediate regime of energy transfer where there is a complex balance between quantum interference among electronic resonances and coupling to the environment causing decoherence[24]. There still remains the question of precisely how quantum coherence can persist for hundreds of femtoseconds in these biological assemblies. In an isolated molecule, electronic decoherence arises from the decay of the overlap $S(t) = \langle v_2(t) \mid v_1(t) \rangle$ between the unobserved vibrational wavepackets $|v_1(t)\rangle$ and $|v_2(t)\rangle$ associated with the lower and upper electronic states respectively[25,26]. In 2DPE experiments, the observable includes both vibrational and electronic components and, as such, decoherence due to decay of $S(t)$ is not manifest in the data. Rather, the slow decay of electronic coherence reflects the interaction of vibronic superposition states with the external environment.

Recent studies have attributed the slow dephasing of electronic coherence to the presence of shared or correlated motions in the surrounding environment[6,7,27,28]. In this context, we note that, unlike most photosynthetic pigments that are non-covalently complexed to their protein environment (chlorophyll via histidine residues, for example), the bilins in PC645 are covalently bound to the protein backbone. Covalent attachment of the chromophores to their protein environment may support or strengthen correlated motions between chromophores and protein and thus be an important factor in slowing down decoherence in cryptophyte antenna proteins at ambient temperature, thereby differentiating them from many other photosynthetic light-harvesting antennae. We also note that the precise manifestation of long-lived quantum-coherence depends on the photo-excitation conditions[22,29,30], and cryptophyte algae are obviously using sunlight that does not arrive in the form of laser pulses as used in our experiments. Nevertheless, the couplings giving rise to the long-lived quantum coherence that we clearly observe at ambient temperature will still be present and strongly suggest that quantum effects facilitate the efficient light-harvesting by cryptophyte algae. That is, long-lived quantum coherence can facilitate energy transfer by 'wiring' together the final energy acceptors (PCB in the case of PC645 and DBV for PE545) across a single protein unit, and thereby help to compensate for the exceptionally large average interchromophore separations in these antenna proteins.

METHODS SUMMARY

Cryptophyte *Chroomonas* sp. (CCMP270 strain, National Culture Collection of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, USA) and *Rhodomonas* sp. (CS24) were cultivated and harvested, and the phycobiliproteins were isolated by usual procedures[20]. 2DPE experiments were performed as described in refs 11 and 27. The laser pulse duration and chirp were measured using transient grating frequency resolved optical gating (TG-FROG) experiments on a solvent (typically ethanol). The time-bandwidth product was estimated to be about 0.53, close to the ideal transform-limited condition for a Gaussian pulse. During data collection for PC645, for any given population time T, τ was scanned from −200 to 200 fs with 0.25-fs steps. Each 2D map at a fixed T is the average of at least three separate scans, and each series of 2D scans at different T was further repeated on different days for comparison. For PE545, τ was scanned from −60 to 60 fs with 0.15-fs steps, and each 2D map is the average of two separate scans. Additionally, PE545 was measured at this wavelength numerous times with different T steps, on different days. The samples were moved after each scan and absorption spectra taken before and after each series of scans confirmed that the sample did not degrade during the measurements.

1. Green, B. R. & Parson, W. W. (eds) *Light-Harvesting Antennas in Photosynthesis* (Kluwer, Dordrecht, 2003).
2. Scholes, G. D. Long-range resonance energy transfer in molecular systems. *Annu. Rev. Phys. Chem.* **54,** 57–87 (2003).
3. Jang, S., Newton, M. D. & Silbey, R. J. Multichromophoric Förster resonance energy transfer from B800 to B850 in the light harvesting complex 2: evidence for subtle energetic optimization by purple bacteria. *J. Phys. Chem. B* **111,** 6807–6814 (2007).
4. Cheng, Y. C. & Fleming, G. R. Dynamics of light harvesting in photosynthesis. *Annu. Rev. Phys. Chem.* **60,** 241–262 (2009).
5. van Grondelle, R. & Novoderezhkin, V. I. Energy transfer in photosynthesis: experimental insights and quantitative models. *Phys. Chem. Chem. Phys.* **8,** 793–807 (2006).
6. Engel, G. S. *et al.* Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature* **446,** 782–786 (2007).
7. Lee, H., Cheng, Y. C. & Fleming, G. R. Coherence dynamics in photosynthesis: protein protection of excitonic coherence. *Science* **316,** 1462–1465 (2007).
8. Mercer, I. P. *et al.* Instantaneous mapping of coherently coupled electronic transitions and energy transfers in a photosynthetic complex using angle-resolved coherent optical wave-mixing. *Phys. Rev. Lett.* **102,** 057402 (2009).
9. Feynman, R. P. Space-time approach to non-relativistic quantum mechanics. *Rev. Mod. Phys.* **20,** 367–387 (1948).
10. Jonas, D. M. Two-dimensional femtosecond spectroscopy. *Annu. Rev. Phys. Chem.* **54,** 425–463 (2003).
11. Brixner, T., Mancal, T., Stiopkin, I. V. & Fleming, G. R. Phase-stabilized two-dimensional electronic spectroscopy. *J. Chem. Phys.* **121,** 4221–4236 (2004).
12. Cho, M. H. Coherent two-dimensional optical spectroscopy. *Chem. Rev.* **108,** 1331–1418 (2008).
13. Abramavicius, D. *et al.* Coherent multidimensional optical spectroscopy of excitons in molecular aggregates; quasiparticle versus supermolecule perspectives. *Chem. Rev.* **109,** 2350–2408 (2009).
14. Spear-Bernstein, L. & Miller, K. R. Unique location of the phycobiliprotein light-harvesting pigment in the cryptophyceae. *J. Phycol.* **25,** 412–419 (1989).
15. van der Weij-De Wit, C. D. *et al.* Phycocyanin sensitizes both photosystem I and photosystem II in cryptophyte *Chroomonas* CCMP270 cells. *Biophys. J.* **94,** 2423–2433 (2008).
16. Wilk, K. E. *et al.* Evolution of a light-harvesting protein by addition of new subunits and rearrangement of conserved elements: crystal structure of a cryptophyte phycoerythrin at 1.63-Ångstrom resolution. *Proc. Natl Acad. Sci. USA* **96,** 8901–8906 (1999).
17. Wedemayer, G. J., Kidd, D. G., Wemmer, D. E. & Glazer, A. N. Phycobilins of cryptophycean algae: occurrence of dihydrobiliverdin and mesobiliverdin in cryptomonad biliproteins. *J. Biol. Chem.* **267,** 7315–7331 (1992).
18. Mirkovic, T. *et al.* Ultrafast light harvesting dynamics in the cryptophyte phycocyanin 645. *Photochem. Photobiol. Sci.* **6,** 964–975 (2007).
19. Scholes, G. D. & Rumbles, G. Excitons in nanoscale systems. *Nature Mater.* **5,** 683–696 (2006).
20. Doust, A. B. *et al.* Developing a structure-function model for the cryptophyte phycoerythrin 545 using ultrahigh resolution crystallography and ultrafast laser spectroscopy. *J. Mol. Biol.* **344,** 135–153 (2004).
21. Scholes, G. D. *et al.* How solvent controls electronic energy transfer and light harvesting. *J. Phys. Chem. B* **111,** 6978–6982 (2007).
22. Rhodes, W. Radiationless transitions in isolated molecules. the effects of molecular size and radiation bandwidth. *J. Chem. Phys.* **50,** 2885–2896 (1969).
23. Cheng, Y. C. & Fleming, G. R. Coherence quantum beats in two-dimensional electronic spectroscopy. *J. Phys. Chem. A* **112,** 4254–4260 (2008).
24. Rackovsky, S. & Silbey, R. Electronic-energy transfer in impure solids. 1. 2 molecules embedded in a lattice. *Mol. Phys.* **25,** 61–72 (1973).
25. Hwang, H. & Rossky, P. J. Electronic decoherence induced by intramolecular vibrational motions in a betaine dye molecule. *J. Phys. Chem. B* **108,** 6723–6732 (2004).
26. Franco, I., Shapiro, M. & Brumer, P. Femtosecond dynamics and laser control of charge transport in trans-polyacetylene. *J. Chem. Phys.* **128,** 244905 (2008).
27. Collini, E. & Scholes, G. D. Coherent intrachain energy migration in a conjugated polymer at room temperature. *Science* **323,** 369–373 (2009).
28. Beljonne, D., Curutchet, C., Scholes, G. D. & Silbey, R. Beyond Förster resonance energy transfer in biological and nanoscale systems. *J. Phys. Chem. B* **113,** 6583–6599 (2009).
29. Langhoff, C. A. & Robinson, G. W. Time decay and untangling of vibronically tangled resonances: naphthalene second singlet. *Chem. Phys.* **6,** 34–53 (1974).
30. Jang, S. Theory of coherent resonance energy transfer for coherent initial condition. *J. Chem. Phys.* **131,** 164101 (2009).

METHODS

Cryptophyte *Chroomonas* sp. (CCMP270 strain, National Culture Collection of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, USA) was cultured at 20 °C under constant low light illumination (12-V white fluorescent tubes, 300 lx at 0.3 m) in a modified 'Fe' medium. The algal cells were harvested and passed twice through a French press cell at a pressure of 1,000 psi. The resultant solution was centrifuged, yielding a supernatant solution containing PC645. PC645 was isolated using gradual ammonium sulphate precipitation from 0 to 80%. The pellets were re-suspended in a minimal volume of 25 mM phosphate buffer at pH 7.1. Further purification was continued using a combination of ion-exchange and size-exclusion chromatography. About 20 ml of pure PC645 was concentrated to approximately 200 µl using a 10-kDa Amicon Centriprep and then frozen using liquid nitrogen before being stored at −80 °C. A buffer solution, prepared from a 25 mM solution of HEPES (4-(2–hydroxyethyl)–1–piperazineethanesulphonic acid) in deionized water, adjusted to pH = 7.5 by the addition of concentrated NaOH solution, was used to prepare dilute PC645 samples for the experiments.

Rhodomonas sp. (CS24) was cultivated and harvested as previously reported[20]. Cell pellets were re-suspended in buffer A (0.05 M Mes (pH 6.5) with 1 mM NaN_3) and homogenized in a Teflon glass homogenizer followed by passage though a French press at a pressure of 1,000 psi. The resultant solution was centrifuged for 30 min at 17,000g, producing a pellet of cell debris containing thylakoid membranes and a supernatant containing phycoerythrin. PE545 was purified from the supernatant as described elsewhere[20].

2DPE experiments were performed as described in refs 11 and 27. A Ti:sapphire regeneratively amplified laser system was used to pump a Noncolinear Optical Parametric Amplifier (NOPA) to produce 25-fs duration pulses centred at 590 nm for PC645 experiments or at 520 nm for PE545 for the results reported here, with a spectral bandwidth of about 25 nm and repetition rate of 1 kHz. The laser pulse duration and chirp were measured using TG-FROG experiments on a solvent (typically ethanol). The time-bandwidth product was estimated to be about 0.53, close to the ideal transform-limited condition for a Gaussian pulse. The pulse from the NOPA was split by a 50% beam splitter and the two resulting beams were overlapped in a diffractive optic, producing two pairs of phase-locked beams in a boxcars phase-matched geometry. The delay time T was controlled by a motorized translation stage inserted in one beam path before the diffractive optic, whereas the delay time τ was introduced by means of movable glass wedge pairs, calibrated by spectral interferometry[11].

During data collection for PC645, for any given population time T, τ was scanned from −200 to 200 fs with 0.25-fs steps. Each 2D map at a fixed T is the average of at least three separate scans, and each series of 2D scans at different T was further repeated on different days for comparison. For PE545, τ was scanned from −60 to 60 fs with 0.15-fs steps, and each 2D map is the average of two separate scans. Additionally, PE545 was measured at this wavelength numerous times with different T steps, on different days. The samples were moved after each scan and absorption spectra taken before and after each series of scans confirmed that the sample did not degrade during the measurements.

To ensure that the local oscillator did not influence the response of the system, its intensity was attenuated by about three orders of magnitude relative to the other beams and the time ordering was set so that the local oscillator always preceded the probe by ∼500 fs. The resulting local oscillator–signal interference intensity was focused into a 0.63-m spectrograph (25-µm slit) and recorded using a 16-bit, 400 × 1,600 pixel, thermo-electrically cooled charge-coupled device (CCD) detector. Subtraction of unwanted scatter contributions, Fourier window-ing, transformation and phase retrieval were performed as reported previously[11].

LETTERS

Migrating tremors illuminate complex deformation beneath the seismogenic San Andreas fault

David R. Shelly[1]

The San Andreas fault is one of the most extensively studied faults in the world, yet its physical character and deformation mode beneath the relatively shallow earthquake-generating portion remain largely unconstrained. Tectonic 'non-volcanic' tremor, a recently discovered seismic signal[1] probably generated by shear slip on the deep extension of some major faults[2–4], can provide new insight into the deep fate of such faults, including that of the San Andreas fault near Parkfield, California[5]. Here I examine continuous seismic data from mid-2001 to 2008, identifying tremor and decomposing the signal into different families of activity based on the shape and timing of the waveforms at multiple stations[6]. This approach allows differentiation between activities from nearby patches of the deep fault and begins to unveil rich and complex patterns of tremor occurrence. I find that tremor exhibits nearly continuous migration, with the most extensive episodes propagating more than 20 kilometres along fault strike at rates of 15–80 kilometres per hour. This suggests that the San Andreas fault remains a localized through-going structure, at least to the base of the crust, in this area. Tremor rates and recurrence behaviour changed markedly in the wake of the 2004 magnitude-6.0 Parkfield earthquake[6,7], but these changes were far from uniform within the tremor zone, probably reflecting heterogeneous fault properties and static and dynamic stresses decaying away from the rupture. The systematic recurrence of tremor demonstrated here suggests the potential to monitor detailed time-varying deformation on this portion of the deep San Andreas fault, deformation which unsteadily loads the shallower zone that last ruptured in the 1857 magnitude-7.9 Fort Tejon earthquake[8].

Earthquakes on the San Andreas fault mostly range between depths of 2 and 15 km and provide information about the location and style of deformation of the fault in this upper-crustal 'seismogenic' zone. Beyond the maximum depth of earthquakes, however, relatively little is known about the character of the fault. While some have proposed that the San Andreas and other faults may evolve into broadly distributed zones of deformation beneath the seismogenic layer[9], recent observations of deep tremor and slow slip suggest that faults can remain localized well below the down-dip limit of earthquakes[10].

Deep tremor was first observed in the subduction zones of southwest Japan[1] and Cascadia[11]. In these locations, the majority of tremor is generated in relatively large bursts of concentrated activity, often accompanied by geodetically detectible slow slip. In southwest Japan, these bursts typically last several days and occur at periods of 3–6 months[12]. By comparison, Cascadia events tend to be larger in area, longer in duration, and less frequent, with typical periodicities of 10–19 months, depending on the location of events along the margin[13].

Tremor has also been identified beneath the strike-slip San Andreas fault near Parkfield, California. This tremor exhibits many of the same characteristics as subduction zone tremor, including a

dominant frequency content of ~2–8 Hz (ref. 5), apparent location on the deep extension of the fault, and waveforms consisting of repeated similar events[4]. In addition, like tremor in Nankai and Cascadia, San Andreas fault tremor is triggered by very small stresses imparted by teleseismic waves[14–16] and tidal fluctuations[17]. Tremor locations aligned with fault strike[4] and triggering by tidally induced right lateral shear stress[17] strongly suggest that San Andreas fault tremor is generated by tectonically driven slip on the deep fault, as has been argued for tremor in Japan[2,18] and Cascadia[3]. Despite the many similarities to subduction zone tremor, no accompanying geodetic signature has yet been detected near Parkfield[19], probably owing to the relatively small spatial dimension and short recurrence times of activity in this area.

As in southwest Japan[18], tremor beneath the San Andreas fault is composed of an overlapping sequence of individual events[4], which have been termed low-frequency earthquakes[20]. While semi-continuous activity and low amplitudes often prevent identification of the P and S wave arrivals that are generally used to locate earthquakes, repeated activity from the same patch of the fault can be identified very effectively using cross-correlation in a multi-channel matched filter[21].

Figure 1 shows estimated locations for the suite of 21 different low-frequency earthquakes from ref. 6. As described in that work, these low-frequency earthquakes are used as template events and are then scanned through continuous borehole seismic data from the Parkfield High Resolution Seismic Network (HRSN) from mid-2001 to the end of 2008, with similarity measured as the sum of the correlation coefficients across all channels[18]. The set of events resembling each template is considered an event family.

This approach has several advantages over amplitude-based analysis of tremor[7], including the ability to identify short-duration tremor and low-amplitude tremor. Furthermore, because it relies on the shape and timing of the waveforms, rather than waveform envelopes, this method permits analysis on much shorter spatial and temporal scales, differentiating activity on many small patches of the deep fault. Similarly, this technique automatically differentiates between tremor and other seismic signals, such as earthquake aftershocks or cultural noise. Thus, there is no need to exclude the times immediately following significant earthquakes, when amplitude detection becomes impractical owing to the difficulty of distinguishing semi-continuous shallow aftershocks from deep fault tremor[7].

One of the striking results of this analysis is the constant motion of the tremor source. Figure 2 shows an example of an extensive migration episode, in which the tremor source propagates ~25 km northwest along fault strike over a period of 90 min. This gives an average migration rate of $17\,\text{km}\,\text{h}^{-1}$ ($4.6\,\text{m}\,\text{s}^{-1}$), but the migration rate appears to vary substantially, and local migration velocities are as high as $80\,\text{km}\,\text{h}^{-1}$, similar to migration rates observed in the up- and down-dip directions in southwest Japan[22]. Despite variations in propagation rate in this example, initial onsets of activity in each event

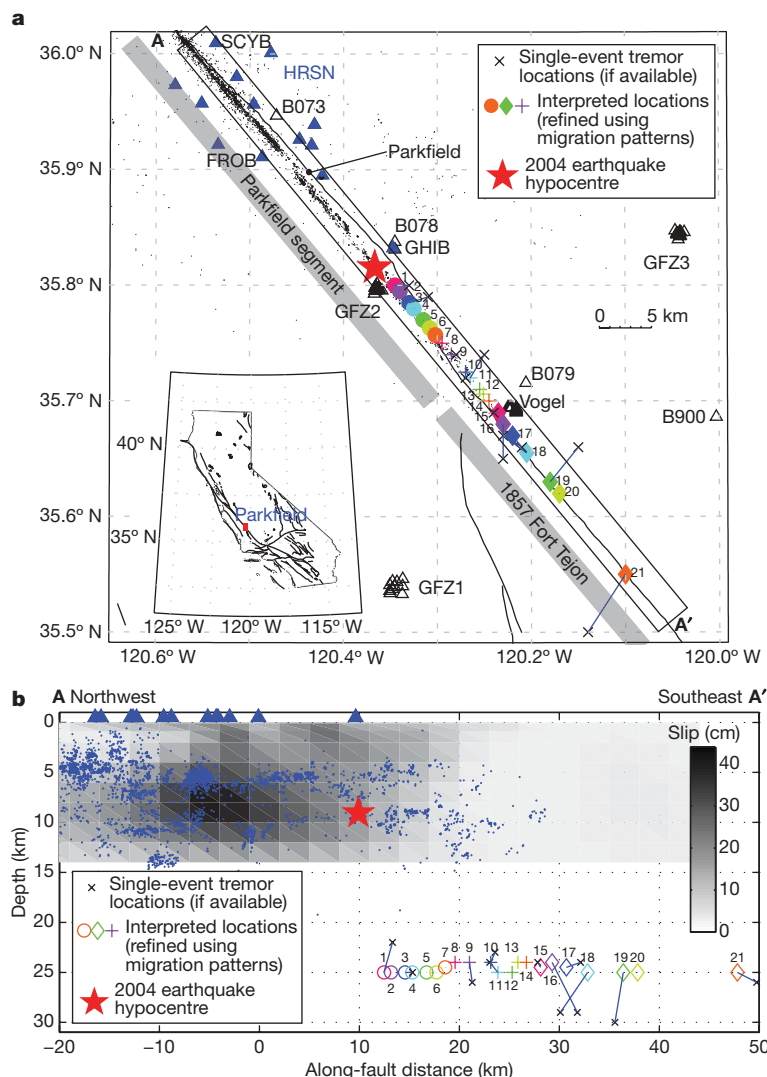[1]US Geological Survey, 345 Middlefield Road, MS 977, Menlo Park, California 94025, USA.

Figure 1 | Map and cross-section showing tremor locations. a, Map view. **b,** Cross-section. The red star shows the hypocentre of the 2004 earthquake. Coloured circle, plus, and diamond symbols show interpreted tremor event-family locations from ref. 6, derived from a combination of arrival time picks and migration patterns. Single-event travel-time-based locations for each family (when available) are shown as black 'x' symbols. Filled blue triangles are stations used in event detection, while black triangles are additional temporary and permanent stations used for template location. Small dots are relocated earthquakes[29]. Shaded grey regions denote the Parkfield segment, which last ruptured in 2004, and the northern end of the segment that last ruptured in the 1857 magnitude-7.9 Fort Tejon earthquake. The greyscale slip model in **b** includes coseismic slip and the first 230 days of afterslip from the 2004 Parkfield earthquake[30]. (Figure modified from ref. 6.)

family show a remarkably systematic progression to the northwest. Activity in a given event family often lingers after the initial front has passed, as seen in Fig. 2. Most San Andreas fault tremor migration episodes traverse shorter distances than in this example (5–10 km is common), yet persistent progressions between event families remain. An additional migration example is shown in Supplementary Fig. 3.

Along-strike tremor migration episodes, such as the one shown in Fig. 2, suggest that the fault exists as a continuous structure at the ~25 km depth of the tremors, near the ~26 km base of the crust in this area[23]. The nature of the fault between 15 and 24 km, however, is unknown, as this area is devoid of earthquakes and does not appear to generate appreciable tremor.

Despite migration episodes that sometimes encompass many event families, different families exhibit very different recurrence patterns, with typical recurrence periods ranging from a few days to a few months. Figure 3 shows cumulative events over time for three representative event families: one with large bursts of hundreds of events every few months, one with small bursts of several events every few days, and one with a hybrid pattern of large and small bursts.

Figure 4 summarizes this behaviour for all 21 families and demonstrates small-scale segmentation within the tremor zone, with groups of families showing coincident, or nearly coincident, bursts of activity. The highest overall level of tremor activity is observed during bursts lasting 1–3 days, recurring every few months. These can be seen in Fig. 4a as the apparent striping (bright bars) in families 8–11 and 13–17. Bursts of activity in these two zones often coincide closely in time, with activity in either zone preceding the other by 1–3 days. Occasionally, however, they become out-of-phase with one another, for example in mid-2006 (Figs 3a and 4a). Other event families typically exhibit small bursts of activity every few days, with varying degrees of regularity. For example, families 17–19 typically exhibit bursts of activity every 2–6 days (Fig. 4b and Supplementary Fig. 4); sometimes showing sympathetically increased activity along with large bursts at neighbouring families (Fig. 4b). This behaviour argues that different patches of the deep fault slip with corresponding variation in frequency, probably reflecting differences in the strength of the deep fault.

The semi-regularity of bursts at each family can be seen in Fig. 4b in the months preceding the 2004 Parkfield earthquake. This regularity suggests a model in which each burst of activity within a given family
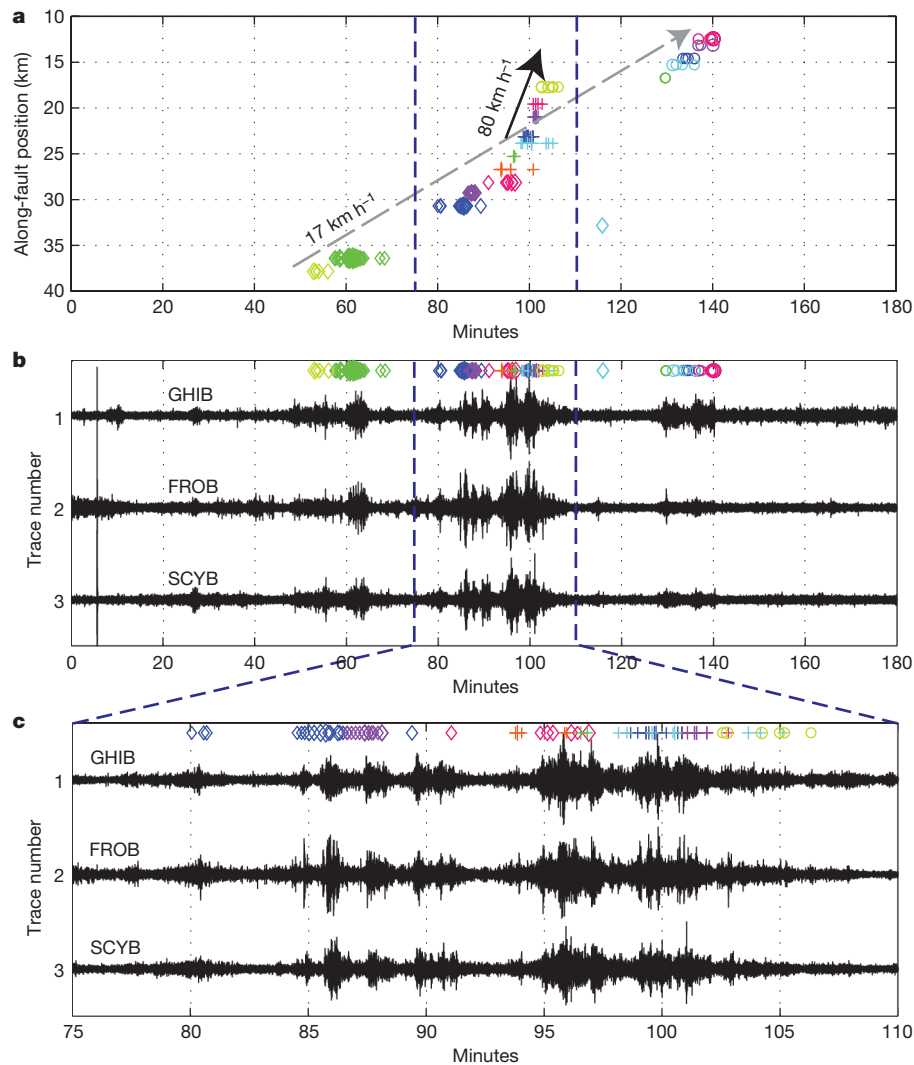
Figure 2 | Tremor migration. a, Example of an extended tremor migration episode, propagating ~25 km along fault strike over 90 min, from event family 20 northwest to event family 1. The average propagation rate in this episode is ~17 km, but local propagation rates vary substantially. Symbols are the same as in Fig. 1. This episode occurred on 26 October 2007 from 12:00–15:00 UTC (Coordinated Universal Time). **b**, Waveforms and event-family matches during the same period at three HRSN borehole stations (horizontal component). **c**, Zoomed view of the highest-amplitude (and fastest-migrating) portion of the sequence.

ruptures the same portion of the deep fault, releasing tectonic stress that is accumulated between bursts. If so, and I assume that slip averages ~3 cm yr^{-1}, I can estimate the amount of slip in each episode. Families with shorter return periods, for example 5 days, would average only 0.4 mm of slip in each burst. Assuming a 1 km radius dislocation (which should contain the members of a given family), the moment magnitude equivalent of such an event is 1.0 (assuming a shear modulus of 30 GPa), which will clearly fall below the detection thresholds of surface geodetic instruments. On the other hand, those families with the longest return periods display about three bursts a year, or ~1 cm of slip per event, and are somewhat more extensive, often 15 km along strike. Assuming a circular dislocation with a radius of 7.5 km, this would be equivalent to a moment magnitude of 4.8. Occurring at these depths (~25 km) over 2–3 days, such an event would probably still evade detection by current instrumentation[24], but might perhaps be observable by a long-baseline laser strainmeter recently installed in the area[25].

Tremor activity was strongly affected following the 2004 earthquake. The nature of the change, however, varies substantially within the tremor zone. The overall occurrence rates increased markedly for some families, but this effect mostly decayed with increasing distance from the earthquake rupture, as illustrated in cumulative event plots for three representative families in Fig. 3. On the other hand, the recurrence frequency of tremor bursts increased more broadly (Fig. 4b and Supplementary Figs 4 and 5), including some event families relatively distant from the rupture whose overall occurrence rates remained nearly constant, such as family 19 (rate is seen in Fig. 3; recurrence is shown in Fig. 4 and Supplementary Fig. 4). As this implies, the increased frequency of event bursts within a family was accompanied by a decrease in the number of events per burst (Figs 3 and 4), suggesting that the deep fault was temporarily weakened following the earthquake. This could imply that static stresses from the earthquake, which decay more quickly with distance, are driving the rate increase, while further-reaching dynamic stresses may temporarily weaken the fault. Although activity returned mostly back to the pre-earthquake pattern after a few months, occurrence rates in some families remained elevated through the end of the study period relative to their occurrence in the year preceding the 2004 earthquake, possibly reflecting an extended post-seismic deformation transient in the lower crust and/or uppermost mantle.

Although growing evidence supports deep fault shear failure as the basic mechanism of tremor, questions remain regarding physical conditions that control the spatial distribution of tremor. Tidal[17] and teleseismic[14–16] triggering of tremor imply that the deep fault is extremely sensitive to small stress perturbations, suggesting the presence of high (near-lithostatic) fluid pressure in the fault zone at these depths. If so,
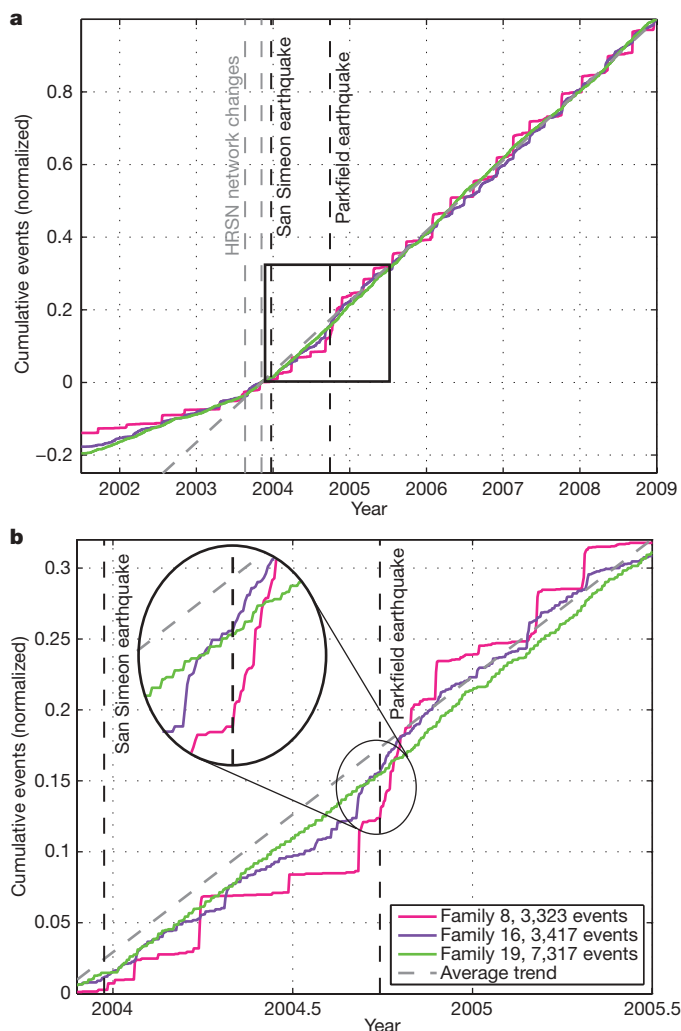
Figure 3 | Event-family recurrence styles. a, Cumulative events with time for three event families with different recurrence patterns. Family 8 (red) exhibits bursts containing many events every few months, while family 19 (green) exhibits smaller bursts every few days. Family 16 (purple) shows a hybrid behaviour of small bursts every few days and a larger burst every few months. Cumulative events are normalized to one between 6 November 2003 (the time of station operational changes) and 31 December 2008. Actual numbers of events are given in the legend. **b**, Zoomed view of **a**, with inset showing details of activity around the Parkfield earthquake, including small frequent bursts in family 19 (green). Notice that bursts in families 8 and 16 are often, but not always, concurrent, suggesting that these patches of the fault can rupture either together or separately.

the fluid source beneath the San Andreas fault is less obvious than in subduction zones, where slab dehydration is expected beneath the tremor zone. Nevertheless, many potential sources of fluid exist. Although a fossil slab may provide water beneath the San Andreas fault[26], the limited spatial extent of the San Andreas fault tremor argues for a more local control. Subducted sediments proposed to exist near the tremor zone[27] might provide a local source of fluids, as would serpentinite or fluid-saturated schist, which have been suggested to exist here on the basis of a receiver function study in this area[23]. High fluid pressures may make the fault extremely weak, possibly permitting patches of the deep fault to deform by brittle rather than ductile failure, even as temperatures approach 600 °C (ref. 28).

Tremor in this area provides an opportunity to examine the relationship between deep fault slip and earthquake occurrence. Whether or not a change in tremor activity might portend a future earthquake rupture is unknown, but high-resolution analysis of tremor provides a new ability to monitor temporal and spatial variations in deformation that might otherwise be concealed. Of particular current interest is tremor activity beneath the segment of the San Andreas fault south of Parkfield, a currently locked portion of the fault that last ruptured in the magnitude-7.9 Fort Tejon earthquake[8]. Although relatively subtle, changes in tremor migration patterns may have signalled accelerated deep fault slip in the three months leading up to the 2004 earthquake[6]. The marked changes in tremor activity following the 2004 earthquake appear to reflect the effects of static and dynamic

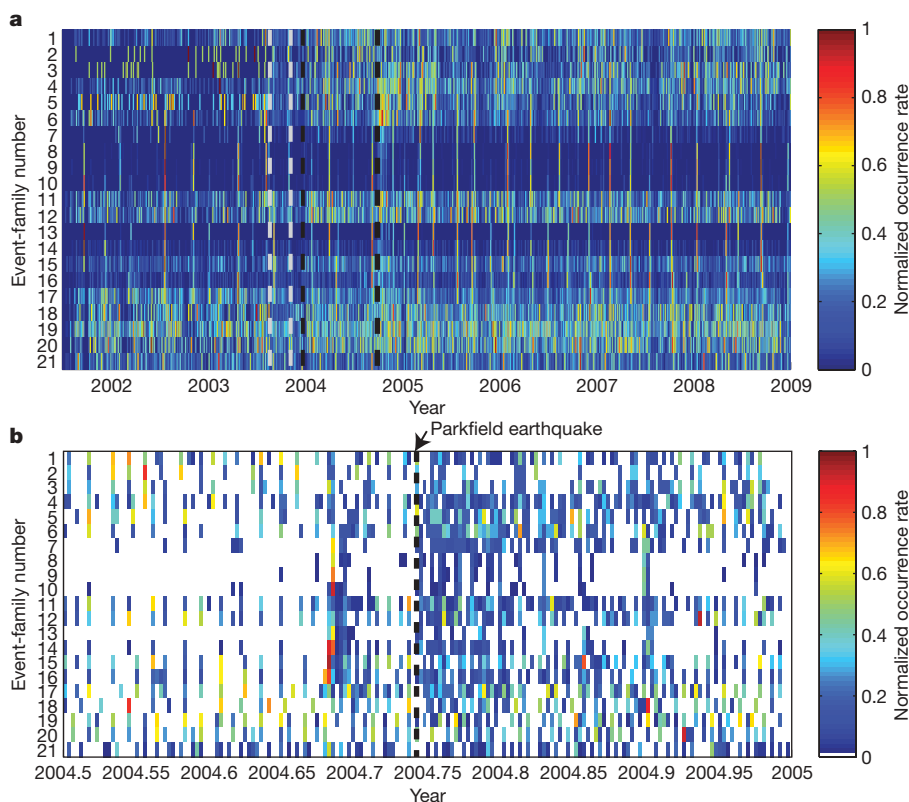**Figure 4 | Recurrence patterns for 21 event families. a**, Number of events in a moving five-day window for each family from mid-2001 to the end of 2008, normalized by the maximum over this period. Events are ordered from northwest (top) to southeast (bottom), with the numbers corresponding to those in Fig. 1. Grey dashed lines in late 2003 indicate times of operational changes to HRSN stations. Black dashed lines indicate the times of San Simeon (22 December 2003) and Parkfield (28 September 2004) earthquakes. Normalization is performed separately before and after 20 August 2003, when the first changes to HRSN operational parameters were made. **b**, Events per day (rather than the five-day window of **a**) for each event family for the second half of 2004, bracketing the Parkfield earthquake. Event numbers are normalized by the maximum from 2001–2008 for each family. Notice the pronounced increase in event frequency for nearly all families following the earthquake. While bursts become more frequent, individual bursts contain fewer events, as indicated by the colours.

stresses imparted by the earthquake, demonstrating its relatively far-reaching influence on ongoing deep fault deformation.

METHODS SUMMARY

The event families and detected catalogue used here were introduced by ref. 6, supporting results describing possible changes in tremor activity preceding the 2004 magnitude-6.0 Parkfield earthquake. This processing consists of three main steps: template event selection, template event location and event detection. Template events are typically selected as relatively impulsive events within tremor in the continuous seismic data. When possible, the location is first estimated using compressional (P) and shear (S) wave arrival time picks for these events (Supplementary Table 1). Persistent migration patterns between event families suggest close spatial relationships and make it possible to refine these locations as well as to estimate locations of events for which phase arrivals are not sufficiently clear[6].

Additional events similar to each template event are detected by cross-correlating the template waveforms with the continuous data stream. Each waveform template consists of 6 s of the three-component waveform at each station, with time delays applied to approximately match the S-wave propagation across the network[18,22]. An example template is shown in Supplementary Fig. 1. This waveform template is then scanned through the continuous data with cross-correlations calculated every 0.05 s, with the time delays of the original template preserved. Similarity is measured by the correlation coefficient sum across all channels and stations of the network. For maximum consistency over time, I used the 25 channels of data from the HRSN network that operated most consistently over the study period and apply a fixed correlation sum threshold of 4.0.

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

Received 29 July; accepted 7 December 2009.

1. Obara, K. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* **296**, 1679–1681 (2002).
2. Ide, S., Shelly, D. R. & Beroza, G. C. Mechanism of deep low frequency earthquakes: further evidence that deep non-volcanic tremor is generated by shear slip on the plate interface. *Geophys. Res. Lett.* **34**, L03308, doi:10.1029/2006GL028890 (2007).
3. Wech, A. G. & Creager, K. C. Cascadia tremor polarization evidence for plate interface slip. *Geophys. Res. Lett.* **34**, L22306, doi:10.1029/2007GL031167 (2007).
4. Shelly, D. R. *et al.* Precise location of San Andreas fault tremors near Cholame, California using seismometer clusters: slip on the deep extension of the fault? *Geophys. Res. Lett.* **36**, L01303, doi:10.1029/2008GL036367 (2009).
5. Nadeau, R. M. & Dolenc, D. Nonvolcanic tremors deep beneath the San Andreas fault. *Science* **307**, 389, doi:10.1126/science.1107142 (2005); published online 9 December 2004.
6. Shelly, D. R. Possible deep fault slip preceding the 2004 Parkfield earthquake, inferred from detailed observations of tectonic tremor. *Geophys. Res. Lett.* **36**, L17318, doi:10.1029/2009GL039589 (2009).
7. Nadeau, R. M. & Guilhem, A. Nonvolcanic tremor evolution and the San Simeon and Parkfield, California earthquakes. *Science* **325**, 191–193 (2009).
8. Sieh, K. E. Central California foreshocks of the great 1857 earthquake. *Bull. Seismol. Soc. Am.* **68**, 1731–1749 (1978).
9. Page, B. M., Coleman, R. G. & Thompson, G. A. Overview: Late Cenozoic tectonics of the central and southern coast ranges of California. *Geol. Soc. Am. Bull.* **110**, 846–876 (1998).
10. Rubinstein, J. L., Shelly, D. R. & Ellsworth, W. L. in *New Frontiers in Integrated Solid Earth Science, International Year of Planet Earth* 287–314 (Springer, doi:10.1007/978-90-481-2737-5_8, 2010).
11. Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* **300**, 1942–1943 (2003).
12. Obara, K. & Hirose, H. Non-volcanic deep low-frequency tremors accompanying slow slips in the southwest Japan subduction zone. *Tectonophysics* **417**, 33–51 (2006).
13. Brudzinski, M. R. & Allen, R. M. Segmentation in episodic tremor and slip all along Cascadia. *Geology* **35**, 907–910 (2007).
14. Gomberg, J. *et al.* Widespread triggering on non-volcanic tremor in California. *Science* **319**, 713 (2008).
15. Peng, Z. *et al.* Strong tremor near Parkfield, CA excited by the 2002 Denali earthquake. *Geophys. Res. Lett.* **35**, doi:10.1029/2008GL036080 (2008).
16. Peng, Z., Vidale, J. E., Wech, A., Nadeau, R. M. & Creager, K. M. Remote triggering of tremor around the Parkfield section of the San Andreas fault. *J. Geophys. Res.* **114**, B00A06, doi:10.1029/2008JB006049 (2009).
17. Thomas, A., Nadeau, R. M. & Bürgmann, R. Tremor-tide correlations and near-lithostatic pore pressure on the deep San Andreas fault. *Nature* **462**, 1048–1051 (2009).
18. Shelly, D. R., Beroza, G. C. & Ide, S. Non-volcanic tremor and low frequency earthquake swarms. *Nature* **446**, 305–307 (2007).
19. Johnston, M. J. S., Borcherdt, R. D., Linde, A. T. & Gladwin, M. T. Continuous borehole strain and pore pressure in the near field of the 28 September 2004 M 6.0 Parkfield, California, earthquake: implications for nucleation, fault response, earthquake prediction, and tremor. *Bull. Seismol. Soc. Am.* **96**, doi:10.1785/0120050822 (2006).
20. Katsumata, A. & Kamaya, N. Low-frequency continuous tremor around the Moho discontinuity away from volcanoes in the southwest Japan. *Geophys. Res. Lett.* **30**, doi:10.1029/2002GL015981 (2003).
21. Gibbons, S. J. & Ringdal, F. The detection of low magnitude seismic events using array-based waveform correlation. *Geophys. J. Int.* **165**, 149–166 (2006).
22. Shelly, D. R., Beroza, G. C. & Ide, S. Complex evolution of transient slip derived from precise tremor locations in western Shikoku, Japan. *Geochem. Geophys. Geosyst.* **8**, Q10014, doi:10.1029/2007GC001640 (2007).
23. Ozacar, A. A. & Zandt, G. Crustal structure and seismic anisotropy near the San Andreas fault at Parkfield, California. *Geophys. J. Int.* **178**, 1098–1104 (2009).
24. Smith, E. F. & Gomberg, J. A search in strainmeter data for slow slip associated with triggered and ambient tremor near Parkfield, California. *J. Geophys. Res.* doi:10.1029/2008JB005848 (in the press).
25. Agnew, D. C. & Wyatt, F. K. Long-base laser strainmeters: a review. ⟨http://repositories.cdlib.org/sio/techreport/2/⟩ (SIO Technical Report, Scripps Institution of Oceanography, 2003).
26. Kirby, S., Wang, K. & Brocher, T. A possible deep, long-term source for water in the northern San Andreas fault system: a ghost of Cascadia subduction past? *Eos* **83**, (Fall Meet. Suppl.) abstract S22B–1038 (2002).
27. Trehu, A. M. & Wheeler, W. H. Possible evidence for subducted sedimentary materials beneath central California. *Geology* **15**, 254–258 (1987).
28. Sass, J. H. *et al.* Thermal regime of the San Andreas fault near Parkfield, California. *J. Geophys. Res.* **102**, 27575–27585 (1997).
29. Thurber, C. *et al.* Three-dimensional compressional wavespeed model, earthquake relocations, and focal mechanisms for the Parkfield, California, region. *Bull. Seismol. Soc. Am.* **96**, S38–S49, doi:10.1785/0120050825 (2006).
30. Murray, J. & Langbein, J. Slip on the San Andreas fault at Parkfield, California, over two earthquake cycles, and the implications for seismic hazard. *Bull. Seismol. Soc. Am.* **96**, S283–S303, doi:10.1785/0120050820 (2006).

METHODS

The event families and detected catalogue used here were introduced by ref. 6, in support of work describing possible changes in tremor activity preceding the 2004 magnitude-6.0 Parkfield earthquake. This processing consists of three main steps: template event selection, template event location and event detection. For convenience, these steps are summarized below.

Selection of waveform templates. The waveform templates used in this study are those used by ref. 6 and are selected from the continuous data. Templates are identified largely by a process of trial and error. To find a new template, I looked for tremor not matched by any current templates. Within this unmatched tremor, I attempted to select relatively strong and isolated events as templates.

To compare tremor rates over time, I must consider the possibility that tremor waveforms may change over periods of years, either due to a change in the source or a change in the earth (for example, changes induced by the San Simeon or Parkfield earthquakes). To test this possibility, I compared the event detection rates from similar waveform templates recorded during different years. I found no significant change over the 7.5 years of this study, and in general used the template that appears to have the highest signal-to-noise ratio.

Template event location. Tremor locations shown in Fig. 1 from ref. 6 are derived from a hybrid process. Whenever possible, I obtained an absolute location for each template event based upon P and S wave arrival times of a low-frequency earthquake with a good signal-to-noise ratio at permanent and temporary stations. These additional stations, shown as open black triangles in Fig. 1a, helped greatly to constrain the locations. To evaluate the range of possible hypocentres and their corresponding residuals, the locations were performed by a grid search (Supplementary Fig. 2) using a one-dimensional approximation to the velocity model of ref. 29 near this location, with a V_P/V_S ratio of 1.78 (where V_P and V_S are the velocities of the P- and S-waves), as done by ref. 4. Uncertainty in these locations was approximately ± 4 km in horizontal and vertical positions. Owing to station distribution, uncertainties tended to be slightly smaller for the northerly tremors (closer to Parkfield) and slightly larger for the southernmost tremors, for example, families 18–21. I obtained absolute locations for 11 of the 21 template events shown in Fig. 1; these coordinates are given in Supplementary Table 1 of ref. 6. Phase arrival time picks (P and S waves) for these events are given here in Supplementary Table 1. Using the absolute locations as a framework, I then refined these locations by observing persistent patterns of migration between event families. For example, if I consistently observed migration from family A to family B to family C, or the reverse of this, I inferred that family B is located between families A and C. I also used migration patterns to infer event locations for which no absolute location can be determined, owing to insufficient identifiable P and S wave arrivals. Because the depth range of single-event locations (22–30 km) can nearly be explained by uncertainty, I plotted all events in the range of 24–25 km, which represents an average depth of the best-constrained events. Small perturbations are plotted to help explain migration patterns and differences in recurrence behaviour, but these differences are neither well constrained nor critical to the analysis presented here.

Event detection. I used the detection catalogue introduced by ref. 6, obtained by a procedure adapted from previous work[18,22]. The waveform template consists of 6 s of the three-component waveform at each station, with time delays applied to approximately match the S-wave propagation across the network. An example template is shown in Supplementary Fig. 1. I cross-correlated each waveform template with the continuous data stream, preserving the time lags across stations. I used data of 20 samples per second, bandpass filtered at 2–8 Hz, and performed the cross-correlation at lag increments of one sample. Similarity is then measured by the correlation sum across all channels and stations of the network. Finally, a threshold was applied. For maximum consistency over time, I used a fixed correlation sum threshold of 4.0 for the 25 selected channels, with a maximum of one event every 4 s.

Comparison with a tremor catalogue determined by amplitude and duration (R. Nadeau, personal communication, 2008) shows good agreement, although I identified many weak or short-duration tremor episodes missing from the amplitude-based catalogue. Specifically, I compared my results with a catalogue of well-located events south of Parkfield for 2006–2007 (R. Nadeau, personal communication, 2008). Although additional templates would be required to match all tremor, I detected at least one event within 99% (305/308) of bursts identified by amplitude and duration. Individual template events match between several hundred and several thousand events over the 7.5 years examined here. Tremor also occurs along the San Andreas fault north of Parkfield[7] and additional template events would be necessary to detect this tremor as well as for additional tremor south of Parkfield.

Stations and components used for consistent detection. To compare tremor rates over time, I used the following stations and channels, which operated most consistently during the study period (STATION.CHANNELS): EADB.123, FROB.123, JCSB.123, LCCB.23, MMNB.3, RMNB.12, SCYB.123, SMNB.123, VCAB.123, GHIB.13 (two-component non-functional at GHIB). The following stations and channels were not used: CCRB.123, JCNB.123, LCCB.1, MMNB.12, RMNB.3, VARB.123 (instrument replaced at new depth in August 2003 at VARB).

Stations used for template event locations. Although the HRSN stations provide optimal long-term tremor detection capability, their concentrated distribution offset from the tremor activity does not provide good location ability. Therefore, I employed additional stations (shown in Fig. 1) to constrain the data. In particular, I used four temporary arrays of ten stations each (GFZ1, GFZ2, GFZ3, Vogel)[4] and six individual three-component stations from the Plate Boundary Observatory and HRSN (B079, B078, B900, B073, GHIB, FROB). Two additional stations from the Southern California Seismic Network (SMM and PHL) were used to help locate the southernmost event family (family 21).

LETTERS

Mammalian biodiversity on Madagascar controlled by ocean currents

Jason R. Ali[1] & Matthew Huber[2]

Madagascar hosts one of the world's most unusual, endemic, diverse and threatened concentrations of fauna[1]. To explain its unique, imbalanced biological diversity, G. G. Simpson proposed the 'sweepstakes hypothesis', according to which the ancestors of Madagascar's present-day mammal stock rafted there from Africa[2]. This is an important hypothesis in biogeography and evolutionary theory for how animals colonize new frontiers[1,3–5], but its validity is questioned[5–9]. Studies suggest that currents were inconsistent with rafting to Madagascar[9] and that land bridges provided the migrants' passage[5–8]. Here we show that currents could have transported the animals to the island and highlight evidence inconsistent with the land-bridge hypothesis. Using palaeogeographic reconstructions and palaeo-oceanographic modelling, we find that strong surface currents flowed from northeast Mozambique and Tanzania eastward towards Madagascar during the Palaeogene period, exactly as required by the 'sweepstakes process'. Subsequently, Madagascar advanced north towards the equatorial gyre and the regional current system evolved into its modern configuration with flows westward[10] from Madagascar to Africa. This may explain why no fully non-aquatic land mammals have colonized Madagascar since the arrival of the rodents and carnivorans during the early-Miocene epoch. One implication is that rafting may be the dominant means of overseas dispersal in the Cenozoic era when palaeocurrent directions are properly considered.

Madagascar is home to one of the most intriguing inventories of flora and fauna anywhere on Earth[1]. It is characterized by unusually high levels of mammalian endemism combined with a uniquely broad diversity from a limited number of orders: lemurs, tenrecs, carnivorans and rodents[1,11–13]. Although our focus here is primarily on mammals, similar patterns are observed for other terrestrial animals on the island, including amphibians and reptiles[14,15], implying a broadly similar colonization mechanism. It is widely acknowledged that Madagascar's mammals are derived from Cenozoic migrants because they share none of the detailed characteristics of the island's late-Cretaceous forms[16,17] and thus could not have evolved from them. The current stock's ancestors journeyed from Africa at various times during the Cenozoic, since 65 Myr ago, and each colonization appears to be the result of a single arrival event[1]. The timing of apparent arrival events of new taxa is a fundamental constraint on hypotheses about how they accomplished the journey[1]. From molecular-clock dating estimates, it is possible to discern four distinct early-Cenozoic to mid-Cenozoic events: the arrival of lemurs between 60 and 50 Myr ago, that of tenrecs between 42 and 25 Myr ago, that of carnivorans between 26 and 19 Myr ago, and that of rodents between 24 and 20 Myr ago[12,13]. Thus, by ~20 Myr ago the major non-volant and non-swimming faunal groups were established, with no further evidence of transfer, except for a few late-Quaternary arrivals such as the pygmy hippopotamus, which is semi-aquatic and known to swim significant ocean distances[18]. As

Madagascar and Africa have been separated for ~120 Myr by the 430-km-wide Mozambique Channel[19,20], the question is whether the non-volant, non-aquatic migrants made their way to the island by walking or by rafting.

One hypothesis is that land bridges might have enabled fauna to walk. It has been argued that substantial portions of the Davie Ridge—a prominent bathymetric feature running north–south down the middle of the Mozambique Channel (Fig. 1)—may have once been sub-aerially exposed and thus could have acted as a quasi-continuous causeway linking Africa and Madagascar[8]. Although this idea has received some support[5–7,21], important weaknesses are acknowledged because it would require a radical rethinking of regional plate tectonics[22]. It also suffers logical flaws: if land bridges were responsible, a greater variety of animals would have crossed and the timing of arrivals would be correlated with the putative maximum extent of the land bridge, neither of which results is supported by data[1,23]. We further note that any islands the Davie Ridge may have generated during the Cenozoic[23] would have been small and separated by open-water gaps several tens to hundreds of kilometres apart[24] (Fig. 1). Consequently, substantial tracts of ocean (>230 km) separated them from the nearest land in eastern Mozambique and western Madagascar; thus, over-water dispersal was unavoidable.

The over-water dispersal mechanism was first mooted by Simpson nearly seventy years ago[2]. He proposed a 'sweepstakes' process by which small mammals—potentially with low metabolic rates and/or a habit of seasonal torpor—were unwittingly rafted to the island on large logs or vegetation mats washed off eastern Africa, either down large rivers or from the coastal strip[25]. Key predictions of Simpson's argument, confirmed by the colonization-history and geological data described above, are the limited number of families that live on the island today, in particular the absence of large-bodied forms (for example antelopes, apes, elephants or lions), and the seemingly random distribution of apparent arrivals in Madagascar from ~60 to ~20 Myr ago.

The main criticism of this hypothesis is that inferred currents and prevailing winds, based on modern observations, are in the opposite direction to those required. As elucidated in ref. 9, if today's surface-water currents in the region are used as a guide (Fig. 1), the strong south-southwest-directed coast-parallel flow of the Mozambique Current would have acted as barrier to eastward transport. Rafts off the shore of Africa would have been entrained in the southward flow and thus could never have beached on Madagascar. Instead they would either return to the African shore or be transported north or south, but never substantially to the east. No quantitative attempts have been made to estimate how currents in the region may have been different in the Cenozoic or what implications this may have for ocean dispersal routes. Past currents remain the major unknown in this controversial issue and many factors, including changing palaeogeographic and palaeo-oceanographic setting, must be considered.

---

[1]Department of Earth Sciences, University of Hong Kong, Pokfulam Road, Hong Kong, China. [2]Earth and Atmospheric Sciences Department and the Purdue Climate Change Research Center, Purdue University, West Lafayette, Indiana 47907, USA.
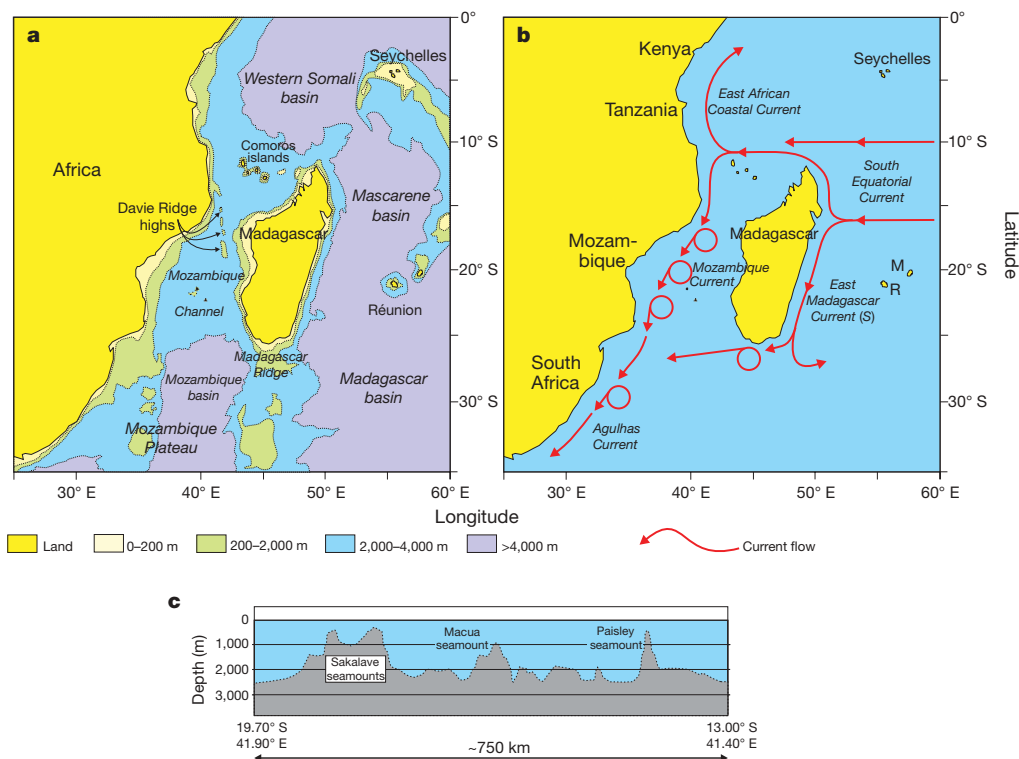
Figure 1 | Key geographical features/phenomena of the southwest Indian Ocean region. a, Simplified bathymetric map. **b**, Present-day surface-water circulation based on refs 25 and 26 (M, Mauritius; R, Réunion). **c**, Bathymetric cross-section along the Davie Ridge. Even in the Eocene, when parts of the feature may have been sub-aerial, the deep troughs[23] separating the peaks would have posed formidable barriers[22].

The Indian Ocean basin has altered considerably as a result of plate tectonics. For instance, since 60 Myr ago Australia and India have respectively migrated ~2,200 and ~4,000 km northwards. Globally, six major ocean gateways have either opened (Tasman–Antarctica, South America/Antarctica, Atlantic–Arctic) or closed (Panama Isthmus, Indonesian, Africa–Arabia–Eurasia). Critically, Africa and Madagascar have both moved some 15° (~1,650 km) closer to the equator[20]. Thus, the potential for surface-water flow in the southwest Indian Ocean to have been markedly different when the faunas are thought to have crossed to Madagascar must be considered.

Today, the entire southern Indian Ocean, from Africa to Australia, lies within a large anticyclonic 'supergyre'[26]. Northern Madagascar lies at the boundary between this gyre and a cyclonic gyre occupying the northern Indian Ocean[26,27]. The southern Indian Ocean gyre's strongest flows are to the southwest of Madagascar, where they form the Agulhas current. A relatively strong eddying flow in the Mozambique Channel between Madagascar and Africa connects the gyre's equatorial currents with the vigorous transports of the Agulhas system[27]. Another important feature of the southern Indian Ocean gyre concerns the fact that Australia spans the latitudes Madagascar does. As a consequence, the zonally integrated wind stress curl, which drives the circulation, is maximized near the southern tip of Africa[27].

To test the hypothesis that changes in the ocean circulation around Madagascar during the Cenozoic explain the observed pattern of land-mammal migrations to the island, an independent knowledge of surface ocean currents and surface wind stress is required. The only means of obtaining this information (as there are no independent proxy records for these variables) is from palaeo-oceanographic modelling. We performed and analysed a suite of experiments using a fully and interactively coupled ocean–atmosphere general circulation model (the Community Climate System Model, version 3 (CCSM3), of the US National Center for Atmospheric Research) with Eocene conditions (Methods). In such models, climate, wind and surface-water currents are predicted and, importantly, are free to evolve according to the equations of motion and thermodynamics. Thus, we have not specified the winds or sea surface temperatures that drive them, nor the ocean currents.

In all of our simulations, the large-scale ocean current systems in the Eocene epoch were robustly different from modern observed and modelled circulations in several crucial ways relevant to the rafting problem. Africa and Madagascar were >10° south of their current positions, which placed Madagascar in the convergence zone at the heart of the subtropical gyre (Fig. 2). Furthermore, because Australia was also much further south, the northern part of that continent did not impede the accumulation of wind stress curl at the latitudes of Madagascar; therefore, the strongest current system in the entire Southern Ocean was just to the east of Madagascar (Fig. 2a). This vigorous eddy off the coast of Madagascar caused water-mass trajectories throughout the region to converge.

As shown in Fig. 2a, the strong anticlockwise gyre directed much of the flow along the African coast eastward towards Madagascar rather than southward through the Mozambique Channel, towards the Agulhas Current, as occurs today. Peak simulated eastward velocities across the Mozambique Channel usually occur in January (Fig. 2b), and monthly mean velocities of >10 cm s^{-1} occur for three to four months within each century of simulation. Trajectories starting in the region of northeast Mozambique and Tanzania sporadically experience enhanced eastward velocities of >20 cm s^{-1} and could therefore have crossed the necessary distance in 25–30 days (Fig. 2c). This vigorous eastward flow was not constant, occurring for only three or four weeks within any century of simulation. Hence, on the long timescales of relevance to this problem (tens of millions of years), these velocities would have occurred many times, but not routinely. Because these are rare events, it is likely that even faster eastward currents occurred, albeit less frequently. Furthermore, flow through the Mozambique Channel was probably strongly eddying, as it is today, but our simulations do not have sufficient resolution to capture the vigorous, high-kinetic-energy properties of these eddies, which today have velocities of ~100 cm s^{-1} (ref. 27). Capturing such mesoscale variability would
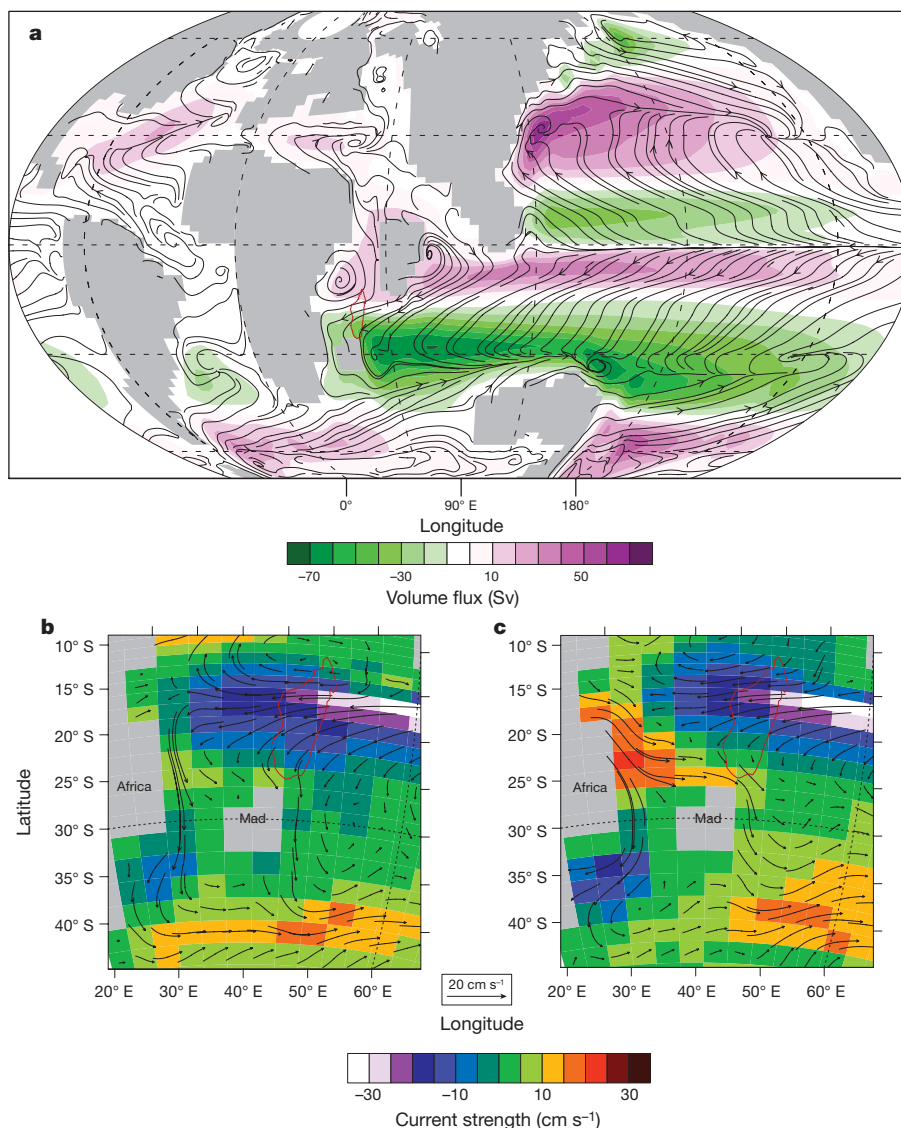
Figure 2 | Eocene ocean currents. a, Simulated annual mean vertically averaged (barotropic) currents (volume flux) and surface ocean currents (streamline vectors). The barotropic currents delineate the average positions of the major Eocene ocean gyres, with Madagascar at the heart of the strongest gyre on Earth, as described in the text. The modern location of Madagascar is shown in red outline, showing how the interplay of continental position and the gyres controls dispersal pathways. **b,** Ensemble-averaged monthly mean ocean surface currents for January, the month in which climatological ocean current directions were optimal for transport towards Madagascar. **c,** During sporadic events, as typified by this ensemble average of the four optimal ocean current events evaluated from model output saved at a temporal resolution of three days, rapid transport directly to Madagascar from Africa was possible at rates of >20 cm s^{-1}. Currents are shown using vectors (with the scale shown), and the magnitude of the east–west current strength is shown in colour. The simulations are described in Methods. Mad, Madagascar.

only increase the estimated probability of the occasional very vigorous eastward transport from Africa to Madagascar, and would enhance transport across the Mozambique Channel in jet-like currents. Consequently, our speed estimate is almost certainly lower than the true maximum eastward rafting velocity.

An additional consideration is the fact that tropical storms are known to generate the large, floating tree 'islands', as well as associated precipitation, that might make a successful ocean voyage of this type possible[4]. It is therefore noteworthy that preliminary analysis of modelled tropical cyclone activity indicates that this region was a locus of such activity, as it is today, and that the tropical cyclone season encompasses the period of highly favourable ocean currents (January). Thus, successful rafting may have involved the fortuitous coincidence of transient storms and ocean current activity.

Thus, all signs point to the Simpson sweepstakes model being correct: ocean currents could have occasionally transported rafts of animals to Madagascar from Africa during the Eocene. Specifically, transport should have been from northeast Mozambique and

Tanzania to the north coast of Madagascar. Given the slow tectonic drift of the island, this configuration probably continued at least through the Oligocene epoch. However, by the early Miocene, Madagascar breached the margin of the subtropical and equatorial gyres. Thereafter, currents were perennially directed westward towards Africa, making the ocean journey for mammals to Madagascar much more difficult, if not impossible.

METHODS SUMMARY

We used the current version of the fully coupled (ocean/atmosphere/land/vegetation/sea ice) global climate model CCSM3. The atmospheric resolution was set at $\sim 3.75° \cdot 3.75°$ (T31 spectral resolution); the oceans had a nominal 3° longitudinal resolution (and variable latitudinal resolution) and 25 vertical levels. This model has been applied to a wide variety of modern[28] and palaeo-climate studies, for example of the Holocene epoch and the Last Glacial Maximum[29] and the Eocene[30].

We carried out a suite of fully coupled simulations for a simulated time of more than 3,000 yr without any acceleration until they had clearly reached equilibrium. A description of the suite of simulations is found in ref. 30; results from

the particular simulation, with a concentration of atmospheric CO_2 of 1,120 p.p.m., used in this study have not been previously described. Here we have concentrated on one simulation appropriate for mid-Eocene to late-Eocene conditions, although none of the results we have discussed are sensitive to the particulars of that choice, as the relevant boundary conditions and main palaeo-currents remained largely unchanged between the end of the Cretaceous period and the early-to-mid Miocene, when current systems shifted towards their modern state[10].

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

1. Yoder, A. D. & Nowak, M. D. Has vicariance or dispersal been the predominant biogeographic in Madagascar? Only time will tell. *Annu. Rev. Ecol. Evol. Syst.* **37,** 405–431 (2006).
2. Simpson, G. G. Mammals and land bridges. *J. Wash. Acad. Sci.* **30,** 137–163 (1940).
3. Heaney, L. R. Is a new paradigm emerging for oceanic island biogeography? *J. Biogeogr.* **34,** 753–757 (2007).
4. Thiel, M. & Haye, P. A. The ecology of rafting in the marine environment. III. Biogeographical and evolutionary consequences. *Oceanogr. Mar. Biol.* **44,** 323–429 (2006).
5. Tattersall, I. in *Elwyn Simons: A Search for Origins* (eds Fleagle, J. G. & Gilbert, C. C.) 397–408 (Springer, 2008).
6. Tattersall, I. Historical biogeography of the strepsirhine primates of Madagascar. *Folia Primatol. (Basel)* **77,** 477–487 (2006).
7. Masters, J. C., de Wit, M. J. & Asher, R. J. Reconciling the origins of Africa, India and Madagascar with vertebrate dispersal scenarios. *Folia Primatol. (Basel)* **77,** 399–418 (2006).
8. McCall, R. A. Implications of recent geological investigations of the Mozambique Channel for the mammalian colonization of Madagascar. *Proc. R. Soc. Lond. B* **264,** 663–665 (1997).
9. Stankiewicz, J., Thiart, C., Masters, J. C. & de Wit, M. J. Did lemurs have sweepstake tickets? An exploration of Simpson's model for the colonization of Madagascar by mammals. *J. Biogeogr.* **33,** 221–235 (2006).
10. von der Heydt, A. & Dijkstra, H. A. Effect of ocean gateways on the global ocean circulation in the late Oligocene and the early Miocene. *Paleoceanography* **21,** PA1011 (2006).
11. Goodman, S. M., Ganzhorn, J. U. & Rakotondravony, D. in *The Natural History of Madagascar* (eds Goodman, S. M. & Benstead, J. P.) 1159–1186 (Chicago Univ. Press, 2003).
12. Yoder, A. D. *et al.* Single origin of Malagasy Carnivora from an African ancestor. *Nature* **421,** 734–737 (2003).
13. Poux, C. *et al.* Asynchronous colonization of Madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes. *Syst. Biol.* **54,** 719–730 (2005).
14. Vences, M. Origin of Madagascar's extant fauna: a perspective from amphibians, reptiles and other non-flying vertebrates. *Ital. J. Zool. (Modena)* **71** (suppl.), 217–228 (2004).
15. Nagy, Z. T., Joger, U., Wink, M., Glaw, F. & Vences, M. Multiple colonization of Madagascar and Socotra by colubrid snakes: evidence from nuclear and mitochondrial gene phylogenies. *Proc. R. Soc. Lond. B* **270,** 2613–2621 (2003).
16. Krause, D. W. Fossil molar from a Madagascan marsupial. *Nature* **412,** 497–498 (2001).
17. Krause, D. W. *et al.* Late Cretaceous terrestrial vertebrates from Madagascar: implications for Latin American biogeography. *Ann. Mo. Bot. Gard.* **93,** 178–208 (2006).
18. Stuenes, S. Taxonomy, habits, and relationships of the subfossil Madagascan hippopotami *Hippopotamus lemerlei* and *H.madagascariensis. J. Vertebr. Paleontol.* **9,** 241–268 (1989).
19. Rabinowitz, P. D., Coffin, M. F. & Falvey, D. The separation of Madagascar and Africa. *Science* **220,** 67–69 (1983).
20. Ali, J. R. & Aitchison, J. C. Gondwana to Asia: plate tectonics, paleogeography and the biological connectivity of the Indian sub-continent from the Middle Jurassic through latest Eocene (166–35 Ma). *Earth Sci. Rev.* **88,** 145–166 (2008).
21. Godinot, M. Lemuriform origins as viewed from the fossil record. *Folia Primatol. (Basel)* **77,** 446–464 (2006).
22. Rabinowitz, P. D. & Woods, S. The Africa–Madagascar connection and mammalian migrations. *J. Afr. Earth Sci.* **44,** 270–276 (2006).
23. Bassias, Y. Petrological and geochemical investigations of rocks from the Davie Fracture Zone (Mozambique Channel) and some tectonic implications. *J. Afr. Earth Sci.* **15,** 321–339 (1992).
24. Krause, D. W., Hartman, J. H. & Wells, N. A. in *Natural Change and Human Impact in Madagascar* (eds Goodman, S. D. & Patterson, B. D.) 3–43 (Smithsonian Inst. Press, 1997).
25. Kappeler, P. M. Lemur origins: rafting by groups of hibernators? *Folia Primatol. (Basel)* **71,** 422–425 (2000).
26. Schott, F. A., Xie, S. P. & McCreary, J. P. Jr. Indian Ocean circulation and climate variability. *Rev. Geophys.* **47,** RG1002 (2009).
27. de Ruijter, W. P. M., Ridderinkhof, H. & Schouten, M. Variability of the southwest Indian Ocean. *Phil. Trans. R. Soc. A* **363,** 63–76 (2005).
28. Yeager, S. G., Shields, C. A., Large, W. G. & Hack, J. J. The low-resolution CCSM3. *J. Clim.* **19,** 2545–2566 (2006).
29. Otto-Bliesner, B. L. *et al.* Last glacial maximum and Holocene climate in CCSM3. *J. Clim.* **19,** 2526–2544 (2006).
30. Liu, Z. *et al.* Global cooling during the Eocene-Oligocene climate transition. *Science* **323,** 1187–1190 (2009).

Author Contributions J.R.A. initiated the study and was primarily responsible for the geologically related aspects. M.H. carried out the palaeo-oceanographic modelling and its interpretation. Both authors contributed to the writing of 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 M.H. (huberm@purdue.edu).

METHODS

Our simulations incorporate the detailed boundary conditions (that is, topography, vegetation and bathymetry) developed previously for Eocene conditions[30–32]. The latitudinal position of the gyres and their strengths are very robust to boundary-condition changes appropriate to the Palaeogene as they are essentially Sverdrupian responses to wind stress changes as modulated by the slowly varying palaeogeographies[31–33].

As described in ref. 9, the rafting problem involves estimating the surface ocean currents with the potential additional influence of surface winds. Currents below the surface are not relevant because typical rafts, such as a tree trunk launched into the ocean from a river mouth during a flood, would not have 'keels' that extended below several metres. Similarly, it is unlikely they had substantial protruding elements that could have formed 'masts', and would thus have been unaffected by wind shear. Here we assume that the raft simply acted like a drifter embedded in the ocean surface current (the surface model level extends down to 8 m). Analysis of the modern simulation using this same model (not shown) produces ocean currents in agreement with observations and the prior work on Madagascar[9].

The Eocene simulation focused on in this study was carried out with an atmospheric CO_2 concentration of 1,120 p.p.m. and all other boundary conditions set at near-modern, pre-industrial values. It is a branch simulation from a simulation with a higher concentration of atmospheric CO_2 that was integrated for a simulated time of more than 3,000 yr. After this spin-up time, the 1,120-p.p.m. simulation was integrated for 3,500 yr. We analysed the monthly mean output from the last 400 yr of simulation to generate Fig. 2a, b. Output at 3-d resolution for the final century of integration was used to generate Fig. 2c. Although the magnitude of the optimal eastward flow is somewhat dependent on the averaging period (peak velocity increases from $\sim$13 cm s^{-1} when monthly mean values are used to $\sim$23 cm s^{-1} when 3-d means are used), the qualitative aspects of the flow are not sensitive to averaging length. In the sampling of both the mean configuration and the extreme events, flow is always westward from Madagascar towards Africa because this is the direction determined by the large-scale gyres in the modern era. Preliminary results from a comparable Miocene (15 Myr ago) simulation show similar results to the modern-day one. All file processing and graphics were performed using the US National Center for Atmospheric Research Command Language (http://ncl.ucar.edu).

31. Huber, M., Sloan, L. C. & Shellito, C. J. in *Causes and Consequences of Globally Warm Climates in the Early Palaeogene* (eds Wing, S. L., Gingerich, P. D., Schmitz, B. & Thomas, E.) 25–47 (GSA Special Paper 369, Geological Society of America, 2003).
32. Huber, M. & Nof, D. The ocean circulation in the southern hemisphere and its climatic impacts in the Eocene. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **231**, 9–28 (2006).
33. Huber, M. *et al.* Eocene circulation of the Southern Ocean: was Antarctica kept warm by subtropical waters? *Paleoceanography* **19**, PA4026 (2004).

Evidence for grid cells in a human memory network

Christian F. Doeller[1,2], Caswell Barry[1,3,4] & Neil Burgess[1,2]

Grid cells in the entorhinal cortex of freely moving rats provide a strikingly periodic representation of self-location[1] which is indicative of very specific computational mechanisms[2–4]. However, the existence of grid cells in humans and their distribution throughout the brain are unknown. Here we show that the preferred firing directions of directionally modulated grid cells in rat entorhinal cortex are aligned with the grids, and that the spatial organization of grid-cell firing is more strongly apparent at faster than slower running speeds. Because the grids are also aligned with each other[1,5], we predicted a macroscopic signal visible to functional magnetic resonance imaging (fMRI) in humans. We then looked for this signal as participants explored a virtual reality environment, mimicking the rats' foraging task: fMRI activation and adaptation showing a speed-modulated six-fold rotational symmetry in running direction. The signal was found in a network of entorhinal/subicular, posterior and medial parietal, lateral temporal and medial prefrontal areas. The effect was strongest in right entorhinal cortex, and the coherence of the directional signal across entorhinal cortex correlated with spatial memory performance. Our study illustrates the potential power of combining single-unit electrophysiology with fMRI in systems neuroscience. Our results provide evidence for grid-cell-like representations in humans, and implicate a specific type of neural representation in a network of regions which supports spatial cognition and also autobiographical memory.

Grid cells recorded in medial entorhinal cortex of freely moving rodents fire whenever the animal traverses the vertices of an equilateral triangular grid covering the environment (see Fig. 1a), and may provide a neural substrate for path integration[1–3,6]. However, it is not known whether or not grid cells exist in humans, or how widespread the network of neurons with grid-like firing is, although the pre- and parasubiculum[7] and posterior parietal cortex[8] have been implicated.

Could grid-cell firing be detected in the functional magnetic resonance imaging (fMRI) signal, which reflects changes in metabolic activity across thousands of individual neurons[9]? The grid patterns of neighbouring cells are offset so as to 'tile' the environment[1], making systematic variation in population activity with position unlikely. However, three factors may allow detection using fMRI. The first factor is that the angular orientation of the grids relative to the environment appears to be constant across cells, whether they are neighbouring[1] or farther apart[5] (Supplementary Fig. 1), and rotate coherently when salient distal cues are rotated[1,6]. The other two factors result from new analyses presented here. The second factor being that firing of 'conjunctive' grid cells, found in the deeper layers of entorhinal cortex[6] and in pre- and parasubiculum[7], is modulated by running direction[6]. Here we show that the directions of modulation of conjunctive grid cells are aligned with the main axes of the grids (Fig. 1a–c). These first two factors will create systematic differences in neural population dynamics for runs aligned or misaligned with the main axes of the grids (Fig. 2). A third factor is running speed: the rate[6] and inter-burst frequency[10] of grid-cell firing, and the frequency with

which individual firing fields in a grid are sampled, increase with running speed. In addition, we show that the spatial organization of grid-cell firing is more strongly apparent during fast running than during slow running and immobility (Fig. 1d; see Methods for details and Supplementary Figs 2–6 for further analyses and recording locations). Thus, entorhinal grid cells form a coherent population in which the common effects of orientation and speed of movement could produce a macroscopic signal visible with fMRI.

We recorded whole-brain fMRI data while human participants navigated within a virtual reality arena, mimicking the foraging task
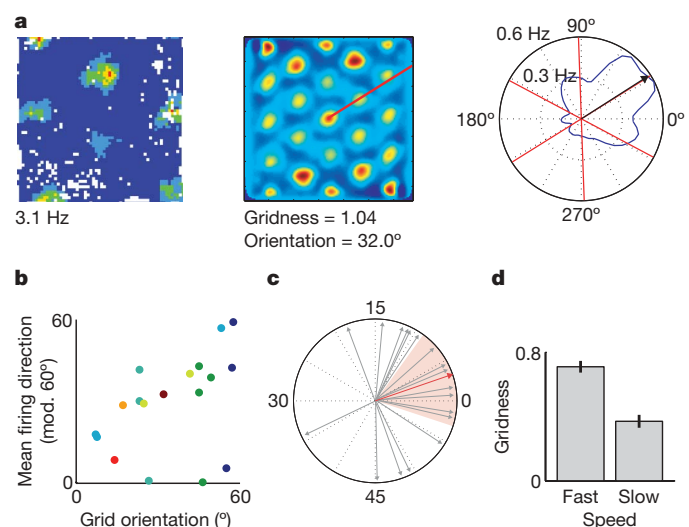

Figure 1 | The mean firing directions of directional grid cells are aligned with the grid. a, The left panel shows the firing rate map of a typical 'conjunctive' directional grid cell showing firing rate as a function of the rat's location within a 1 m² box (red, high firing rate; blue, low rate; white, unvisited location; peak rate 3.1 Hz). The middle panel shows the spatial autocorrelogram constructed from the rate map. The right panel shows the polar firing rate map for the same cell. The black arrow indicates mean firing direction. Red lines indicate the main axes of the grid firing pattern identified from the spatial autocorrelogram (see middle). **b**, Scatter plot of all directional grid cells (*n* = 18) showing grid orientation versus circular mean firing direction, modulo 60°. Cells from different rats (*n* = 8) are coloured differently. **c**, Angular difference between the circular mean firing direction of each cell and the nearest axis of its grid-like firing pattern is not distributed uniformly (Rayleigh test of uniformity; *P* = 0.007) and is significantly clustered around zero (Monte Carlo simulation; *P* < 0.001). The red arrow shows the mean difference (3.15°); red shaded area indicates the 95% confidence interval. **d**, The spatial organization of grid cell firing is less strongly apparent during slow movement and immobility than during fast movement. The bar graph shows mean gridness (a measure of six-fold spatial periodicity, see Methods) score for 113 grid cells, separately for fast and slow movements (median split); paired *t*-test, *P* = 2.2 × 10⁻¹¹. Error bars show s.e.m. over cells.

[1]UCL Institute of Cognitive Neuroscience, London WC1N 3AR, UK. [2]UCL Institute of Neurology, London WC1N 3BG, UK. [3]UCL Department of Cell and Developmental Biology, London WC1E 6BT, UK. [4]UCL Institute of Behavioural Neuroscience, University College London, London WC1H 0AP, UK.

in rodents (see Methods for details). The arena comprised a grassy plain bounded by a cylindrical cliff, containing a single intra-maze landmark and surrounded by a background scene providing orientation cues rendered at infinity (Fig. 2a). Participants collected and replaced objects found within the arena. Memory was measured as the proximity of the replacement location to the correct location (Supplementary Fig. 7 and ref. 11).

If orientationally aligned grid-cell firing is present, there should be an effect of running direction with six-fold rotational symmetry, reflecting the differences between running aligned or misaligned to the grid axes, and this effect should be modulated by running speed (Fig. 2b–d). We defined an anatomical region of interest (ROI) on the entorhinal cortices, and examined one-half of the fMRI data for modulation by running direction and speed to find the orientation of potential grids within each participant's entorhinal cortex. Briefly, we used a quadrature filter to identify the orientation of any six-peaked sinusoidal modulation of signal as a function of running direction and speed: providing the potential grid orientation in each voxel (Supplementary Fig. 8 and Methods). The potential grid orientations during fast runs were significantly clustered across voxels in the ROI in 34 out of 42 participants (Rayleigh test, accounting for spatial smoothing, on each participant; $P \ll 0.00001$ for the population, Monte Carlo simulation; Supplementary Fig. 9). The mean grid orientation for the entorhinal ROI (φ) was defined as the population vector (vector sum of orientations weighted by amplitude), and calculated for data corresponding to fast, medium and slow runs.

We then examined the second half of the data for differential activation for fast, medium and slow runs aligned versus misaligned to each participant's mean grid orientation. We looked for sinusoidal modulation of activation with six-fold rotational symmetry aligned with each participant's mean grid orientation across the whole brain
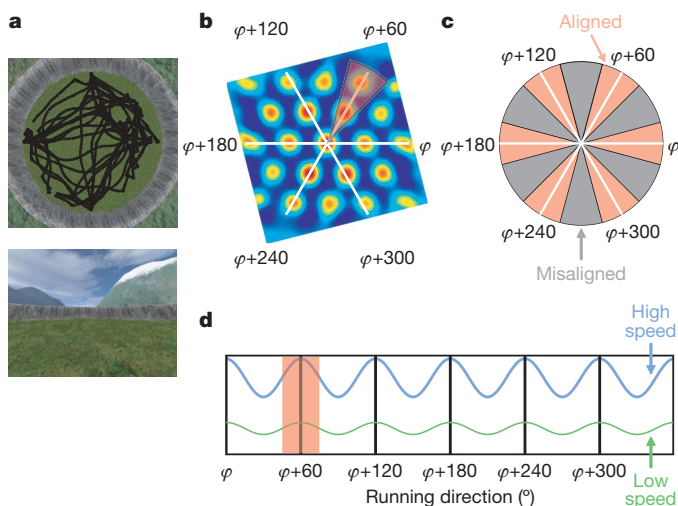

Figure 2 | fMRI: virtual reality arena and experimental logic. a, Human participants ($n = 42$) explored a circular virtual reality environment, bounded by a cliff and surrounded by orientation cues (mountains), finding objects and having to replace them in the correct locations. Top panel: aerial view, including one participant's virtual trajectory (black line). Bottom panel: participant's view. **b,** Spatial autocorrelogram of a typical grid cell showing the three main axes of the grid (white lines) and a 30° sector aligned with the grid (red). φ indicates mean grid orientation to which the grid shown is aligned. **c,** Schematic of running directions aligned (red) and misaligned (grey) with the grid. **d,** Given the alignment of directionally modulated grid cells with the grid and the constant grid orientation across cells (Fig. 1b), we predicted a sinusoidal modulation of fMRI signal by running direction with six-fold rotational symmetry, and a stronger effect for faster (blue) than slower (green) runs (see Fig. 1d). Note that the absolute orientation of the pattern (mean grid orientation, φ) will not be known a priori.

(using the regressor $\cos(6[\theta(t) - \varphi])$ where $\theta(t)$ is running direction; Supplementary Fig. 8). We found a significant activation in the right entorhinal cortex for fast runs (Fig. 3a, b). Direct comparison of activation for fast runs aligned versus misaligned to each participant's mean grid orientation showed a similar effect (Fig. 3d). Similar sinusoidal modulations were present in left entorhinal cortex, although slightly less reliably than on the right, and in the mean response of the entorhinal ROI (Supplementary Fig. 10). No activation elsewhere in the brain reached the significance level for this whole-brain analysis. Equivalent analyses for medium or slow runs, or looking for four- or eight-fold (Fig. 3b, c), or even five- or seven-fold, rotational symmetry (Supplementary Fig. 11) failed to show significant activation. No significant activation was found when the analyses applied to the entorhinal ROI were performed to control ROIs in the posterior right hippocampal location associated with memory performance in this task[11] or the visual cortical location showing adaptation to running direction (below) (see Supplementary Fig. 12). Supplementary Figs 13–15 show control analyses of behaviour in rats and humans.

Consistent with grid cells, the mean grid orientations in the six-fold symmetry model varied randomly between participants, ruling out a role for specific visual features of the environment (Supplementary Fig. 16). Notably, the coherence of the potential grid orientations in each participant's right entorhinal cortex was significantly correlated with that participant's spatial memory performance (Fig. 3f). This parallels findings that entorhinal cortex activation can predict memory performance[12,13], and provides a first indication that grid-like representations might usefully guide behaviour in mammals.

An independent indicator of variation in the dynamics of neuronal activity is provided by fMRI adaptation[14–17], that is, reduced signal from activation of a representation that was recently active compared to a previously inactive one. During runs aligned with the grid, a smaller proportion of grid cells will fire but will do so at a higher rate, whereas more grid cells will be active during misaligned runs but at a lower rate (Supplementary Fig. 17). Given the complex nonlinear dynamical relationship between neuronal activation and fMRI adaptation[9] we predicted that differences in the dynamics of neuronal activity when running aligned to the grids compared to misaligned would cause differences in fMRI adaptation, and that these differences would also vary with running speed.

We first checked whether the more basic effect of fMRI adaptation to running direction occurred within our virtual navigation paradigm. We made separate adaptation regressors (the logarithm of the time since last running in the current direction) for fast, medium and slow runs. We found significant adaptation to running direction in visual cortex extending into posterior parahippocampal/lingual and retrosplenial cortices, with larger effects for fast than for slow runs (see Fig. 4a, b). The adaptation effect is independent of variations in angular velocity (see Methods) and is consistent with adaptation to the background visual scene (which is rendered at infinity; viewing direction is the same as running direction), as previously found with static scenes in similar areas[14,16]. It is also consistent with view-specific single-unit responses in humans[18] or with the presence of head-direction cells[19], which have been reported in retrosplenial cortex[20].

Given the robust fMRI adaptation to virtual running direction, we tested our specific prediction that adaptation in entorhinal cortex is modulated by running direction with 60° periodicity and by running speed. Accordingly we looked for regions showing a reduction in fMRI signal according to how recently participants were running at 60° to the current direction, which was greater for fast than for slow running. Unlike our analysis of whole-brain activation as a function of mean grid orientation in the entorhinal ROI (Fig. 3), this analysis can be directly applied to the whole brain without reference to specific ROIs. As predicted, right entorhinal cortex (extending into subiculum) showed 60° shifted directional adaptation, independent of any effect of basic directional adaptation, as did regions in posterior parietal, lateral temporal and medial prefrontal cortices (see Fig. 4c). Medial
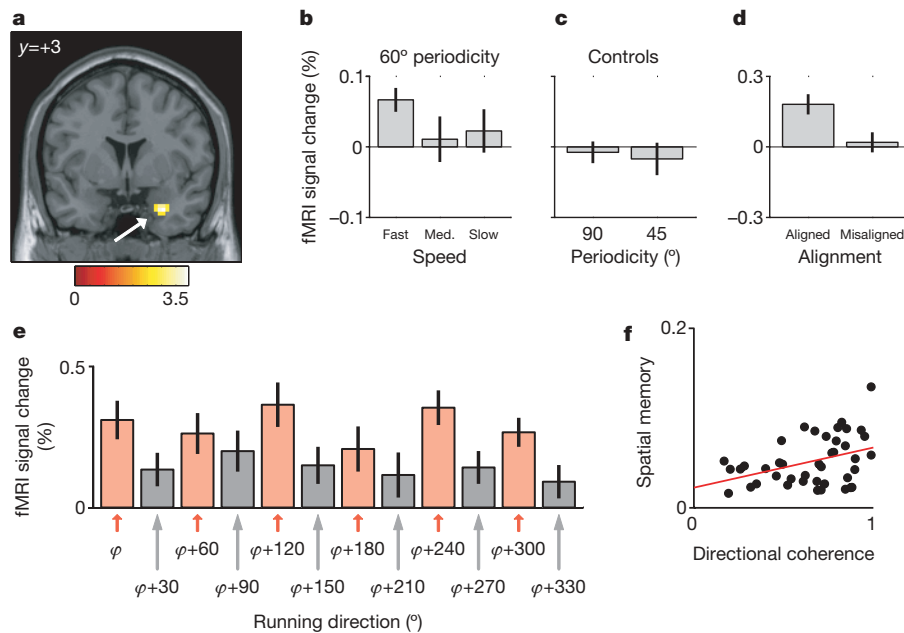
Figure 3 | Modulation of entorhinal cortical activity by running direction with six-fold rotational symmetry, and correlation with spatial memory. **a**, Sinusoidal modulation of activity by running direction with six-fold rotational symmetry. The orientation of potential grids in each participant's entorhinal cortex was estimated on one-half of the data, and the correspondingly aligned sinusoidal regressor was fitted to the other half of the data, showing a significant modulation only in right entorhinal cortex. Plot shows fMRI activation for fast runs on an aligned structural template; colour bar indicates t-statistic; all reported effects of whole-brain analyses are significant at $P < 0.001$ uncorrected; the t-image is thresholded at $P < 0.01$ for display purposes; peak Montreal Neurological Institute (MNI) coordinates: 30/3/−30; peak z-score = 3.59. **b**, Directional modulation depends on running speed, being present for fast runs, but not for medium (Med.) or slow runs. **c**, Directional modulation during fast runs has $60°$ directional periodicity, not $90°$ or $45°$ (that is, six-fold rather than four- or eight-fold rotational symmetry). Bar plots in **b** and **c** show the mean amplitude of sinusoidal modulation for the peak voxel in **a**. **d**, Activation for aligned (within $15°$ of the main axes of the grid, see Fig. 2c) and misaligned fast runs relative to baseline (epochs of no movement in the environment) in the peak voxel shown in **a**, confirming the sinusoidal modulation effect. **e**, To examine the pattern underlying these effects, we plotted the average fMRI signal over the entire time series of all voxels in the entorhinal ROI for all directions of aligned (red) and misaligned (grey) fast runs, relative to baseline (see Fig. 2c). **f**, The coherence of the potential grid orientations in each participant's right entorhinal cortex (mean length of resultant direction vector) correlated significantly with that participant's spatial memory (1/mean distance of object replacement locations from correct locations; range = 7.4–61.7 virtual metres; mean = 25.4, Spearman's rho = 0.32, $P = 0.039$). Each dot represents one participant. All bars show mean and s.e.m. over participants.

parietal cortex also showed $60°$ shifted directional adaptation, in addition to the basic directional adaptation described above, consistent with the presence of both visual/head-direction and grid-like representations (see Supplementary Fig. 18). The adaptation effect in all of these regions was strongly modulated by the participants' virtual running speed, being higher for fast compared to slow runs (Fig. 4d); the effect was specific to $60°$ rather than $45°$ or $90°$ shifted directional adaptation (Fig. 4e).

Do these remote regions showing $60°$ shifted directional adaptation also show the sinusoidal modulation of activation seen in entorhinal cortex during fast running, consistent with the presence of grid-like representations in these regions too? We examined the mean response to the sinusoidal regressor defined in entorhinal cortex in 8-mm-radius ROIs centred on the locations of peak adaptation. The responses in medial prefrontal (one-tailed t-test; $P = 0.008$), medial parietal ($P = 0.002$) and lateral temporal (left, $P = 0.019$; right, $P = 0.0005$) ROIs, but not the posterior parietal ROI ($P = 0.058$), reached significance for this specific prediction. Overall, these results suggest a network of regions containing coherently aligned neural representations with six-fold rotational symmetry, although the responses in remote regions were weaker than in right entorhinal cortex (for example, not reaching significance in the initial whole-brain split-half analysis), possibly reflecting a lower concentration of grid-like cells.

Our results provide the first evidence that human entorhinal cortex encodes virtual movement direction with six-fold symmetry, consistent with a coherently-oriented population of grid cells similar to those found in rat entorhinal cortex[1,5,6] and pre- and parasubiculum[7]. The dependence of directional modulation on running speed is consistent with the effects of speed on the firing rate[6,10] and apparent

spatial organization (Fig. 1d) of grid cells, whereas effects of speed on other aspects of behaviour may also contribute (see for example, Supplementary Fig. 14). The relationship to spatial memory of the directional coherence of potential grids across entorhinal cortex provides a first indication that grid-like representations might usefully guide behaviour.

Because we can only measure effects of direction and speed (not location) in the fMRI signal, our findings could reflect the presence of grid cells, or movement-related responses from head direction[19], or 'conjunctive' directional grid[6,7] cells, if they form coherent populations whose firing has six-fold rotational symmetry. We showed that conjunctive grid cells from rat entorhinal cortex have such an organization (Fig. 1). Our finding of similar and aligned fMRI responses from subicular, posterior/medial parietal, lateral temporal, and medial prefrontal cortices indicate that populations with similar properties also exist elsewhere, a prediction directly testable in rodents (see also refs 7, 8). These results outline a circuit for navigation, consistent with suggestions that medial and lateral temporal, posterior and medial parietal and medial prefrontal areas cooperate to support spatial cognition[2,8,11,21,22], and implicate a specific type of underlying neural representation.

Our study illustrates the ability to infer neural representations in humans by using fMRI in conjunction with single-unit recording in behaving animals, promising a coherent understanding of behaviour at the neural and systems levels. The observed grid-like representations support spatial memory (Fig. 3f) and are found in a circuit of regions which markedly overlaps the network for autobiographical memory and imagery (for example, refs 21, 23, 24). These types of regularly repeating representation may provide a clue to the neural basis of autobiographical memory, perhaps encoding temporal as
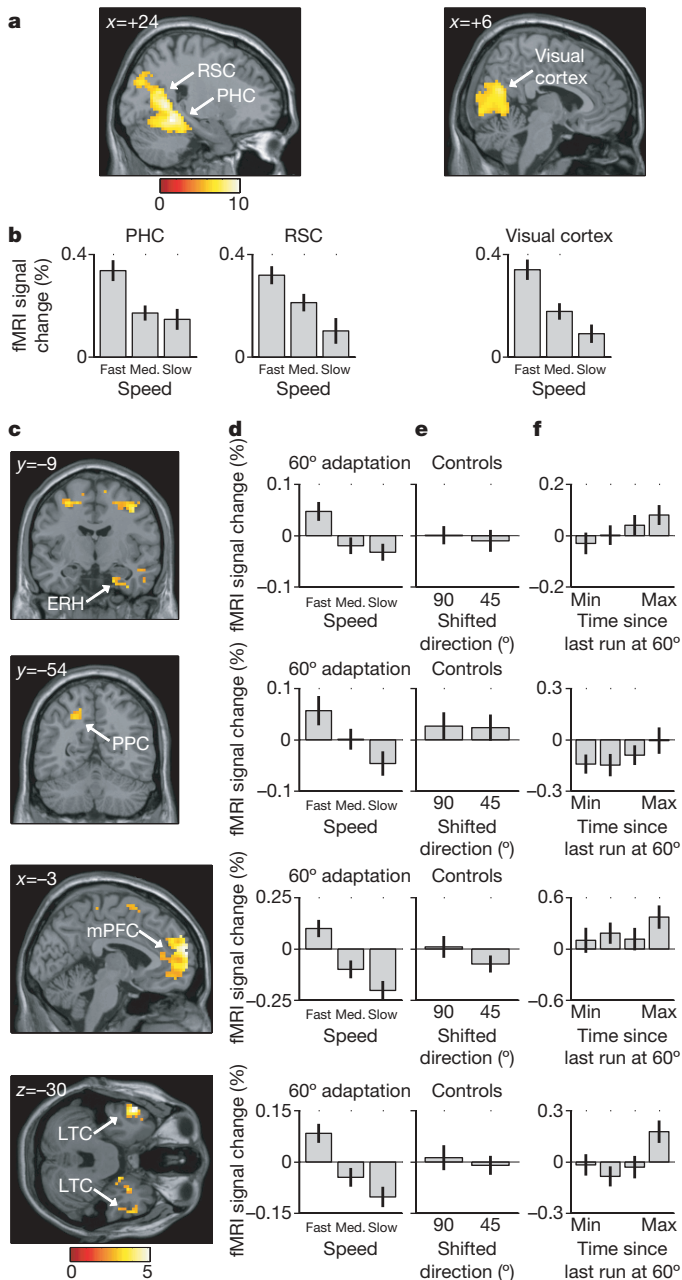
Figure 4 | fMRI adaptation to running direction and to runs at 60° from it. a, Activity in parahippocampal (PHC; 24/−48/-12; $z = 7.13$), retrosplenial (RSC; 18/−57/18; $z = 7.16$) and visual (peak at 18/−69/15; $z = 7.11$) cortices shows adaptation to absolute running direction (regressor: log[time since last run in current direction]). **b**, Adaptation is greater for faster runs, showing the adaptation effect for the peak voxels in the three regions for fast, medium (Med.) and slow runs. **c**, A speed-dependent fMRI adaptation to runs at 60° from the current direction (regressor: log[time since last run at 60° from the current direction]) is seen in a network of regions, including entorhinal cortex extending into subiculum (ERH; 21/−9/−30; $z = 3.28$), anterior entorhinal/perirhinal (33/0/−27; $z = 3.69$), posterior parietal (PPC; −18/−54/45; $z = 3.24$), medial prefrontal (mPFC; −3/63/15; $z = 4.96$), lateral temporal cortices (LTC; left: −54/9/−30; $z = 4.99$; right: 42/15/−36; $z = 3.48$) and precentral gyrus/superior frontal gyrus/motor cortex. These effects are independent of any basic (360°) directional adaptation (images are exclusively masked by the effects of basic directional adaptation at threshold $P < 0.05$, uncorrected). **d**, The 60° adaptation effect is specific to fast runs. **e**, No significant adaptation is seen for fast running at 45° or 90° from the current direction. **f**, fMRI activity as a function of time since last fast run at 60° from the current direction (log[time] binned in quartiles), illustrating the adaptation effect. Signal in **d–f** is shown for the peak voxels in the four regions. All effects significant at $P < 0.001$, uncorrected; for display purposes, t-images are thresholded at $P < 0.000001$ in **a** and $P < 0.01$ in **c**. Error bars show mean and s.e.m. over participants.

well as spatial context[3,4,25–27] for combination with parallel networks representing non-spatial information[25,28–30].

METHODS SUMMARY

Single-unit recording experiment in rats. Single-unit recordings were made from the medial entorhinal cortex of 17 male, Lister hooded rats, in accordance with appropriate Home Office Project and Personal licences. Each animal received a single microdrive carrying four tetrodes. While animals foraged for sweetened rice in a familiar environment, their positions and concomitant action potentials were captured using a standard procedure[5]. Grid cells were identified on the basis of their hexagonal spatial firing pattern, the regularity of which was assessed using a gridness measure[1,5]. Post-recording, animals were killed, perfused with phosphate-buffered saline and then with 4% paraformaldehyde. Histology was conducted to confirm recording locations. See Methods and ref. 5 for details.

Functional neuroimaging experiment in humans. Forty-two male participants took part in this experiment (aged 18–31, mean age 23.1 years). The study was approved by the local Research Ethics Committee. Functional images were acquired on a 3T Siemens Allegra scanner using a gradient-echo echo-planar imaging (EPI) pulse sequence. Imaging data were analysed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/), including standard pre-processing procedures (spatial realignment, correction for differences in slice acquisition time, spatial normalization and smoothing) and modelling the data by a set of general linear models. During the virtual reality task, participants were presented with a first-person perspective of a grassy plane surrounded by a circular cliff. Participants moved the viewpoint by using their right hand to operate keys to move forward and turn left or right. Throughout the experiment, participants collected and replaced everyday objects within the arena. (See Methods and ref. 11 for details.)

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

1. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
2. McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. & Moser, M. B. Path integration and the neural basis of the 'cognitive map'. *Nature Rev. Neurosci.* **7**, 663–678 (2006).
3. Burgess, N., Barry, C. & O'Keefe, J. An oscillatory interference model of grid cell firing. *Hippocampus* **17**, 801–812 (2007).
4. Hasselmo, M. E. A model of episodic memory: mental time travel along encoded trajectories using grid cells. *Neurobiol. Learn. Mem.* **92**, 559–573 (2009).
5. Barry, C., Hayman, R., Burgess, N. & Jeffery, K. J. Experience-dependent rescaling of entorhinal grids. *Nature Neurosci.* **10**, 682–684 (2007).
6. Sargolini, F. *et al.* Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**, 758–762 (2006).
7. Boccara, C. N. *et al.* Grid cells in presubiculum and parasubiculum. *FENS Abstr.* **4**, 128.21 (2008).
8. Whitlock, J. R., Sutherland, R. J., Witter, M. P., Moser, M. B. & Moser, E. I. Navigating from hippocampus to parietal cortex. *Proc. Natl Acad. Sci. USA* **105**, 14755–14762 (2008).
9. Logothetis, N. K. What we can do and what we cannot do with fMRI. *Nature* **453**, 869–878 (2008).
10. Jeewajee, A., Barry, C., O'Keefe, J. & Burgess, N. Grid cells and theta as oscillatory interference: electrophysiological data from freely moving rats. *Hippocampus* **18**, 1175–1185 (2008).
11. Doeller, C. F., King, J. A. & Burgess, N. Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proc. Natl Acad. Sci. USA* **105**, 5915–5920 (2008).
12. Fernandez, G., Brewer, J. B., Zhao, Z., Glover, G. H. & Gabrieli, J. D. Level of sustained entorhinal activity at study correlates with subsequent cued-recall performance: a functional magnetic resonance imaging study with high acquisition rate. *Hippocampus* **9**, 35–44 (1999).
13. Olsen, R. K. *et al.* Performance-related sustained and anticipatory activity in human medial temporal lobe during delayed match-to-sample. *J. Neurosci.* **29**, 11880–11890 (2009).
14. Epstein, R., Graham, K. S. & Downing, P. E. Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron* **37**, 865–876 (2003).
15. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).
16. Park, S., Intraub, H., Yi, D. J., Widders, D. & Chun, M. M. Beyond the edges of a view: boundary extension in human scene-selective visual cortex. *Neuron* **54**, 335–342 (2007).
17. Bakker, A., Kirwan, C. B., Miller, M. & Stark, C. E. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).
18. Ekstrom, A. D. *et al.* Cellular networks underlying human spatial navigation. *Nature* **425**, 184–188 (2003).

19. Taube, J. S. Head direction cells and the neuropsychological basis for a sense of direction. *Prog. Neurobiol.* **55**, 225–256 (1998).

20. Chen, L. L., Lin, L. H., Green, E. J., Barnes, C. A. & McNaughton, B. L. Head-direction cells in the rat posterior cortex. I. Anatomical distribution and behavioral modulation. *Exp. Brain Res.* **101**, 8–23 (1994).

21. Byrne, P., Becker, S. & Burgess, N. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychol. Rev.* **114**, 340–375 (2007).

22. Wolbers, T., Wiener, J. M., Mallot, H. A. & Buchel, C. Differential recruitment of the hippocampus, medial prefrontal cortex, and the human motion complex during path integration in humans. *J. Neurosci.* **27**, 9408–9416 (2007).

23. Maguire, E. A. Neuroimaging studies of autobiographical event memory. *Phil. Trans. R. Soc. Lond. B* **356**, 1441–1451 (2001).

24. Schacter, D. L., Addis, D. R. & Buckner, R. L. Remembering the past to imagine the future: the prospective brain. *Nature Rev. Neurosci.* **8**, 657–661 (2007).

25. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).

26. Howard, M. W., Fotedar, M. S., Datey, A. V. & Hasselmo, M. E. The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychol. Rev.* **112**, 75–116 (2005).

27. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsaki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).

28. Gaffan, D. Scene-specific memory for objects: a model of episodic memory impairment in monkeys with fornix transection. *J. Cogn. Neurosci.* **6**, 305–320 (1994).

29. Knierim, J. J. Neural representations of location outside the hippocampus. *Learn. Mem.* **13**, 405–415 (2006).

30. Eichenbaum, H., Yonelinas, A. P. & Ranganath, C. The medial temporal lobe and recognition memory. *Annu. Rev. Neurosci.* **30**, 123–152 (2007).

METHODS

Single-unit recording experiment in rats: animals and surgery. Seventeen male Lister hooded rats (250–400 g at implantation) each received a single microdrive carrying four tetrodes[31] of twisted 17–25 μm HM-L coated platinum-iridium wire (90–10%) (California Fine Wire). In most cases 17 μm wire was platinum plated to reduce impedance to 200–300 kΩ at 1 kHz. The surgical procedure and housing conditions were the same as those described previously[5]. All work was conducted within the terms of appropriate Home Office Project and Personal licences.

Recording and behavioural training. Training and screening was performed post-surgically after a 1 week recovery period. An Axona recording system was used to acquire the single unit and positional data (see ref. 5 for details). The position of the rat was captured using an overhead video camera to record the position of the one or two LEDs on the animal's head-stage. In all cases head direction was inferred from the animal's trajectory. Animals were trained to forage for sweetened rice in a 1 m by 1 m square environment. Environment floors were made of either Perspex or polyvinyl and were wiped down with water between trials. Training consisted of at least three trials, each in excess of 15 min, distributed over 3 days. After this period animals would run continuously as they searched for sweetened rice. Activity was recorded while the animals foraged in the now familiar environment, with the exception of three rats trained in a 110-cm diameter cylinder and tested in 110-cm or 100-cm cylinders. In all cases the laboratory was visible above the arena wall. All recording trials were 20 min long. On days in which multiple recording trials were made only a single trial, usually the first, was used. The winding protocol, spatial firing correlates, and spike amplitudes were used to exclude multiple recordings of the same cell such that only the first recording was used. Potential grid cells were selected visually by the experimenter on the basis that they exhibited the regular triangular firing pattern diagnostic of grid cells. In this way 140 cells were subject to further analysis.

Data analysis: spike sorting and binning. Spike sorting was performed offline using a data analysis suite (Tint, Axona) and using Matlab (Mathworks). Action potentials were assigned to potential cells based on amplitude, waveform and temporal autocorrelation criteria applied elsewhere to entorhinal grid cells[1,5]. The animal's position and spike locations were binned into a 64×64 bin array covering the camera's field of view; each bin being the equivalent of 8×8 pixels, roughly 2×2 cm. Unsmoothed rate maps were calculated by dividing the number of spikes assigned to a bin by the cumulative occupancy of the bin. For smoothed rate maps the firing rate for bin i was the number of spikes in a 5×5 kernel centred on i divided by the cumulative occupancy of the same bins. Similarly, polar rate maps were constructed using 6 degree radial bins and were smoothed with a Gaussian kernel ($\sigma = 10°$).

Gridness and grid orientation. Spatial autocorrelograms were estimated from unsmoothed rate maps[1,6], smoothed (using a two-dimensional Gaussian kernel, $\sigma = 2.5$ bins), and used to assess the periodicity, regularity and orientation of cells with multiple firing fields. The spatial autocorrelogram was defined as:

$$r(\tau_x, \tau_y) = \frac{n \sum \lambda_1(x,y)\lambda_2(x-\tau_x, y-\tau_y) - \sum \lambda_1(x,y) \sum \lambda_2(x-\tau_x, y-\tau_y)}{\sqrt{n\sum \lambda_1(x,y)^2 - \left(\sum \lambda_1(x,y)\right)^2}\sqrt{n\sum \lambda_2(x-\tau_x, y-\tau_y)^2 - \left(\sum \lambda_2(x-\tau_x, y-\tau_y)\right)^2}}$$

where $r(\tau_x, \tau_y)$ is the autocorrelation between bins with spatial offset of τ_x and τ_y. $\lambda_1(x,y)$ and $\lambda_2(x,y)$ are the mean firing rate in bin (x,y) of the firing rate map and n is the number of pairs of visited bins. Six 'central peaks' were identified as the local maxima closest to, but excluding, the central peak in the spatial autocorrelogram. The extent of each peak was defined as the contiguous set of bins around the peak with a value greater than half the peak value. The orientation of each cell was calculated as the angle between a nominal east reference line and an axis defined by the centre of the spatial autocorrelogram and the peak closest to the reference line in an anticlockwise direction from east. Gridness, a measure of spatial periodicity, was calculated by defining an area of the spatial autocorrelation centred on but excluding the central peak bounded by a circle passing around the outermost edge of the six central peaks. This area was rotated in 30° increments up to 150°, and for each rotation the correlation coefficient with the un-rotated area was found. Gridness was defined as the lowest correlation for rotations of 60° and 120° minus the highest correlation for 30°, 90° or 150°. Cells with gridness of 0 or greater were classified as grid cells (see Supplementary Fig. 6). A total of 113 cells from 16 of the 17 rats were judged to be grid cells.

Directionality and directional measures. The directionality of firing exhibited by each grid cell was defined as the KL divergence between the polar rate map and a uniform circular distribution with equal mean:

$$D_{KL} = \sum_i \frac{\tau_1(i)\log(\tau_1(i))}{\tau_2(i)}$$

where $\tau_1(i)$ is the value in the ith bin of a polar rate map normalized to have area 1 and $\tau_2(i)$ is the value in the ith bin of a uniform probability distribution with the

same number of bins as τ_1. Cells with $D_{KL} \geq 0.15$ were considered directional, consistent with the experimenter's intuition, producing a total of 18 directional cells from 8 rats (see Supplementary Figs 3–6). For each directional grid cell the circular mean of the polar rate map was calculated according to:

$$\mu_{circ} = \arctan\left(\frac{\sum_i \tau_1(i)\sin\theta_i}{\sum_i \tau_1(i)\cos\theta_i}\right)$$

Where θ_i is the angle of the centre of radial bin i and $\tau_1(i)$ is defined above.

Histology. After the experiment rats received an overdose of Euthatal (sodium pentobarbital) and were transcardially perfused with phosphate-buffered saline and then with 4% paraformaldehyde (PFA) solution. The brains were removed and stored in 4% PFA for at least 1 week before sectioning (in some cases 20% sucrose cryo-protection was used). 40 μm frozen sagittal sections were cut, mounted on gelatine-coated glass slides and stained with cresyl violet. High resolution images were acquired using either an Olympus microscope with Xli digital camera (XL Imaging Ltd.) or Nikon Labophot-2 microscope with Leica DFC420c camera. The depth and layer at which cells were recorded was extrapolated from the record of tetrode movements after taking account of brain shrinkage. All recordings were confirmed to the dorso-lateral extent of the medial entorhinal cortex (see Supplementary Fig. 2 for examples of recording locations).

Functional neuroimaging experiment in humans: participants. Forty-two male participants took part in this experiment (aged 18–31, mean age 23.1 years). Participants gave written consent and were paid for participating, as approved by the local Research Ethics Committee. All were right-handed with normal or corrected-to-normal vision and reported to be in good health with no history of neurological disease. All had previous experience of playing first person perspective video games. The current report includes data from a previous study[11] where we report the results of unrelated analyses.

Virtual reality environment and task. We used UnrealEngine2 Runtime software (Epic Games) to present a first-person perspective view of a grassy plane surrounded by a circular cliff with a background of mountains, clouds and the sun (created using Terragen, Planetside Software) projected at infinity, to provide orientation but not location within the arena. One rotationally symmetric intramaze landmark was always present in the arena. Participants moved the viewpoint by using their right hand to operate keys to move forward and turn left or right. The viewpoint is ~2 virtual metres above ground, the arena is ~180 virtual metres in diameter and the virtual heading and location were recorded every 100 ms. Participants received training in an unrelated virtual environment before performing the experiment and then familiarized themselves with the main arena in the scanner. Throughout the experiment, participants collected and replaced everyday objects within the arena. They collected each object once during an initial trial, by running over it. In each subsequent trial they saw an image of one of the objects and had to move to where they thought the object had been in the arena. After their response by a button press, feedback was provided (that is, the object appeared in its correct position and participants collected it again). Memory was measured in terms of the proximity of the replace response location to the correct object location (see Supplementary Fig. 7 and ref. 11).

fMRI acquisition. Blood-oxygenation-level-dependent (BOLD) T2*-weighted functional images were acquired on a 3T Siemens Allegra scanner using a gradient-echo echo-planar imaging (EPI) pulse sequence with the following parameters: Repetition time (TR) = 2,600 ms, echo time (TE) = 30 ms, flip angle = 90°, slice thickness = 2 mm, interslice gap = 1 mm, in-plane resolution = 3×3 mm, field of view (FoV) = 192 mm², 40 slices per volume. The first five volumes were discarded to allow for T1 equilibration. The sequence was optimized to minimize signal dropouts in the medial temporal lobes[32]. In addition, a field map using a double echo FLASH sequence was recorded for distortion correction of the acquired EPI images[32] (see below). For 32 of the participants we also acquired a standard T1-weighted structural MRI scan (3D MDEFT, resolution = $1 \times 1 \times 1$ mm).

Image pre-processing. The main imaging analyses were performed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/) within Matlab. Additional analyses were conducted outside SPM using Matlab. Functional images were spatially realigned to the first image in the times series, were corrected for distortions based on the field map[33] and the interaction of motion and distortion using the Unwarp routines in SPM[33,34], and sinc-interpolated in time to correct for differences in slice acquisition time. The structural images underwent segmentation (into grey matter, white matter and cerebro-spinal fluid), bias correction and spatial normalization to the MNI template using 'unified segmentation'[35]. For participants with a structural scan, functional images were normalized based on the spatial parameters derived from the normalization of their structural images. For participants without a structural scan, images were normalized to an EPI template specific to our sequence and scanner that was aligned to the MNI T1 template.

Finally, the normalized functional images were spatially smoothed with an isotropic 8-mm full-width-half-maximum Gaussian kernel.

Analysis of fMRI time series. After pre-processing, fMRI time series were modelled by a set of general linear models (GLMs). All models included regressors for the entire time of virtual reality navigation, separated by the participant's speed of movement (three equal tertiles for high, medium and slow speed), various parametric modulations of these regressors (see details below), and regressors of no interest modelling phases without navigation in the environment (see ref. 11). All parametric modulations are normalized to have zero mean and thus be orthogonal to the un-modulated regressor[36]. All regressors (parametrically modulated or not) were convolved with the canonical haemodynamic response function (HRF) in SPM before entering the GLM. Data were high-pass filtered at 1/128 Hz. Coefficients for each regressor were estimated for each participant using maximum likelihood estimates to account for serial correlations in the data. Linear contrasts of coefficients for each participant were entered into second level random-effects analyses. The statistical reliability (across participants) of differences of parameter estimates from each other, or from zero, are shown and referred to as fMRI 'activation'. On the basis of our strong a priori hypotheses we have chosen an uncorrected statistical threshold of $P < 0.001$ for whole-brain analyses. Analyses of single comparisons (for example, concerning the mean activation across an ROI) are tested at $P = 0.05$. Coordinates of brain regions are reported in MNI space.

Entorhinal cortex ROI. We defined an ROI including both entorhinal cortices following anatomical boundaries as described previously[37]. In rats, grid cells are found in the medial but not the lateral entorhinal cortex. However, the homologues of these subdivisions in humans are not known. Eight subdivisions of the human entorhinal cortex have been identified[38]. The medio-lateral orientation of most of these subdivisions rules out a simple correspondence between medial entorhinal areas in humans and medial entorhinal cortex in rodents. In rats, medial entorhinal cortex receives direct input from presubiculum. In monkeys, the more posterior subdivisions of entorhinal cortex receive such input (M. Witter, personal communication), which might also be the case in humans. We note that many, but not all, of our entorhinal cortex activations are more posterior, and that the coherence of potential grid orientation is slightly stronger in posterior than anterior entorhinal cortex. However, future high-resolution imaging in humans and anatomical studies in non-human primates will be required to resolve this issue fully.

Quadrature filter and mean grid orientation analysis. To estimate the potential grid orientations in entorhinal cortex, we used a quadrature filter, a well-known signal processing technique, on one-half of the data. We separated each participant's virtual navigation fMRI data arbitrarily into 12 scanning runs of equal length. For one-half of the data (odd runs), we calculated the cosine and sine of the participant's virtual running direction $\theta(t)$ with a periodicity of 60°, that is, $\cos(6\theta(t))$ and $\sin(6\theta(t))$, arbitrarily aligned to 0° of the virtual reality environment as defined by the background cues (see Supplementary Fig. 8) separately for fast, medium and slow runs. The factor 6 means that these regressors are sensitive to activation showing a six-fold rotational symmetry in running direction (that is, activation with six evenly spaced peaks as a function of running direction will produce parameter estimates for the two regressors (β_1 and β_2) with large amplitude $\left(\sqrt{\beta_1{}^2 + \beta_2{}^2}\right)$). These cosine and sine regressors (separately for the three speed levels) were included in the GLM used to analyse the fMRI data. The resulting parameter estimates β_1 and β_2 were used to calculate the potential grid orientation in each voxel Φ (varying between 0 and 59°) in the entorhinal ROI as $\Phi = [\arctan(\beta_2/\beta_1)]/6$, where arctan is mapped into the range [0–360°] according to the signs of β_2 and β_1, separately for the three speed levels. The 'mean grid orientation' for the region (φ) was defined as the population vector (vector average of grid orientations Φ in each voxel weighted by the amplitude of sinusoidal modulation in that voxel, or $[\arctan(<\beta_2>/<\beta_1>)]/6$ where $<>$ denotes the mean across voxels), separately for the three speed levels.

In a next step, we looked for sinusoidal modulation of activation with six-fold rotational symmetry in the other half of the data (even runs). We conducted the following analyses: (1) calculating the six-fold cosine of the direction of virtual navigation aligned to the mean grid orientation (that is, $\cos\{6[\theta(t)-\varphi]\}$) resulting in one aligned cosine parametric regressor per speed level (see Fig. 3 and Supplementary Fig. 8). (2) For visualization of the effect, we separated runs aligned (within 15° of the nearest main axis of the grid) from those misaligned to the main grid axes (more than 15° from a grid axis), resulting in two (aligned versus misaligned) non-parametric regressors per speed level (see Fig. 3 and Supplementary Fig. 10).

Control models. To test for the specificity of the six-fold symmetric sinusoidal modulation, we used the same approach of (1) estimating the mean grid orientation of entorhinal cortex and (2) looking for sinusoidal modulation in control models, however with directional periodicities of 90° and 45° (that is, four-fold and eight-fold rotational symmetry; Fig. 3) and of 72° and 51.4° (that is, five-fold and seven-fold rotational symmetry) respectively (see Supplementary Fig. 11).

Adaptation analyses. In the basic analysis of adaptation to running direction we separated the whole directional range of the participant's virtual path into 36 equal 10° bins and then looked for fMRI correlates of how recently participants were running in the current direction bin (regressor: log[time since last running in the current direction]), separately for the three speed levels. To make sure that the adaptation effect is not confounded by angular velocity, the model also included a regressor reflecting angular velocity. In a next step we used the same logic as the above analysis but looked for adaptation to running directions shifted by 60° from the current direction (regressor: log[time since last running at 60° to the current direction]). For the visualization of the adaptation effect, we grouped trials into four bins reflecting quartiles of the log[time] regressor, estimated a new GLM, and calculated the per cent signal change separately for each bin. (See bar plots in Fig. 4f.) We also analysed control models, with adaptation to directions shifted by 90° and 45° from the current direction (regressor: log[time since last running at 90° and 45° to the current direction]).

Showing fMRI effect sizes. The mean fMRI signal is expressed in terms of per cent signal change compared to the 'baseline' level of fMRI activation (the constant term in the GLM, changes from which are modelled by our regressors) for the peak voxel in a region found in the respective analyses and averaged across participants. The adaptation effect is also expressed as per cent signal change: the amplitude of the adaptation regressor fitted by the GLM (that is, the difference between the amplitude of response when the previous sample was most recent versus longest ago) compared to the baseline level. In our analyses 'sample' corresponds to running in the current direction (Fig. 4a, b) or ±60° from it (Fig. 4c, d).

31. Recce, M. & O'Keefe, J. The tetrode: a new technique for multiunit extracellular recording. *Soc. Neurosci. Abstr.* **15**, 1250 (1989).

32. Weiskopf, N., Hutton, C., Josephs, O. & Deichmann, R. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage* **33**, 493–504 (2006).

33. Hutton, C. *et al.* Image distortion correction in fMRI: A quantitative evaluation. *Neuroimage* **16**, 217–240 (2002).

34. Andersson, J. L., Hutton, C., Ashburner, J., Turner, R. & Friston, K. Modeling geometric deformations in EPI time series. *Neuroimage* **13**, 903–919 (2001).

35. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26**, 839–851 (2005).

36. Büchel, C., Holmes, A. P., Rees, G. & Friston, K. J. Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage* **8**, 140–148 (1998).

37. Fischl, B. *et al.* Predicting the location of entorhinal cortex from MRI. *Neuroimage* **47**, 8–17 (2009).

38. Insausti, R., Tunon, T., Sobreviela, T., Insausti, A. M. & Gonzalo, L. M. The human entorhinal cortex: a cytoarchitectonic analysis. *J. Comp. Neurol.* **355**, 171–198 (1995).

LETTERS

Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons

Valeria Specchia[1], Lucia Piacentini[2], Patrizia Tritto[3], Laura Fanti[2], Rosalba D'Alessandro[3], Gioacchino Palumbo[3], Sergio Pimpinelli[2] & Maria P. Bozzetti[1]

The canalization concept[1] describes the resistance of a developmental process to phenotypic variation, regardless of genetic and environmental perturbations, owing to the existence of buffering mechanisms. Severe perturbations, which overcome such buffering mechanisms, produce altered phenotypes that can be heritable and can themselves be canalized by a genetic assimilation process. An important implication of this concept is that the buffering mechanism could be genetically controlled. Recent studies on Hsp90, a protein involved in several cellular processes and development pathways[2–5], indicate that it is a possible molecular mechanism for canalization and genetic assimilation. In both flies and plants, mutations in the Hsp90-encoding gene induce a wide range of phenotypic abnormalities, which have been interpreted as an increased sensitivity of different developmental pathways to hidden genetic variability[6,7]. Thus, Hsp90 chaperone machinery may be an evolutionarily conserved buffering mechanism of phenotypic variance, which provides the genetic material for natural selection. Here we offer an additional, perhaps alternative, explanation for proposals of a concrete mechanism underlying canalization. We show that, in *Drosophila*, functional alterations of Hsp90 affect the Piwi-interacting RNA (piRNA; a class of germ-line-specific small RNAs) silencing mechanism leading to transposon activation and the induction of morphological mutants. This indicates that Hsp90 mutations can generate new variation by transposon-mediated 'canonical' mutagenesis.

In *Drosophila*, primary spermatocytes of Hsp90 mutant males exhibit crystalline aggregates[8,9] (see also Fig. 1b), usually absent in wild-type testes (Fig. 1a), the formation of which is due to the transcriptional activation, in male germ line, of the repeated *Stellate* (*Ste*) elements[10–13]. Such elements encode a protein, similar to the β-subunit of casein kinase 2 (CK2) (refs 12, 14), that is the main component of the crystalline aggregates[14]. To test the specificity of Hsp90 mutations in inducing crystalline aggregates, we analysed spermatocytes of wild-type males treated with the Hsp90 inhibitor geldanamycin and found
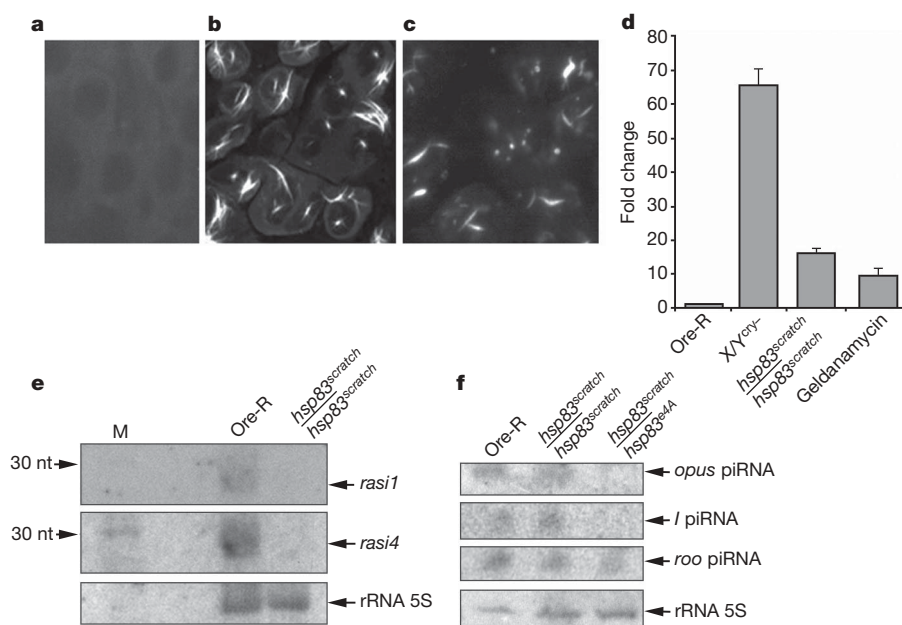

Figure 1 | *hsp83* mutations activate *Stellate* sequences and affect *crystal* and transposon piRNA production. **a**, Testes of Ore-R males immunostained with anti-Stellate. **b**, Testes immunostained with anti-Stellate in a *hsp83scratch* mutant. **c**, Testes immunostained with anti-Stellate in geldanamycin-treated Ore-R males. The magnification for **a**–**c** is 40×. **d**, Quantitative RT–PCR of *Stellate* euchromatic transcript in testes of different genotypes. Error bars denote the coefficient of variation. **e**, Total RNA from testes hybridized with probes complementary to *crystal* piRNAs, *rasi1* and *rasi4*. 30 nt indicates the 30-nucleotide molecular marker RNA. **f**, Total RNA from ovaries hybridized with probes complementary to piRNA from *opus*, *roo* and *I* transposons. Note that for all transposons, the piRNAs are reduced in homozygotes whereas they disappear in trans-heterozygotes except for *roo*, where they appear only strongly reduced.

[1]Dipartimento di Scienze e Tecnologie Biologiche ed Ambientali (DiSTeBA), University of Salento, 73100 Lecce, Italy. [2]Istituto Pasteur, Fondazione Cenci-Bolognetti and Dipartimento di Genetica e Biologia Molecolare, University of Rome 'La Sapienza', 00185 Roma, Italy. [3]Dipartimento di Genetica e Microbiologia (DIGEMI), University of Bari, 70126 Bari, Italy.
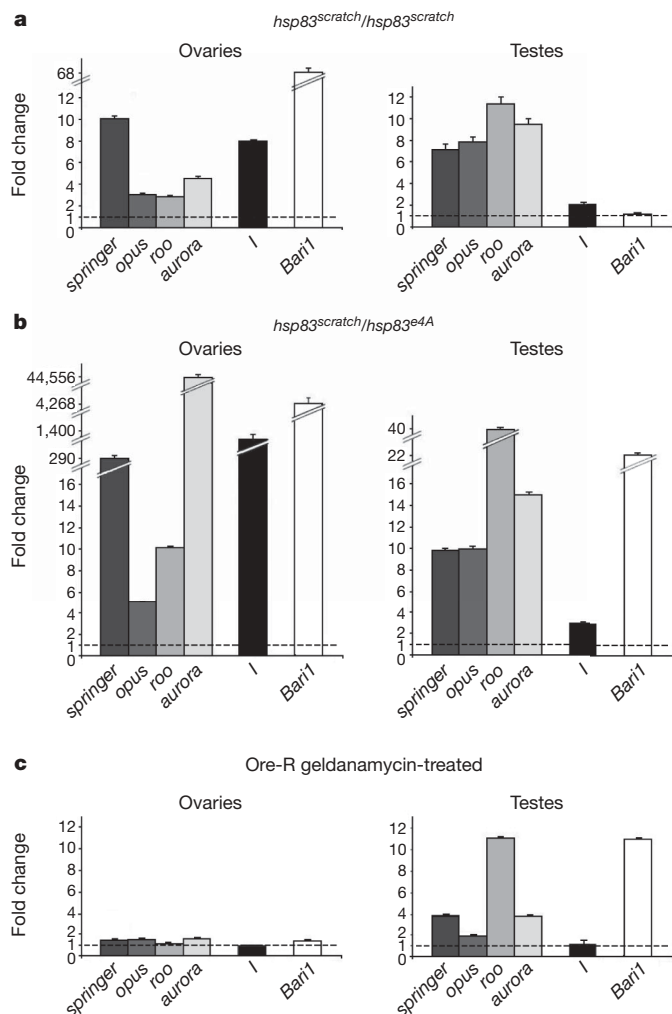
Figure 2 | hsp83 mutations and geldanamycin activate transposon transcription. The relative amount of transposon transcripts with respect to *rp49* transcript was determined by quantitative RT–PCR in total RNA extracted from ovaries (left) and testes (right) of the indicated genotypes. **a**, *hsp83^scratch* homozygous ovaries (left side) and testes (right side) compared to *hsp83^scratch* heterozygous ovaries and testes. **b**, *hsp83^scratch*/ *hsp83^e4A* trans-heterozygous ovaries (left side) and testes (right side) compared to *hsp83^scratch* heterozygous ovaries and testes. **c**, Ovaries (left side) and testes (right side) of Ore-R flies treated with geldanamycin. Values are expressed on an arbitrary scale described in Methods; the dashed line represents the value of 1, arbitrarily assigned to the control genotype. Error bars denote the coefficient of variation.

crystalline aggregates (Fig. 1c) and a significant amount of *Stellate* transcript (Fig. 1d).

The silencing of the *Stellate* sequences is mediated by RNA interference (RNAi) mechanisms[9,15–17] and mutations in genes involved in RNAi, such as *aubergine*, *armitage* and *spindle E* (also called *homeless*), activate *Stellate* in testes of mutant males[9,18,19]. More recently, it has been shown that *Stellate* is repressed by a piRNA-mediated mechanism that is specific for repetitive sequences and transposon silencing[20,21]. Consistently, we found that Hsp90 mutations affect the biogenesis of piRNAs specific for *Stellate* and transposons (Fig. 1e, f). These results prompted us to test, in ovaries and testes, for possible effects of Hsp90 mutations on the expression of long terminal repeat (LTR) *springer*, *opus*, *roo* and *aurora*, non-LTR *I* elements that transpose by an RNA intermediate, and invert repeat (IR) *Bari1*, an element that transposes by a DNA intermediate[20,22]. We analysed homozygous *hsp83^scratch* and trans-heterozygous *hsp83^scratch*/*hsp83^e4A* mutants. Hsp83 is the denomination of Hsp90 in *Drosophila*; *hsp83^scratch* is a viable hypomorphic male sterile mutation[8] and *hsp83^e4A* is a lethal amorphic mutation[3,23]. We also
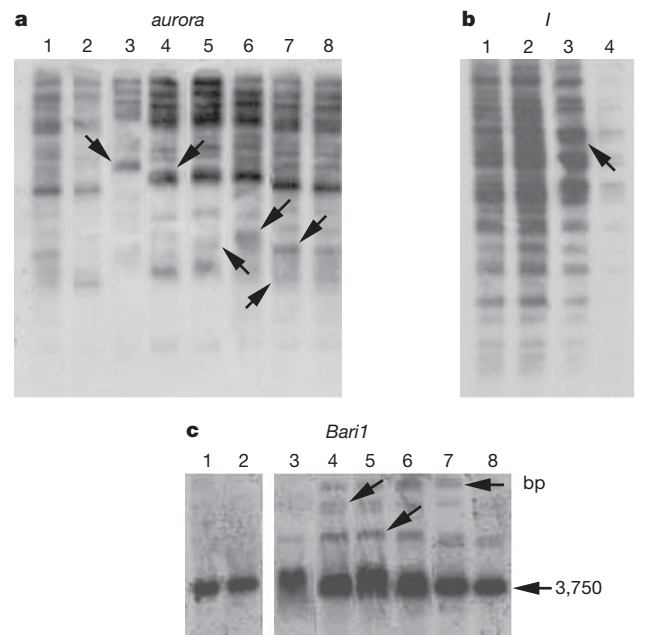

Figure 3 | Southern blot analysis of DNA from single flies with different genotypes. **a**, Single fly Southern blot analysis with *aurora* probe. Lane 1, DNA from *hsp83^scratch+*/*hsp83^scratch* male parent; lane 2, DNA from *hsp83^scratch+*/*hsp83^scratch* female parent; lanes 3–5, DNA from *hsp83^scratch*/ *hsp83^scratch* F₁ female progeny; lanes 6–8, DNA from *hsp83^scratch*/*hsp83^scratch* F₁ male progeny. **b**, Single fly Southern blot analysis with *I* probe. Lane 1, DNA from *hsp83^scratch+*/*hsp83^scratch* male parent; lane 2, DNA from *hsp83^scratch+*/*hsp83^scratch* female parent; lane 3, DNA from *hsp83^scratch*/ *hsp83^scratch* F₁ female progeny; lane 4, DNA from *hsp83^scratch*/*hsp83^scratch* F₁ male progeny. **c**, Single fly Southern blot analysis with *Bari1* probe. Lane 1, DNA from *hsp83^scratch+*/*hsp83^scratch* male parent; lane 2, DNA from *hsp83^scratch+*/*hsp83^scratch* female parent; lanes 3–5, DNA from *hsp83^scratch*/ *hsp83^scratch* F₁ female progeny; lanes 6–8, DNA from *hsp83^scratch*/*hsp83^scratch* F₁ male progeny. Arrows (**a**–**c**) indicate new bands in the progeny DNA.

tested the expression of the same transposons in ovaries and testes of wild-type Oregon-R (Ore-R) flies treated with geldanamycin. As shown in Fig. 2, in mutant ovaries and testes the amount of transcripts increases significantly, although differentially, for all the transposons. The increase is more abundant in trans-heterozygous (Fig. 2b) than in homozygous mutants (Fig. 2a) and in ovaries than in testes. Treatment with geldanamycin (3 μg ml⁻¹) induces a significant increase of all the transposon transcripts, except for the *I* element, but only in testes (Fig. 2c).

We used Southern blot analysis to look at the effect of Hsp90 mutations on transposon mobility. To compare genomes homogeneously, we extracted DNA from single heterozygous flies (the parents) and from single homozygous flies of F₁ progeny (males and females); all the DNA genomic samples were digested with HindIII. Figure 3 shows the hybridization patterns with *aurora* (Fig. 3a), *I* (Fig. 3b) and *Bari1* (Fig. 3c), where significant differences between parents and F₁ progeny are evident for each element. We found new bands in the DNA of the F₁ progeny compared to the parental DNA, thus suggesting a mutation-induced mobilization of these elements. We obtained similar results with *springer*, *opus* and *roo* (data not shown). To ensure that the effects on transposons are due exclusively to the homozygosis of the *hsp83^scratch* mutation and not to the genetic background of the mutagenized III chromosome, we did the same experiments with *aurora*, *I* and *Bari1* in trans-heterozygous *hsp83^scratch*/*hsp83^e4A* mutants with similar results (data not shown). In addition, we did not find any significant mobilization or increased transcription of the same elements in a homozygous *hsp83^rev22* revertant strain (data not shown).

Because active transposons are mutagenic, these data suggest that the phenotypic variation observed in Hsp90 mutants could be due to
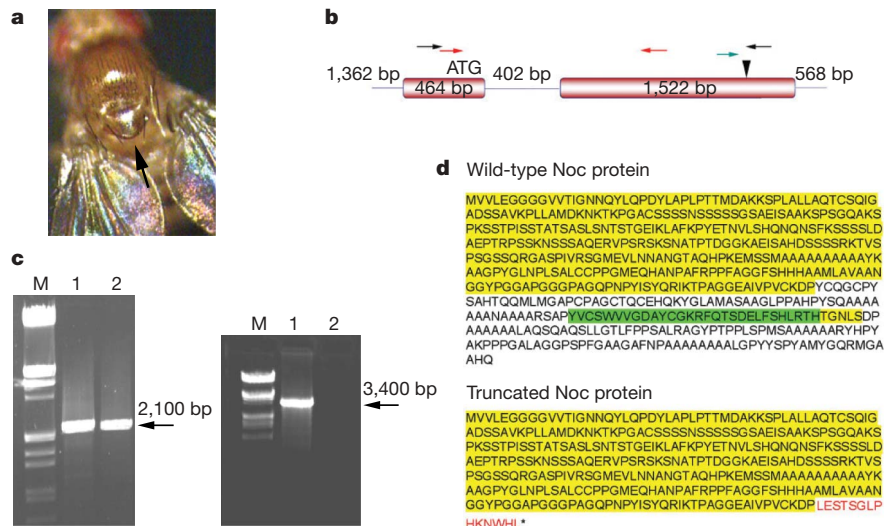
Figure 4 | Analysis of the morphological mutant identified among *hsp83*[scratch] **homozygous flies. a**, Fly showing a *Scutoid*-like phenotype. **b**, Schematic structure of the *noc* gene. Black arrows correspond to primers that amplify a 3,400-bp-long fragment in wild-type DNA (for specific positions see Methods); red arrows correspond to primers that amplify a 2,100-bp-long fragment in wild-type DNA (for specific positions see Methods); the green arrow corresponds to the primer used in the amplification of the DNA extracted from the *Sco*-like mutant coupled with a lower primer localized in the *I* element (for sequence and position see Methods). The arrowhead indicates the insertion site of the *I* element (position 1394 of the *noc* cDNA). **c**, 1% agarose gel showing the presence of the 2,100-bp amplicon in both wild type (left lane 1) and *Sco*-like mutant (left lane 2) whereas the 3,400-bp amplicon is only present in wild type (right lane 1). M, marker. **d**, The wild-type sequence of the Noc protein with the zinc-finger motif highlighted in green is shown in the top panel; sequence of the truncated Noc protein from the *Sco*-like mutant as deduced from the gene sequence is shown in the bottom panel.

de novo mutations produced by activated transposons. To test this possibility, we screened 3,220 *hsp83*[scratch] homozygous flies and found 30 flies with morphological abnormalities, corresponding to a frequency of about 1%. This is similar to the frequency previously reported in some mutant stocks[6]. We also analysed 3,220 flies from an Ore-R stock without finding any morphological abnormalities. Among the abnormalities we did observe, there was a fly resembling the dominant mutation *Scutoid* (Fig. 4a). Because the *Scutoid* phenotype is caused by mutations in the *noc* gene (Fig. 4b), we molecularly analysed this gene in the phenotypic variant. We choose two pairs of primers that in the wild-type gene amplify two DNA fragments of 2,100 bp and 3,400 bp (positions are indicated in Fig. 4b). As shown in Fig. 4c, the 2,100-bp fragment is present in both the *Sco*-like mutant and in Ore-R, whereas the 3,400-bp fragment is detected only in the wild type. Inverse polymerase chain reaction (PCR) analysis showed that the *noc* gene is interrupted at nucleotide 1394 of the cDNA sequence by an *I* element-like sequence (Fig. 4b). Consistent with this observation is the amplification of a DNA fragment obtained by using an upper primer—the position of which in the *noc* gene is indicated in Fig. 4b—and a lower primer corresponding to the *I* element at the 1243 nucleotide of the M14954 sequence. After sequencing we found that this fragment encodes a Noc truncated protein, which cannot function as a transcriptional factor owing to the loss of the zinc-finger domain (Fig. 4d). We also analysed the *noc* sequence in the DNA extracted from about 1,000 phenotypically wild-type flies collected during our screening, and found only the normal *noc* sequence. This indicates that the *Scutoid* phenotype that we found was caused by a *de novo* mutation instead of being the expression of a pre-existing cryptic mutation. The impairment of piRNA biogenesis by Hsp83 indicated that other mutations affecting transposon activity might also induce phenotypic variation. To test this possibility, we made a screen for morphological variants in a stock carrying *spindle E*[e00786], a male sterile mutation at the *spindle E* gene. It has helicase activity, de-represses *Stellate* sequences and transposons[19], and is involved in piRNA biogenesis[24–26]. We screened 1,500 flies and found 19 morphological variants (see Supplementary Fig. 1a–c for examples), a similar rate to that found in some Hsp90 mutants strains (1–2%). Supplementary Fig. 1d shows that the transposons are activated in the germ line of male and female *spindle E*

mutants. This further indicates that the expression of morphological variability could be related to the disruption of the piRNA silencing mechanism.

These data clearly show a novel function of Hsp90 with important implications for the current hypothesis about this protein as a capacitor of morphological evolution. Hsp90 is involved in stress responses and the expression of wide ranging morphological changes in *Drosophila* and other organisms. A reduced amount of Hsp90 can induce abnormal developmental phenotypic variations; these morphogenetic changes can become fixed and stably transmitted even if wild-type Hsp90 function is restored in subsequent generations[6]. The current interpretation is that Hsp90 buffers pre-existing genetic variation that is not expressed and accumulates in neutral conditions; its mutations will then induce the expression of this variation[6]. This stress-sensitive storage and release of genetic variation by Hsp90 would favour adaptive evolution. The hypothesis of pre-existing genetic variation has been based on the observation that, when flies carrying Hsp83 alleles were outcrossed to different laboratory stocks, the observed defects were typical for each outcross, thus suggesting that the defects depended on specific genetic backgrounds[6]. We think that those data could be also explained in the light of our results. There is evidence that different genetic backgrounds may induce different transposon insertions[27].

Our demonstration that Hsp90 is involved in the control of transcription and mobilization of transposable elements in germ cells by affecting piRNA biogenesis strongly suggests that the reduction of Hsp90 causes a stress-response-like activation and transposition of mobile elements affecting piRNA silencing. This in turn would induce *de novo* gene mutations that affect development pathways and that can be expressed and fixed across subsequent generations. This explanation agrees with the suggestion that transposable element activity could be a response to stress[28]. Our results do not exclude, however, that Hsp90 could be both a buffering factor as well as a suppressor of transposable element (TE)-induced mutations.

We think that our data indicate an additional, if not alternative, mechanism to the canalization and assimilation hypothesis based on a link between stress and transposon activity through piRNA-mediated silencing. This mechanism also potentially provides another molecular interpretation with respect to the vague capacitor concept. We do not yet know if Hsp90 also has a direct role in piRNA biogenesis, or is

only involved in triggering the stress response leading to transposon activation.

METHODS SUMMARY

Drosophila wild-type stocks, balancers and fly media are described at http://flybase.org.

hsp83^scratch is a viable, male sterile and female semifertile allele obtained from a P (*lacZ ry^+*) element mutagenesis[23]. It is maintained in the stock *hsp83^scratch*/TM3 or in *hsp83^scratch*/TM6B. *hsp83^e4A* is a lethal allele[3], maintained in the stock *hsp83^e4A*/TM6B.

spindle E^c00786 is a male sterile allele, maintained in the stock *spindle E^c00786*/TM6B (Bloomington number 10274).

Drosophila testes were manually dissected in Ringer's solution. A total of 20 μg of total RNA, isolated from testes using Trizol (Invitrogen), was resolved by 10% denaturing polyacrylamide/urea gel electrophoresis. After electrophoresis, the polyacrylamide gel was transferred to Hybond N+ (GE Healthcare) and hybridized with probes corresponding to the oligonucleotides antisense to *crystal*-specific and transposon piRNAs. They were 5′-^{32}P-radiolabelled with polynucleotide kinase (Roche) and 30 μCi [γ-^{32}P]ATP.

In the first-strand cDNA synthesis, 5 μg of total RNA were used as a template for oligonucleotide dT primed reverse transcription using SuperScript III RNaseH-reverse transcriptase (Invitrogen), according to manufacturer's instructions. Real-time RT–PCR was performed in the SmartCycler Real-time PCR (Cepheid). Relative abundance of the different transcripts in germinal tissues was determined using Fluo Cycle for SYBR green (Celbio) according to manufacturer's protocol. For quantification of the transcripts we used the $2^{\Delta\Delta Ct}$ method[29]. The primers we used are reported in the Methods.

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

1. Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565 (1942).
2. Ding, D., Parkhurst, S. M., Halsell, S. R. & Lipshitz, H. D. Dynamic *hsp83* localization during *Drosophila* oogenesis and embryogenesis. *Mol. Cell. Biol.* **13**, 3773–3781 (1993).
3. Cutforth, T. & Rubin, G. M. Mutations in *Hsp83* and *cdc37* impair signaling by the sevenless receptor tyrosine kinase in *Drosophila*. *Cell* **77**, 1027–1036 (1994).
4. Hartl, F. U. Molecular chaperones in cellular protein folding. *Nature* **381**, 571–580 (1996).
5. van der Straten, A., Rommel, C., Dickson, B. & Hafen, E. The heat shock protein 83 (Hsp83) is required for Raf-mediated signalling in *Drosophila*. *EMBO J.* **16**, 1961–1969 (1997).
6. Rutherford, S. L. & Lindquist, S. HSP90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
7. Queitsch, C., Sangster, T. A. & Lindquist, S. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**, 618–624 (2002).
8. Yue, L. *et al.* Genetic analysis of viable Hsp90 alleles reveals a critical role in *Drosophila* spermatogenesis. *Genetics* **151**, 1065–1079 (1999).
9. Tritto, P. *et al.* Structure, regulation and evolution of the *crystal–Stellate* system of *Drosophila*. *Genetica* **117**, 247–257 (2003).
10. Hardy, R. W. *et al.* Cytogenetic analysis of a segment of the Y chromosome of *Drosophila melanogaster*. *Genetics* **107**, 591–610 (1984).
11. Livak, K. J. Organization and mapping of a sequence on the *Drosophila melanogaster* X and Y chromosome that is transcribed during spermatogenesis. *Genetics* **107**, 611–634 (1984).
12. Livak, K. J. Detailed structure of the *Drosophila melanogaster Stellate* genes and their transcript. *Genetics* **124**, 303–316 (1990).
13. Palumbo, G. *et al.* Interaction systems between heterochromatin and euchromatin in *Drosophila melanogaster*. *Genetica* **94**, 267–274 (1994).
14. Bozzetti, M. P. *et al.* The *Ste* locus, a component of the parasitic *cry–Ste* system of *Drosophila melanogaster*, encodes a protein that forms crystals in primary spermatocytes and mimics properties of the β-subunit of casein kinase 2. *Proc. Natl Acad. Sci. USA* **92**, 6067–6071 (1995).
15. Gvozdev, V. A. *et al.* Paralogous *Ste* and *Su(Ste)* repeats: evolution and ability to silence a reporter gene. *Genetica* **109**, 131–140 (2000).
16. Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* **11**, 1017–1027 (2001).
17. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5**, 337–350 (2003).
18. Schmidt, A. *et al.* Genetic and molecular characterization of *sting*, a gene involved in crystal formation and meiotic drive in the male germ line of *Drosophila melanogaster*. *Genetics* **151**, 749–760 (1999).
19. Stapleton, W., Das, S. & McKee, B. A role of the *Drosophila homeless* gene in repression of Stellate in male meiosis. *Chromosoma* **110**, 228–240 (2001).
20. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
21. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature Rev. Genet.* **10**, 94–108 (2009).
22. Caizzi, R., Caggese, C. & Pimpinelli, S. Bari-1, a new transposon-like family in *Drosophila melanogaster* with a unique heterochromatic organization. *Genetics* **133**, 335–345 (1993).
23. Castrillon, D. H. *et al.* Toward molecular genetic analysis of spermatogenesis in *Drosophila melanogaster*: characterization of male sterile mutants generated by single P element mutagenesis. *Genetics* **135**, 489–505 (1993).
24. Nishida, K. M. *et al.* Gene silencing mechanisms mediated by Aubergine–piRNA complexes in *Drosophila* male gonad. *RNA* **13**, 1911–1922 (2007).
25. Li, C. *et al.* Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**, 509–521 (2009).
26. Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137**, 522–535 (2009).
27. Soriano, S., Fortunati, D. & Junakovic, N. Evidence for the host contribution in the definition of preferential insertion sites of the elements of Bari 1 transposon family in *Drosophila melanogaster*. *J. Mol. Evol.* **55**, 606–615 (2002).
28. McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
29. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using realtime quantitative PCR and the $2^{\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).

Author Contributions V.S., L.P., P.T., L.F., R.D. and G.P. performed research; S.P. and M.P.B. designed research; S.P. and M.P.B. analysed data and wrote the paper; all authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P.B. (maria.bozzetti@unisalento.it).

METHODS

Microscopy and histology. Testes were dissected in Ringer's solution (182 mM KCl, 46 mM NaCl, 3 mM CaCl$_2$, 10 mM Tris-HCl, pH 7.5), and immediately visualized under phase contrast optics in a Zeiss photomicroscope. The antibody staining was done according to refs 13 and 14.

DNA extraction and Southern blotting. Single fly DNA extraction was done as described previously[11]. Restriction enzyme digestions, electrophoresis, blotting to Hybond N filters and hybridization procedures were done as described[30]. As probes for Southern analysis we used clones of *aurora*, *I* and *Bari1* with the following characteristics: the *aurora* clone is a 4,500-bp genomic fragment in an EcoRI site of the pUC19 vector; the *I* clone is a 3,500-bp genomic fragment in an EcoRI site of the pUC19 vector; the *Bari1* clone is a 1,100-bp genomic fragment in a HindIII site of the pUC8 vector.

Northern blot analysis for *crystal* and transposon piRNA detection. *Drosophila* testes were manually dissected in Ringer's solution. Total RNA, isolated from testes using Trizol (Invitrogen), was quantified by absorbance at 260 nm, and 20 μg of total RNA was resolved by 10% denaturing polyacrylamide/urea gel electrophoresis. After electrophoresis, the polyacrylamide gel was transferred to Hybond N+ (GE Healthcare) in 0.5× TBE by semidry transfer (Hoefer SemiPhor) at 20 V for 3 h. The RNA was crosslinked to the membrane by ultraviolet irradiation (1,200 μjoules cm^{-2}; Hoefer) and pre-hybridized in a buffer containing 6× SSPE, 5% Denhardt's solution, 0.5% SDS and 100 μg ml^{-1} salmon sperm DNA, for 4 h at 37 °C. The oligonucleotides, complementary to *rasi1* (5′-GAACUGUCGAACACCGGCUUCAG-3′), *rasi4* (5′-CCCGAGGAAC UGUCGAACACCG-3′), *opus* (5′-AAUAGUCAAGAUAGUUAAGAGGG-3′), *I* (5′-UGAUUAGAAUAAGGAGUUGGUUG-3′) and *roo* (5′-UCAUAUGAUC UGAGAAUAUGUUU-3′) elements were 5′-^{32}P-radiolabelled with polynucleotide kinase (Roche) and 30 μCi [γ-^{32}P]ATP. The ^{32}P-radiolabelled probes were purified using a Sephadex G-25 spin column (GE Healthcare) and hybridized for 16 h at 37 °C. After hybridization, membranes were washed twice with 2× SSC/ 0.1% (w/v) sodium dodecyl sulphate (SDS) and once with 1× SSC/0.1% (w/v) SDS for 15 min. The membranes were then exposed to autoradiography for 4–24 h. piRNA signals were normalized against that of the 5S rRNA.

Total RNA extraction for qRT–PCR. Total RNA was extracted from 30 mg of male and female germinal tissues using the RNAqueos-4 PCR Kit (AMBION) reagent following the manufacturer's protocol. To remove any DNA from the preparation, the samples were incubated with DNase I RNase free (AMBION) (1 U mg^{-1} RNA) at 37 °C for 15 min, in a total volume of 100 μl. After this treatment, the enzyme was inactivated with the DNase inactivation reagent (AMBION). DNase-treated RNA was precipitated at −80 °C overnight, and after centrifugation it was dissolved in 30 μl of nuclease-free water. The RNA concentration and purity were determined photometrically by measuring its absorbance at 260 nm and the A$_{260}$/A$_{280}$ ratio.

Quantitative real-time RT–PCR of *Stellate* and transposable elements. In the first-strand cDNA synthesis, 5 μg of total RNA were used as a template for oligonucleotide dT primed reverse transcription using SuperScript III RNaseH-reverse transcriptase (Invitrogen), according to manufacturer's instructions. Real-time RT–PCR was performed in the SmartCycler Real-time PCR (Cepheid). Relative abundance of the *Stellate* euchromatic (*Ste^eu*) transcript and the different transcripts of transposons in germinal tissues was determined using Fluo Cycle for SYBR green (Celbio) according to the manufacturer's

protocol. For quantification of the transcripts we used the 2^{ΔΔCt} method[29]. The primers we used are reported in Supplementary Table 1.

PCR amplification was performed in a final volume of 25 μl using standard cycling parameters (10 min 95 °C; 30 s 95 °C; 30 s specific annealing temperature; 30 s 72 °C with the latter three steps repeated 45 times). For *Stellate*, for each transposon and for *rp49* we constructed the standard curve and we calculated the primer amplification efficiency $E = [10^{-1/slope}] - 1$ using the standard curve. The efficiencies of amplification of the target genes and the control were approximately the same. We also found the amplicon melting curve to ensure that the desired amplified fragment was detected.

For the standard curve of *rp49* we used a clone provided by M. Berloco from the Genetics laboratory (University of Bari). For the *Stellate* and for the transposon standard curves we used the amplicon obtained by RT–PCR. The melting temperature of the amplicons was the same in each experiment.

In the 2^{ΔΔCt} method, the relative amount of specific transcript (fold change) was calculated as follows: $X_{test}/X_{control} = 2^{ΔΔCt}$; $ΔΔCt = (Ct_X - Ct_R)_{control} - (Ct_X - Ct_R)_{test}$ (where Ct_X is the Ct of the gene of interest and Ct_R is the Ct of the reference gene (*rp49*); test refers to the different cDNAs to analyse and control refers to the cDNA of reference). The error (coefficient of variation, CV) was calculated as $CV = \sqrt{cv_{test,X}^2 + cv_{test,R}^2}$.

PCR of genomic DNA. For PCR amplifications of genomic DNA, 1 μl of the single fly DNA preparation was used as a template with Platinum Taq (Invitrogen) in 50 μl reactions, according to the manufacturer's recommendations. The cycling conditions were as follow: 94 °C for 3 min; 25 cycles of 30 s denaturation at 94 °C, 30 s at specific annealing temperature, and 30 s at 72 °C; 72 °C for 10 min.

The primers used are: noc3400 upper, 5′-AGCGAAGCGAGGCGATGGAA-3′, position 14490264 in the *noc* gene on the II chromosome (2L); noc3400 lower, 5′-AATTGTTGGTTTGTTGTATGC-3′, position 14493702 in the *noc* gene (annealing temperature 54 °C); noc2100 upper, 5′-GGAAATGGTTGAAGAGT GTG-3′, position 14490280 in the *noc* gene; noc2100 lower, 5′-ATGGT GTGAGAATCCTCCGG-3′, position 14492380 in the *noc* gene (annealing temperature 55 °C).

Inverse PCR. A total of 5 μl of the single fly DNA preparation were used in restriction enzyme digestion with HindIII enzyme in 50 μl reaction. The obtained fragments were used in the 400 μl ligase (Roche) reaction at 4 °C overnight to promote intramolecular ligation. The ligase reaction was precipitated at −80 °C overnight and after centrifugation it was dissolved in 5 μl of the nuclease-free water and used for the PCR amplification with the *noc* new upper 5′-TCCCCTGTCCGCCCTGTGCT-3′ (position 14492309 in the *noc* gene) and IPCR2 5′-CTCCGATTTGGCTTCCTG-3′ (position 14490599 in the *noc* gene) primers (annealing temperature 57 °C).

PCR amplification of genomic DNA from the *Sco*-like mutant. To confirm that an *I* element is inserted in the *noc* gene we performed a DNA genomic amplification using the *noc* new upper primer, 5′-TCCCCTGTCCGCCCTGTGCT-3′ (see above), and the *I* lower primer, 5′-GGTGTTGGTGTGGTTGGTTG-3′, position 1343 in the *I*-element sequence M14954 (annealing temperature 55 °C).

30. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory, 1989).

LETTERS

Large, rare chromosomal deletions associated with severe early-onset obesity

Elena G. Bochukova[1]*, Ni Huang[2]*, Julia Keogh[1], Elana Henning[1], Carolin Purmann[1], Kasia Blaszczyk[1], Sadia Saeed[1], Julian Hamilton-Shield[3], Jill Clayton-Smith[4], Stephen O'Rahilly[1], Matthew E. Hurles[2] & I. Sadaf Farooqi[1]

Obesity is a highly heritable and genetically heterogeneous disorder[1]. Here we investigated the contribution of copy number variation to obesity in 300 Caucasian patients with severe early-onset obesity, 143 of whom also had developmental delay. Large (>500 kilobases), rare (<1%) deletions were significantly enriched in patients compared to 7,366 controls ($P < 0.001$). We identified several rare copy number variants that were recurrent in patients but absent or at much lower prevalence in controls. We identified five patients with overlapping deletions on chromosome 16p11.2 that were found in 2 out of 7,366 controls ($P < 5 \times 10^{-5}$). In three patients the deletion co-segregated with severe obesity. Two patients harboured a larger de novo 16p11.2 deletion, extending through a 593-kilobase region previously associated with autism[2–4] and mental retardation[5]; both of these patients had mild developmental delay in addition to severe obesity. In an independent sample of 1,062 patients with severe obesity alone, the smaller 16p11.2 deletion was found in an additional two patients. All 16p11.2 deletions encompass several genes but include SH2B1, which is known to be involved in leptin and insulin signalling[6]. Deletion carriers exhibited hyperphagia and severe insulin resistance disproportionate for the degree of obesity. We show that copy number variation contributes significantly to the genetic architecture of human obesity.

Although the rise in obesity prevalence is driven by environmental factors, there is considerable evidence that weight is highly heritable[1,7]. Genome-wide association studies (GWAS) have identified common single nucleotide polymorphisms (SNPs) associated with increased body mass index (BMI)[8–11]; however, together these account for a small percentage of the inherited variation in BMI. Studies in patients with severe early-onset obesity have led to the detection of rare variants, many of which have an impact on the leptin–melanocortin system involved in energy homeostasis[7,12,13]. To explore the contribution of copy number variants (CNVs) to obesity, we studied 300 Caucasians with severe early-onset obesity. Rare CNVs can cause developmental diseases[14]; as such we enriched for patients with developmental delay in addition to severe obesity (n = 143) in this discovery set (Supplementary Information). Genomic DNA was hybridized to the Affymetrix 6.0 array and the frequency of CNVs in patients was compared to 7,366 controls of European ancestry. Data were analysed using Affymetrix Power Tools and Birdsuite software[15]. We performed extensive post-calling quality control (QC) and filtering to arrive at a final set of CNVs (Supplementary Information). We detected 15,407 CNVs in the 284 patients passing QC and 403,098 CNVs in the 7,366 controls passing QC. The median number of CNV calls per individual was similar between patients and controls (53 and 55, respectively), as was the median size of CNVs (22 kilobases (kb) and 23.5 kb, respectively).

Following criteria set previously[16], we defined large, rare deletions as present in <1% of individuals and >500 kb in size. We observed a twofold enrichment of large, rare deletions in patients compared to controls (Table 1). We assessed the significance of this observation using conditional permutation tests controlling for differences in total number of CNV calls or data quality between patients and controls, and found it to be highly significant ($P = 0.0005$, Supplementary Information). Patients with developmental delay in addition to severe early-onset obesity were more likely to harbour rare, large deletions (Table 1). The trend towards enrichment of large, rare deletions remains in patients with severe obesity alone, but is no longer statistically significant.

We identified several rare CNVs recurrent among patients and enriched relative to controls (Table 2 and Supplementary Fig. 1). We confirmed these by multiplex ligation-dependent probe amplification (MLPA) and investigated co-segregation in families. We identified four patients with the 16p11.2 (29.5–30.1 megabases (Mb)) deletion reported in autism and mental retardation[2–4] (Table 2). These patients had mild developmental delay requiring special educational needs support; two had autistic spectrum behaviour. This frequency (~1% of all obesity patients, ~2% of patients with obesity and developmental delay) is slightly higher than the frequency (0.5–1%) in published studies[2,3,5], possibly due to the inclusion of patients with developmental delay in addition to severe obesity in our study (n = 143). The literature[17] suggests that 16p11.2 (29.5–30.1 Mb) deletions may be associated with being overweight[5] and with severe obesity in one patient[4]. Two of three patients

Table 1 | Global CNV burden analysis in patients with severe early-onset obesity

Samples	Type	Case rate	Case/control ratio	P value*	P value†
All patients	Losses and gains	0.2500	1.2996	0.0201	0.0433
	Losses	0.1127	2.0906	0.0005	0.0007
	Gains	0.1373	0.9917	0.4776	0.5800
Severe early-onset obesity only	Losses and gains	0.2089	1.0857	0.3150	0.3905
	Losses	0.0696	1.2917	0.2389	0.2884
	Gains	0.1392	1.0055	0.4790	0.5332
Severe early-onset obesity and developmental delay	Losses and gains	0.2937	1.5417	0.0098	0.0195
	Losses	0.1667	3.1318	0.0003	0.0001
	Gains	0.1270	0.9252	0.5701	0.6801

* Single-tailed P values derived from permutation conditioned on total number of calls per sample. Alternative hypothesis: ratio >1.
† Single-tailed P values derived from permutation conditioned on median absolute deviation (MAD). Alternative hypothesis: ratio >1.

[1]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK. [3]Bristol Children's Hospital, Bristol BS2 8BG, UK. [4]Genetic Medicine, St Mary's Hospital, Oxford Road, Manchester M13 9WL, UK.
*These authors contributed equally to this work.

Table 2 | Rare recurrent CNVs found in patients with severe early-onset obesity

Locus	CNV type	Frequency Cases	Frequency Controls	P value*	Subject	Genomic coordinates	BMI s.d.s./additional features	Inheritance
3p11.2	Gain	3	0	5.07×10^{-5}	1	89,245,197–89,343,751	3.8/−	n/a
					2	89,250,592–89,343,751	3.7/mild DD	n/a
					3	89,250,592–89,319,536	5.5/behavioural abnormalities	From obese parent
6p12.1	Loss	2	0	1.37×10^{-3}	4	52,875,284–52,892,054	4.2/DD, dysmorphic features	n/a
					5	52,875,284–52,892,054	4.0/−	n/a
8q24.3	Gain	2	0	1.37×10^{-3}	6	143,067,903–143,688,840	4.4/−	n/a
					7	143,268,033–143,634,461	6.0/disproportionate hyperinsulinaemia	n/a
10p15.3	Gain	2	1	4.02×10^{-3}	8	432,207–876,909	7.1/−	n/a
					9	541,873–818,440	6.6/DD, seizures	n/a
11q22.2†‡	Loss	2	0	3.71×10^{-2}	10	103,489,260–106,419,349	5.9/DD	n/a
				3.71×10^{-2}	11	105,716,030–108,818,442	5.3/DD, dysmorphic features	From obese parent
11q13.4	Gain	2	0	1.37×10^{-3}	12	71,980,493–72,106,059	4.4/−	n/a
					13	72,013,333–72,089,312	5.2/−	From obese parent
15q13.2-q13.3‡	Gain	2	2	7.84×10^{-3}	14	28,675,248–30,231,488	4.0/impaired speech, dyspraxia	n/a
					15	28,700,879–30,577,010	3.8/dyspraxia, memory impairment, DD	Found in obese sibling
16p11.2	Loss	3	2	1.37×10^{-6}	16	28,731,428–28,951,376	3.8/−	From obese parent
					17	28,655,035–28,951,376	3.9/−	From obese parent
					18	28,731,428–28,951,376	6.0/−	From obese parent
16p11.2	Loss	2	0	1.37×10^{-3}	19	28,291,978–30,085,309	4.1/DD	*De novo*
					20	28,443,702–30,099,397	3.9/DD	*De novo*
16p11.2‡	Loss	4	4	4.60×10^{-7}	9	29,448,112–30,227,810	6.6/DD, seizures	n/a
					21	29,448,112–30,134,433	4.9/DD, speech delay	From normal weight parent
					22	29,487,535–30,134,444	3.0/DD, autistic	From obese parent
					23	29,487,535–30,235,818	4.0/DD, autistic	n/a
17p13.3	Gain	2	2	7.84×10^{-3}	12	2,218,774–2,259,124	4.4/−	n/a
					24	2,224,814–2,256,880	3.3/DD, syndactyly, dysmorphic features	From obese parent
22q13	Gain	2	0	1.37×10^{-3}	25	49,246,176–49,313,898	3.7/mild DD	n/a
					12	49,246,176–49,349,366	4.4/−	n/a

Rare recurrent and MLPA validated CNVs found in patients with severe early-onset obesity and controls are shown. Details about the CNV locus (CNV type, chromosomal location, size, inheritance pattern) are listed; genes contained within each CNV are listed in the Supplementary Information. Phenotypic information on the subjects carrying the CNV is included where available. Severity of obesity is reported as BMI standard deviation score (s.d.s.). A minus (−) indicates no additional phenotype reported. All those with developmental delay were in the subgroup of patients with 'obesity and developmental delay' in Table 1. DD, developmental delay; n/a, not available.
* P value derived from two-sided Fisher's exact test.
† The two CNVs at this locus overlap at <50% of the sequence.
‡ CNVs identified in previous studies (prevalence in cases and controls in these studies is provided in the Supplementary Information for comparison).

harbouring duplications in this region were reported as underweight[17]. However, our study design did not permit detailed investigation of the contribution of 16p11.2 (29.5–30.1 Mb) deletions to altered body weight.

A large 15q11-q13 duplication is well-known to be associated with autism[4,18]; we found smaller, distinct 15q13.2-q13.3 duplications in two patients, which have previously been observed in controls[19], as well as patients with mental retardation, autism and other phenotypes[20] (Table 2). We also identified two patients with a 11q22-q23 deletion overlapping minimally with a CNV associated with mental retardation[20,21] on a single brain-expressed gene *GUCY1A2* (Table 2). Some of the CNV carriers in these studies were obese[20]. The finding of these rare recurrent CNVs in our cohort may point to a shared mechanistic basis for these neurodevelopmental diseases. Rare CNVs encompassing brain-expressed genes such as *TSNARE1* and *PDE2A* were unique to patients with severe early-onset obesity and require further investigation.

The commonest CNV enriched in patients with severe obesity was identified in five unrelated patients harbouring deletions on 16p11.2 (Table 2) with a minimal overlapping segment of 220 kb (Fig. 1a). This 220-kb deletion (28.73–28.95 Mb) was found in 2 out of 7,366 controls (Supplementary Information); BMI data were unavailable. The minimal deleted region contains genes involved in neurological diseases (*TUFM*, *ATP2A1*), immunity (*CD19*, *NFATC2IP*, *LAT*) and genes of unknown function (*ATXN2L*, *RABEP2*, *SPNS1*), as well as *SH2B1*, which encodes an adaptor protein involved in leptin and insulin signalling. Disruption of *Sh2b1* in mice results in obesity and severe insulin resistance[22]. As CNVs can unmask a recessive allele, we sequenced *SH2B1* in these patients but did not find any coding/splice site mutations (Supplementary Information).

The 220-kb 16p11.2 deletion (28.73–28.95 Mb) was seen in three patients and co-segregated with severe early-onset obesity alone (Fig. 1b). A longer ~1.7-Mb deletion detected in two patients (Fig. 1c) encompassed the 220-kb deletion and extended through a 593-kb region (29.5–30.1 Mb) where deletions are associated with autism and mental retardation[2–5,17] (Fig. 1c). These longer deletions occurred *de novo*; both these patients had mild developmental delay in addition to their severe obesity. These findings are consistent with a role for the *SH2B1*-containing 220-kb region (28.73–28.95 Mb) in severe obesity and the 29.5–30.1-Mb region in brain development. We used MLPA to screen for this rare *SH2B1*-containing (28.73–28.95 Mb) deletion in an independent set of 1,062 Caucasian patients with severe obesity alone from the same cohort, as no other comparable cohort of the same ancestry and size exists. We found two additional patients with this deletion. The deletion co-segregated with obesity in one family whose samples were available (Supplementary Information). The prevalence of the *SH2B1*-containing deletion was lower in the replication cohort (2 out of 1,062) compared to the discovery set with severe obesity alone (3 out of 157). Overestimation of case frequency in the primary screen is a recognized phenomenon especially with the investigation of rarer variants[23]. Overall the prevalence of the *SH2B1*-containing deletion in patients with severe early-onset obesity alone (5 out of 1,219; 0.41%) is significantly greater than in controls (2 out of 7,366; 0.027%) ($P < 0.001$).
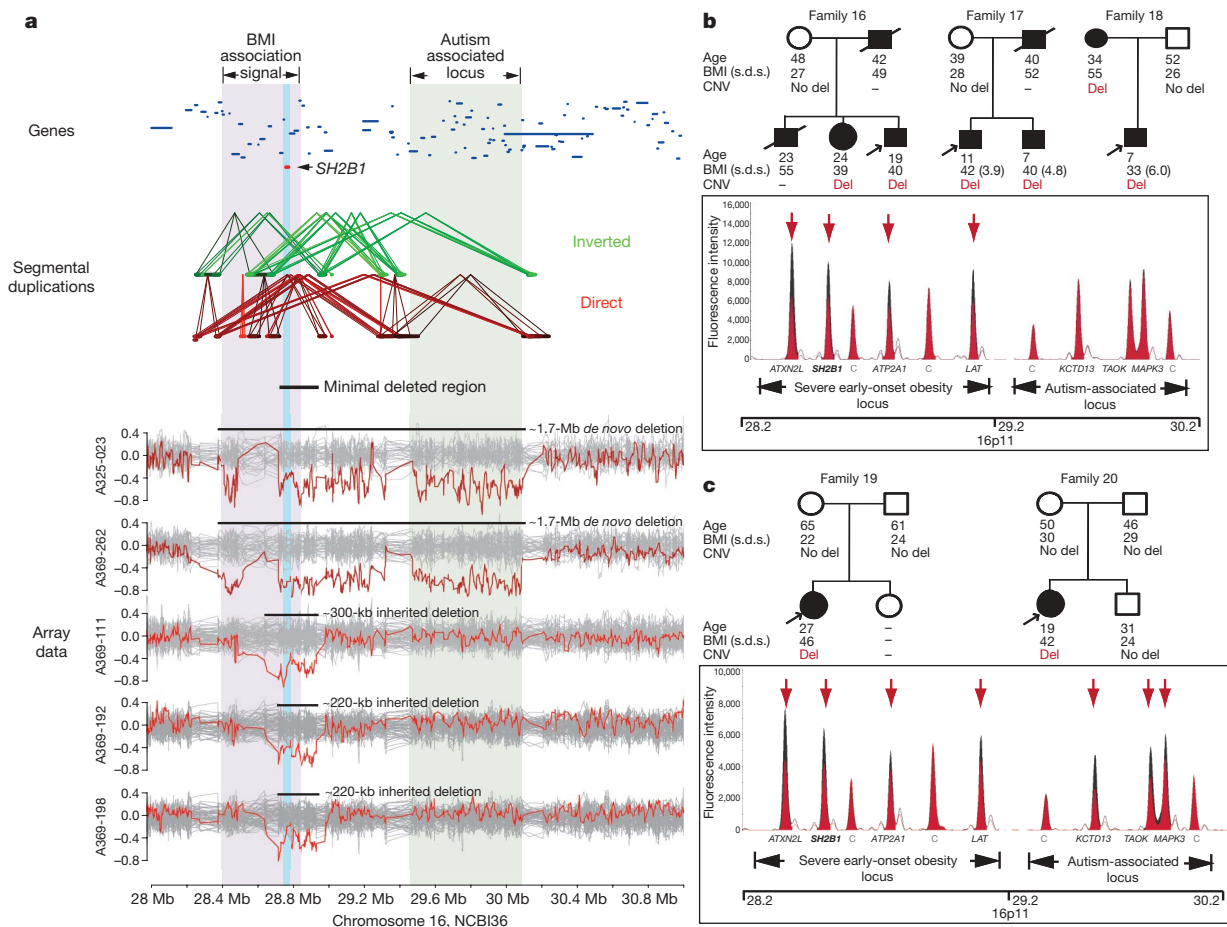
Figure 1 | Discovery of 16p11.2 CNV associated with severe early-onset obesity. a, Affymetrix 6.0 array data for five patients with deletions at 16p11.2 is shown. Log$_2$ ratios of the five samples are highlighted in dark red, with other samples in the same genotyping plate shown in grey. The structure of extensive segmental duplication that extends to the flanking regions is shown. Annotation of the segmental duplications was taken from UCSC Genome Browser and the darkness of colour coding represents sequence similarity between the duplicated pairs. Protein-coding genes are represented by dark-blue lines; *SH2B1* is highlighted in red and by blue vertical shading. The light-pink vertical shading indicates the range of a previous BMI-association signal found in two genome-wide association studies[10,11]; the light-grey vertical shading indicates the reported autism-associated CNV region[2–4,17]. **b, c**, Pedigrees are shown for the five patients with 16p11.2 deletions. Families are numbered as in Table 2. Filled symbols represent patients with severe early-onset obesity; arrows indicate probands. Age, BMI (body mass index) and BMI s.d.s. (standard deviation score) for children are included where available. Presence (del) or absence (no del) of the 16p11.2 *SH2B1*-containing CNV is shown where known. Representative MLPA data are shown. MLPA probes for genes in the region of interest are shown. The MLPA target regions labelled as C are control probes located either on chromosome 16 but outside the deleted region or on other chromosomes (Supplementary Information). Patient MLPA traces are in red, overlaid upon the normal control MLPA traces in black. Arrows point to the deleted probes. **b**, Three probands in whom the 16p11.2 *SH2B1*-containing deletion co-segregates with severe early-onset obesity alone. **c**, Two probands harbouring larger *de novo* 16p11.2 deletions that also encompass a known autism-associated locus and are associated with developmental delay and severe early-onset obesity.

The breakpoints of both classes of deletion are embedded within complex, segmentally duplicated regions of 16p11.2 containing directly-oriented, highly-similar (>98% sequence similarity) duplicated sequences greater than 15 kb in length (Fig. 1a). This observation strongly supports the hypothesis that these deletions arise through non-allelic homologous recombination between duplicated sequences[14].

We characterized the phenotype of patients with the *SH2B1*-containing deletion. Patients rapidly gained weight in the first years of life (Fig. 2a); their excess weight was predominantly fat mass (Fig. 2b). They exhibited hyperphagia with increased *ad libitum* food intake (Fig. 2c). Fasting plasma insulin levels were disproportionately elevated compared to age- and obesity-matched controls and patients with other obesity syndromes including MC4R deficiency, where we have reported disproportionate hyperinsulinaemia[12] (Fig. 2d). Fasting plasma glucose concentrations were in the normal range. Adult patients with the *SH2B1*-containing deletion had delayed, exaggerated insulin secretion in response to an oral glucose load compared to equally obese controls (Fig. 2e).

The similarities with the human leptin-receptor-deficient phenotype are striking[13] and consistent with a role for central disruption of SH2B1 (ref. 6). Deletion of peripheral *Sh2b1* in rodents impairs insulin signalling in muscle, liver and adipose tissue[24], which might explain the severe and disproportionate insulin resistance in these patients. As heterozygous deletion of *Sh2b1* in mice leads to obesity on a high-fat diet[6], haploinsufficiency may be a plausible mechanism underlying the phenotype seen in humans (Supplementary Information).

Common SNPs near the *SH2B1* locus have been associated with BMI in GWAS[10,11]. We investigated CNVs upstream or downstream of known obesity genes or other BMI-associated loci. We identified a 153-kb deletion 60 kb downstream from *MC4R* (chromosome 18: 55,975,710–56,128,721) in a single patient (Supplementary Fig. 2a), inherited from an obese father (Supplementary Fig. 3). This deletion removes a BMI association interval[25] (Supplementary Fig. 2a) and its distal end partially overlaps with a CNV found in a single control (chromosome 18: 55,804,160–56,063,854). Further studies will be required to ascertain the significance of this finding. A 100-kb
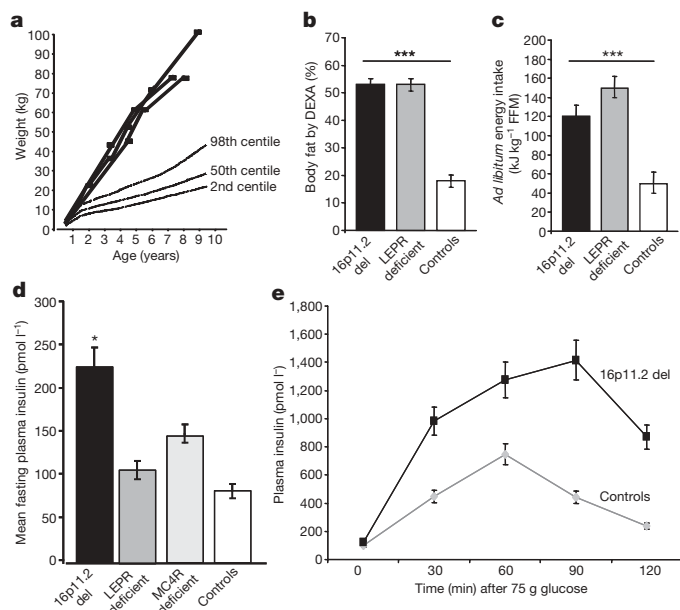
Figure 2 | Metabolic phenotype of carriers of the 16p11.2 deletion.
a, Weights for three boys with the 16p11.2 deletion are plotted on growth charts based on UK reference data. **b, c,** Percentage body fat measured by dual energy X-ray absorptiometry (DEXA) (**b**) and *ad libitum* energy intake at a test meal (adjusted for fat free mass (FFM) in kilograms) (**c**) is shown for 16p11.2 deletion carriers ($n = 5$), leptin-receptor-deficient subjects ($n = 10$) and normal weight controls ($n = 35$). Means ± s.e.m. (error bars) are indicated. **d,** Fasting plasma insulin levels for 16p11.2 deletion carriers ($n = 6$), leptin-receptor-deficient subjects ($n = 10$), MC4R-deficient subjects ($n = 35$), and age and BMI s.d.s. matched controls from the GOOS cohort ($n = 535$) are shown. Plasma insulin values were analysed after $\log_{10}$-transformation and are presented as geometric mean (1 s.e. range) after back-transformation. **e,** Plasma insulin levels in response to a 75 g oral glucose load are shown in adult 16p11.2 deletion carriers ($n = 3$) and severely obese adult controls matched for fasting plasma insulin levels ($n = 10$) (Supplementary Information). Plasma insulin values were analysed after $\log_{10}$-transformation and are presented as geometric mean (1 s.e. range) after back-transformation. Results were compared using unpaired t-tests. All P values were two-sided. $P < 0.05$ was considered statistically significant. *$P < 0.01$; ***$P < 0.001$.

duplication encompassing part of *LEPR* was found in a single patient but also in 16 controls (Supplementary Fig. 2b). Although we did not find any rare CNVs close to *FTO*[8,9], we detected the common 45-kb deletion upstream of *NEGR1*[10,11] at comparable prevalence in patients and controls (Supplementary Fig. 2c).

We report here the first rare CNVs associated with severe early-onset obesity. We show that deletion of 16p11.2 is associated with highly penetrant familial severe early-onset obesity and severe insulin resistance. Although the contribution of other genes or non-coding genetic material cannot be excluded, the phenotype is consistent with a role for SH2B1 in human energy homeostasis and glucose metabolism. We did not identify any coding/splice site mutations in *SH2B1* in 500 Genetics of Obesity Study (GOOS) patients (Supplementary Information). Segmental duplications on 16p11.2 and concomitant increase in deletion rate through non-allelic homologous recombination may explain why deletions are a relatively more frequent source of loss of function in *SH2B1* than at other monogenic obesity loci. Novel CNVs unique to patients with severe early-onset obesity and the increased CNV burden (which is not fully accounted for by *SH2B1* deletions) suggest a role for additional genes in the aetiology of severe obesity. Our findings indicate the presence at the same loci of both common variants influencing susceptibility to common obesity and more highly penetrant rare variants, including CNVs, associated with severe early-onset forms of the disease. Strategies aimed at looking for rare variants near common susceptibility loci may well prove to be fruitful in other common complex diseases.

METHODS SUMMARY

A discovery set of 300 UK Caucasian patients from the Genetics of Obesity Study (GOOS) cohort were randomly selected for this study (143 had developmental delay). All patients had severe obesity defined as a BMI standard deviation score (s.d.s.) >3 and onset of obesity before 10 years of age (Supplementary Information). Mutations in *LEPR*, *POMC* and *MC4R* were excluded by direct nucleotide sequencing and a karyotype performed. DNA samples were run on Affymetrix Genome-Wide Human SNP Array 6.0 by Aros, Inc., and compared to control data on the same platform obtained on over 7,000 controls recruited from the Wellcome Trust Case Control Consortium 2 (WTCCC2) and the Genetic Association Information Network (GAIN). The WTCC2 data set contains 6,000 individuals, 3,000 from the 1958 British Birth Cohort and 3,000 from the UK Blood Service Collection, used as common controls in genome-wide association studies of 13 disease conditions undertaken by Wellcome Trust Case Control Consortium 2 (http://www.wtccc.org.uk/ccc2). The GAIN data set contains 1,442 individuals of European ancestry used as part of the control cohorts in genome-wide association studies of schizophrenia and bipolar disorders. The data set was obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nih.gov/gap through dbGaP accession numbers phs000021.v1.p1, phs000021.v2.p1 and phs000017.v1.p1. Further methods for CNV calling, analysis and MLPA are included in the Supplementary Information. All clinical studies were conducted in accordance with the principles of the Declaration of Helsinki using methods that have previously been reported[13].

1. Friedman, J. M. Obesity in the new millennium. *Nature* **404**, 632–634 (2000).
2. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
3. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
4. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
5. Bijlsma, E. K. *et al.* Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur. J. Med. Genet.* **52**, 77–87 (2009).
6. Ren, D. *et al.* Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *J. Clin. Invest.* **117**, 397–406 (2007).
7. Walley, A. J., Asher, J. E. & Froguel, P. The genetic contribution to non-syndromic human obesity. *Nature Rev. Genet.* **10**, 431–442 (2009).
8. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
9. Dina, C. *et al.* Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genet.* **39**, 724–726 (2007).
10. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).
11. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genet.* **41**, 18–24 (2009).
12. Farooqi, I. S. *et al.* Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N. Engl. J. Med.* **348**, 1085–1095 (2003).
13. Farooqi, I. S. *et al.* Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. *N. Engl. J. Med.* **356**, 237–247 (2007).
14. Lee, J. A. & Lupski, J. R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52**, 103–121 (2006).
15. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genet.* **40**, 1253–1260 (2008).
16. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
17. Fernandez, B. A. *et al.* Phenotypic spectrum associated with *de novo* and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J. Med. Genet.* doi:10.1136/jmg.2009.069369 (15 September 2009).
18. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
19. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genet.* **40**, 322–328 (2008).
20. van Bon, B. W. *et al.* Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *J. Med. Genet.* **46**, 511–523 (2009).
21. Poot, M. *et al.* Recurrent copy number changes in mentally retarded children harbour genes involved in cellular localization and the glutamate receptor complex. *Eur. J. Hum. Genet.* doi:10.1038/ejhg.2009.120 (22 July 2009).
22. Ren, D., Li, M., Duan, C. & Rui, L. Identification of SH2-B as a key regulator of leptin sensitivity, energy balance, and body weight in mice. *Cell Metab.* **2**, 95–104 (2005).

23. Hirschhorn, J. N. & Altshuler, D. Once and again—issues surrounding replica-tion in genetic association studies. *J. Clin. Endocrinol. Metab.* **87**, 4438–4441 (2002).

24. Morris, D. L., Cho, K. W., Zhou, Y. & Rui, L. SH2B1 enhances insulin sensitivity by both stimulating the insulin receptor and inhibiting tyrosine dephosphorylation of IRS proteins. *Diabetes* **58**, 2039–2047 (2009).

25. Loos, R. J. *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genet.* **40**, 768–775 (2008).

Author Contributions I.S.F., M.E.H., E.G.B. and N.H. designed the study, analysed the data and wrote the paper. All the authors contributed to data collection and/or processing, as well as the current version of 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 I.S.F. (isf20@cam.ac.uk) or M.E.H. (meh@sanger.ac.uk).

A new highly penetrant form of obesity due to deletions on chromosome 16p11.2

A list of authors and their affiliations appears at the end of the paper

Obesity has become a major worldwide challenge to public health, owing to an interaction between the Western 'obesogenic' environment and a strong genetic contribution[1]. Recent extensive genome-wide association studies (GWASs) have identified numerous single nucleotide polymorphisms associated with obesity, but these loci together account for only a small fraction of the known heritable component[1]. Thus, the 'common disease, common variant' hypothesis is increasingly coming under challenge[2]. Here we report a highly penetrant form of obesity, initially observed in 31 subjects who were heterozygous for deletions of at least 593 kilobases at 16p11.2 and whose ascertainment included cognitive deficits. Nineteen similar deletions were identified from GWAS data in 16,053 individuals from eight European cohorts. These deletions were absent from healthy non-obese controls and accounted for 0.7% of our morbid obesity cases (body mass index (BMI) $\geq 40\,\mathrm{kg\,m^{-2}}$ or BMI standard deviation score ≥ 4; $P = 6.4 \times 10^{-8}$, odds ratio 43.0), demonstrating the potential importance in common disease of rare variants with strong effects. This highlights a promising strategy for identifying missing heritability in obesity and other complex traits: cohorts with extreme phenotypes are likely to be enriched for rare variants, thereby improving power for their discovery. Subsequent analysis of the loci so identified may well reveal additional rare variants that further contribute to the missing heritability, as recently reported for *SIM1* (ref. 3). The most productive approach may therefore be to combine the 'power of the extreme'[4] in small, well-phenotyped cohorts, with targeted follow-up in case-control and population cohorts.

The extent to which copy-number variants (CNVs) might contribute to the missing heritability of common disorders is currently under debate[2]. Because most common simple CNVs are well tagged by single nucleotide polymorphisms (SNPs), it has recently been suggested that common CNVs are unlikely to contribute substantially to the missing heritability[5]. However, rare variants or recurring CNVs that have arisen on multiple independent occasions are unlikely to be captured by SNP tagging, and their identification will require alternative approaches.

We have previously proposed that cohorts with extreme phenotypes that include obesity may be enriched for rare but very potent risk variants[4,6]. Here we investigate 312 subjects, from three centres in the UK and France, presenting with congenital malformations and/or developmental delay in addition to obesity as defined previously[6,7] (see Methods). Known syndromes (for example, Prader–Willi and fragile X) were excluded. A combination of array comparative genomic hybridization (aCGH), genotyping arrays, quantitative PCR (qPCR) and multiplex ligation-dependent probe amplification (MLPA) was used to identify and confirm the presence of a heterozygous deletion on 16p11.2 in nine individuals (2.9%). These deletions, estimated to be a total of 740 kilobases (kb) in size (one copy of a segmental duplication plus 593 kb of unique sequences; Fig. 1a), have previously been associated to varying extents with autism, schizophrenia and developmental delay[8–11]; however, the observed
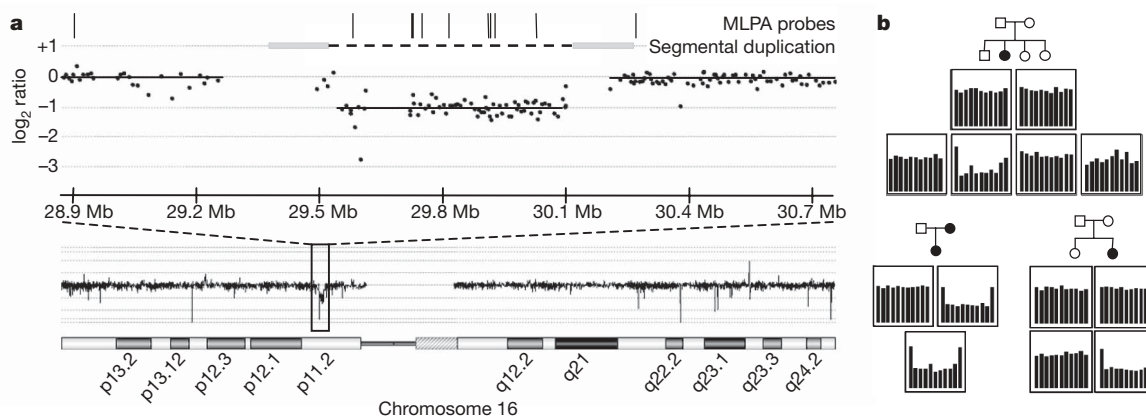

Figure 1 | Identification and validation of deletions at 16p11.2. a, aCGH data showing the location of the 16p11.2 deletion. The data show the log$_2$ intensity ratio for a deletion carrier compared with that for an undeleted control sample. Grey bars connected by a broken line denote the segmental duplication flanking the deletion region. Vertical bars indicate the positions of the probe pairs used for MLPA validation. Note that CGH and genotyping array probes targeted against segmental duplications may not accurately report copy number as a result of the increased number of homologous sequences in the diploid state. Genome coordinates are in accordance with the hg18 build of the reference genome. **b,** MLPA validation of 16p11.2 deletions. Representative MLPA results are shown, illustrating one instance of maternal transmission and two instances of *de novo* deletions. Genotyping data excluded the possibility of non-paternity. Full results for MLPA validation and inheritance analysis are shown in Supplementary Fig. 1. Each panel shows the relative magnitude of the normalized, integrated signal at each probe location, in order of chromosomal position of the MLPA probe pairs as indicated in **a**. Each panel corresponds to its respective position on the associated pedigree, as shown.

frequency of deletions in our cohort is appreciably higher than the reported frequencies in the cohorts from the previous studies (less than 1%), which did not include obesity as an inclusion criterion.

A parallel, independent survey of aCGH and SNP-CGH data from eight cytogenetic centres in France, Switzerland and Estonia, involving 3,947 patients with developmental delay and/or malformations but this time without selection for obesity, revealed 22 unrelated cases with similar deletions (0.6%). This is a frequency consistent with those found in the previous studies[8–11], but is significantly lower than for the above cohort, which only included obese subjects ($P = 2.2 \times 10^{-4}$, Fisher's exact test).

Analysis of the available clinical data for these 22 new carriers indicated that, in addition to the ascertained cognitive deficits or behavioural abnormalities (including hyperphagia, specifically identified in at least nine cases; see Supplementary Table 1), a 16p11.2 deletion gave rise to a strongly expressed obesity phenotype in adults, with a more variable phenotype in childhood. All four teenagers and adults carrying a deletion were obese, whereas child carriers were also frequently either obese (4 of 15) or overweight (2 of 15), a tendency that has previously been noted[11]; the very young (under 2 years old) were of normal weight. This age-dependent penetrance was observed in all instances of deletions for which phenotypic data were available, whether from this study or from previously published reports[10–15], and regardless of ascertainment (Fig. 2; see Supplementary Tables 2 and 3).

Taken together, the data from these parallel studies suggest a possible direct association of deletions at 16p11.2 with obesity, distinct from their cognitive phenotype. Also identified in these cohorts were instances of the reciprocal duplication, which has also been implicated in neurodevelopmental disorders, but with a variable phenotype and lower penetrance[9,10,12]. The frequency of the duplication in the two cohorts (12 of 4,183 (0.3%)) was consistent with previous reports for patients with cognitive deficits (0.3–0.7%)[10,12]. Carriers of the duplication neither were obese nor had reported hyperphagia.

To investigate further the association of 16p11.2 deletions with obesity, and to estimate the extent to which it is observed independently of ascertainment for neurodevelopmental symptoms, we

Figure 2 | Dependence of BMI on age in subjects having a deletion at 16p11.2. Data are shown for all individuals carrying a deletion for whom phenotypic data were available. Similar data from this study only are shown in Supplementary Figs 2 and 3. Lines denote the thresholds corrected for age and gender (solid, male; broken, female) for obesity and morbid obesity. Squares, male; circles, female; black, ascertained for developmental delay; grey, not ascertained for developmental delay; filled, ascertained for obesity; open, not ascertained for obesity; diamonds, first-degree relative of proband; crosses, previously published data[10–15]. The 31-year-old male with a BMI of about $20 \, \text{kg m}^{-2}$ was diabetic, as determined by a fasting blood glucose of more than 7 mM.

performed algorithmic and statistical analyses of genome-wide SNP genotyping data (see Table 1) from Swiss (CoLaus[16]), Finnish (NFBC1966 (ref. 17)) and Estonian (EGPUT[18]) general population cohorts (11,856 subjects in total), from child obesity and adult morbid obesity case-control cohorts[6,19,20] (1,224 and 1,548 subjects, respectively), from an extreme early-onset obesity cohort (SCOOP, 931 subjects) and from 141 patients undergoing bariatric weight-loss surgery (see Methods); in total, we identified 17 instances of deletions (and four duplications) with no significant gender bias (Table 1). In addition, we identified two further unrelated carriers of a deletion from 353 members of 149 families with sibling pairs discordant for obesity (SOS Sib Pair Study[21]). When DNA was available for further analysis (15 of 19 samples), the presence of a deletion was validated by using MLPA (Fig. 1b) or qPCR; the remaining deletions were validated by applying a second independent algorithm to the data. With the exception of a single individual who is apparently diabetic (fasting blood glucose more than 7 mM), all adult carriers of such deletions were obese, the majority being morbidly obese; similarly, each of the seven child or adolescent carriers had a BMI in the top 0.1% of the population range for their age and gender. None of the individuals ascertained on the basis of their obesity had any reported developmental delay or cognitive deficit; four subjects were reported as having hyperphagia.

To enable sufficient statistical power to give robust conclusions, we combined data from the population and obesity cohorts in an overall case-control association analysis (the samples from sib-pair families were excluded to avoid complications due to their relatedness). In comparison with lean or normal weight subjects (see Table 1 and Methods), 16p11.2 deletions were associated with obesity ($P = 5.8 \times 10^{-7}$, Fisher's exact test; odds ratio 29.8, 95% confidence limits 3.9 and 225) and morbid obesity ($P = 6.4 \times 10^{-8}$; odds ratio 43.0, 95% confidence limits 5.6 and 329) at or near genome-wide levels of significance. Expanding the control group to include all non-obese individuals increased the significance to $P = 4.2 \times 10^{-9}$ (obese) and $P = 6.1 \times 10^{-10}$ (morbidly obese).

Previous reports have indicated that these deletions are frequently not inherited from either parent but arise *de novo*, possibly by non-allelic homologous recombination between the more than 99% sequence-identical segmental duplications flanking the deleted region[11,14]. Therefore, where possible we investigated the parents of carriers of deletions, identifying 11 cases of maternal transmission and 4 of paternal transmission. The available data showed that all first-degree relatives carrying a deletion were also obese (Supplementary Table 1). In ten instances the deletion was apparently *de novo* (see Fig. 1b). Extrapolation to our full data set indicates that about 0.4% of all morbidly obese cases are due to an inherited 16p11.2 deletion. The frequency of *de novo* events is consistent with a previous report, in which ascertainment was for developmental delay and/or congenital anomalies[11]; by contrast, deletions are reported to be almost exclusively *de novo* in autistic subjects[8–10].

Although they may be heterogeneous in nature, these deletions are highly likely to be the causal variants, representing the second most frequent genetic cause of obesity after point mutations in *MC4R*[22,23]. Their repeated *de novo* occurrence is likely to result in a lack of linkage disequilibrium with any other flanking variant—no consistent haplotype has been identified by analysis of the available surrounding genotypes. To assess the effect of a deletion on the expression of nearby genes (for example, the obesity GWAS-associated *SH2B1* locus 800 kb distant[24]), we analysed available transcript data for subcutaneous adipose tissue samples from the discordant sibling cohort. Comparisons of the two subjects carrying a deletion with their corresponding non-obese siblings, and with other obese and non-obese subjects (Supplementary Fig. 4 and Supplementary Tables 4 and 5), showed that many, although not all, transcripts from within the deletion had a markedly decreased abundance (0.4–0.7-fold). In contrast, no clear evidence was found for consistent *cis* effects of the deletion on the abundance of messenger RNAs encoded by genes
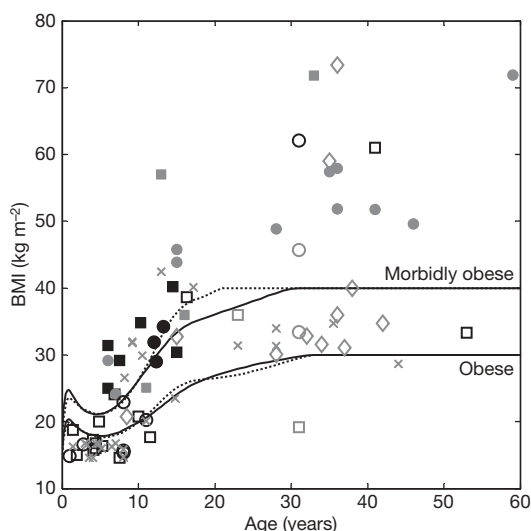
Table 1 | Frequency of detected 16p11.2 deletions in multiple cohorts

Cohort	Deletions/total					Technology
	Lean/normal	Overweight	Obese	Morbidly obese	Total	
Ascertained for cognitive deficits/malformations and obesity						
Lille/Strasbourg*					8/279	qPCR, aCGH
London*					1/33	aCGH, MLPA
Ascertained for cognitive deficits/malformations						
French–Swiss cytogenetic clinical diagnostic group*					21/3,870	aCGH, QMPSF, qPCR, FISH
Estonian cases of cognitive deficit*					1/77	Illumina CNV370-Duo, qPCR
Ascertained for obesity						
Swedish families with discordant siblings†§	0/140	0/54	0/115	2/44	2/353	Illumina 610K-Quad, MLPA
French adult case-control†	0/669	0/174	–	4/705	4/1,548	Illumina CNV370-Duo, MLPA
French child case-control‡	0/530	0/51	1/260	3/383	4/1,224	Illumina CNV370-Duo, MLPA
British extreme early-onset obesity (SCOOP)‡	–	–	–	3/931	3/931	Affymetrix 6.0, MLPA
French bariatric weight-loss surgery†	–	–	0/15	2/126	2/141	Illumina 1M-duo, MLPA
Population cohorts (origin)						
NFBC1966 (Finnish)†	1/3,148	0/1,622	1/434	1/42	3/5,246	Illumina CNV370-Duo
CoLaus (Swiss)†	0/2,675	0/2,049	0/830	0/58	0/5,612	Affymetrix 500K
EGPUT (Estonian)†	0/412	0/358	1/213	0/15	1/998	Illumina CNV370-Duo, qPCR
Total without ascertainment for cognitive deficits/malformations§	1/7,434	0/4,254	3/1,752	13/2,260		

For each cohort, 16p11.2 deletions were identified and validated with the indicated technologies. Where full phenotypic data were available, members of cohorts were categorized in accordance with the appropriate obesity criteria (see Supplementary Information): *not categorised, complete phenotypic data not available; †BMI thresholds for overweight, obese and morbidly obese were at least 25 kg m^{-2}, at least 30 kg m^{-2} and at least 40 kg m^{-2}, respectively; ‡BMI thresholds for overweight, obese and morbidly obese were the 90th centile, 97th centile and 4 standard deviations above the mean, respectively, corrected for age and gender. §Discordant siblings were not included in totals because of relatedness. QMPSF, quantitative multiplex PCR of short fluorescent fragments; FISH, fluorescence *in situ* hybridization.

flanking the deletion. In addition, global analysis of this data set has not identified any *trans*-acting expression quantitative trait loci either within or nearby the deletion.

Thus, although we cannot completely exclude the possibility that a 16p11.2 deletion affects the expression of nearby genes (for instance, its impact may be different in other tissues), the expression analysis described strongly indicates that the observed phenotypes are likely to be due to haploinsufficiency of one or more of the about 30 genes within the deleted region. Indeed, rather than being due to a single haploinsufficiency, the phenotype may well result from the deletion of multiple genes with an impact on pathways central to the development of obesity (see Supplementary Table 5). Functional network analysis of the deleted genes has led to the suggestion of a similar multigene effect for the cognitive phenotype[8]. The extent to which there is overlap between the genes involved in the obesity and cognitive phenotypes remains to be elucidated.

There is a strong correlation between developmental and cognitive disabilities and the prevalence of obesity: patients with autism or who have learning disabilities have a greatly increased risk of obesity[25], and the severely obese exhibit significant cognitive impairment[26]. Possible explanations include a direct causal relationship between obesity and developmental delay, the involvement of the same or related regulatory pathways, or different outcomes of the same set of behavioural disorders with complex pleiotropic effects and variable ages of onset and expressivities. The higher frequency of 16p11.2 deletions in the cohort ascertained for both phenotypes (2.9%), compared with cohorts ascertained for either phenotype alone (0.4% and 0.6%, respectively), confirms their impact on both obesity and developmental delay, adding to the evidence that these two phenotypes may be fundamentally interrelated.

METHODS SUMMARY

Obesity. Definitions for overweight, obesity and morbid obesity were based on previous studies[6,7]: for adults, BMI $\geq$ 25, 30 and 40 kg m^{-2} respectively; for children, BMI respectively above the 90th, 97th centiles and at least four standard deviations above the mean, calculated according to their age and gender from a French reference population[27,28].

Statistics. All reported statistical tests used Fisher's exact test[29], performed on contingency tables constructed for the number of subjects carrying or lacking a 16p11.2 deletion versus the obesity status or ascertainment of the individual. Because no homozygous deletions were observed, it was unnecessary to make a prior distinction between recessive, additive and dominant models of disease risk. Odds ratios and 95% confidence limits were calculated as described[30].

CNV discovery. Subjects ascertained for cognitive deficit/malformations with or without obesity were selected from those clinically referred for genetic testing; 16p11.2 deletions were identified in these individuals by standard clinical diagnostic procedures. Algorithmic analyses of GWAS data were performed variously using the cnvHap algorithm, a moving-window average-intensity procedure, a Gaussian mixture model, QuantiSNP, PennCNV, BeadStudio GT module, and Birdseed. When experimental validation was not possible, at least two independent algorithms were used for each data set.

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

1. Walley, A. J., Asher, J. E. & Froguel, P. The genetic contribution to non-syndromic human obesity. *Nature Rev. Genet.* **10**, 431–442 (2009).
2. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Stutzmann, F. *et al.* Loss-of-function mutations in SIM1 cause a specific form of Prader–Willi-like syndrome. *Diabetologia* **52**, S104 (2009).
4. Froguel, P. & Blakemore, A. I. F. The power of the extreme in elucidating obesity. *N. Engl. J. Med.* **359**, 891–893 (2008).
5. Conrad, D. F. *et al.* Origins and functional effect of copy number variation in the human genome. *Nature* advance online publication doi:10.1038/nature08516 (7 October 2009).
6. Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genet.* **41**, 157–159 (2009).
7. Farooqi, I. S. & O'Rahilly, S. Recent advances in the genetics of severe childhood obesity. *Arch. Dis. Child.* **83**, 31–34 (2000).
8. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
9. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
10. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
11. Bijlsma, E. K. *et al.* Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur. J. Med. Genet.* **52**, 77–87 (2009).
12. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet.* **41**, 1223–1227 (2009).
13. Ghebranious, N., Giampietro, P. F., Wesbrook, F. P. & Rezkalla, S. H. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. *Am. J. Med. Genet. A.* **143A**, 1462–1471 (2007).
14. Fernandez, B. A. *et al.* Phenotypic spectrum associated with *de novo* and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J. Med. Genet.*, doi:10.1136/jmg.2009.069369 (in the press).
15. Shimojima, K., Inoue, T., Fujii, Y. & Ohno, K. Yamamoto, T. A familial 593 kb microdeletion of 16p11.2 associated with mental retardation and hemivertebrae. *Eur. J. Med. Genet.* **52**, 433–435 (2009).

16. Firmann, M. *et al.* Prevalence of obesity and abdominal obesity in the Lausanne population. *BMC Cardiovasc. Disord.* **8**, 330, doi:10.1186/1471-2261-8-6 (2008).

17. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35–46 (2009).

18. Nelis, M. *et al.* Genetic structure of Europeans: a view from the north-east. *PLoS One* **4**, e5472, doi:10.1371/journal.pone.0005472 (2009).

19. Balkau, B., Eschwege, E., Tichet, J. & Marre, M. Proposed criteria for the diagnosis of diabetes: evidence from a French epidemiological study (D.E.S.I.R.). *Diabetes Metab.* **23**, 428–434 (1997).

20. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

21. Jernås, M. *et al.* Regulation of carboxylesterase 1 (CES1) in human adipose tissue. *Biochem. Biophys. Res. Commun.* **383**, 63–67 (2009).

22. Yeo, G. S. *et al.* A frameshift mutation in *MC4R* associated with dominantly inherited human obesity. *Nature Genet.* **20**, 111–112 (1998).

23. Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. A frameshift mutation in human *MC4R* is associated with a dominant form of obesity. *Nature Genet.* **20**, 113–114 (1998).

24. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).

25. Chen, A. Y., Kim, S. E., Houtrow, A. J. & Newacheck, P. W. Prevalence of obesity among children with chronic conditions. *Obesity (Silver Spring)*, **18**, 210–213 (2010).

26. Boeka, A. G. & Lokken, K. L. Neuropsychological performance of a clinical sample of extremely obese individuals. *Arch. Clin. Neuropsychol.* **23**, 467–474 (2008).

27. Poskitt, E. M.; European Childhood Obesity Group. Defining childhood obesity: the relative body mass index (BMI). *Acta Paediatr.* **84**, 961–963 (1995).

28. Rolland-Cachera, M. F. *et al.* Body mass index variations: centiles from birth to 87 years. *Eur. J. Clin. Nutr.* **45**, 13–21 (1991).

29. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of *P. J. R. Stat. Soc. A* **85**, 87–94 (1922).

30. Woolf, B. On estimating the relation between blood group and disease. *Ann. Hum. Genet.* **19**, 251–253 (1955).

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

R. G. Walters[1,2]*, S. Jacquemont[3]*, A. Valsesia[4–6], A. J. de Smith[1], D. Martinet[3], J. Andersson[1], M. Falchi[1], F. Chen[7], J. Andrieux[8], S. Lobbens[9], B. Delobel[10], F. Stutzmann[9], J. S. El-Sayed Moustafa[1], J.-C. Chèvre[9], C. Lecoeur[9], V. Vatin[9], S. Bouquillon[8], J. L. Buxton[1], O. Boute[11], M. Holder-Espinasse[11], J.-M. Cuisset[12], M.-P. Lemaitre[12], A.-E. Ambresin[13], A. Brioschi[14], M. Gaillard[3], V. Giusti[15], F. Fellmann[3], A. Ferrarini[3], N. Hadjikhani[7,16], D. Campion[17], A. Guilmatre[17], A. Goldenberg[18], N. Calmels[19], J.-L. Mandel[19], C. Le Caignec[20,21], A. David[20], B. Isidor[20], M.-P. Cordier[22], S. Dupuis-Girod[22], A. Labalme[22], D. Sanlaville[22,23], M. Béri-Dexheimer[24], P. Jonveaux[24], B. Leheup[24,25], K. Ôunap[26], E. G. Bochukova[27], E. Henning[27], J. Keogh[27], R. J. Ellis[28], K. D. MacDermot[28], M. M. van Haelst[28]†, C. Vincent-Delorme[29], G. Plessis[30], R. Touraine[31], A. Philippe[32], V. Malan[32], M. Mathieu-Dramard[33], J. Chiesa[34], B. Blaumeiser[35], R. F. Kooy[35], R. Caiazzo[36,37], M. Pigeyre[37], B. Balkau[38], R. Sladek[39,40], S. Bergmann[4,6], V. Mooser[41], D. Waterworth[41], A. Reymond[42], P. Vollenweider[43], G. Waeber[43], A. Kurg[44], P. Palta[44], T. Esko[45,46], A. Metspalu[45,46], M. Nelis[45,46], P. Elliott[2], A.-L. Hartikainen[47], M. I. McCarthy[48,49], L. Peltonen[50–52], L. Carlsson[53], P. Jacobson[53], L. Sjöström[53], N. Huang[50], M. E. Hurles[50], S. O'Rahilly[27], I. S. Farooqi[27], K. Männik[44], M.-R. Jarvelin[2,54,55], F. Pattou[36,37], D. Meyre[9], A. J. Walley[1], L. J. M. Coin[2], A. I. F. Blakemore[1], P. Froguel[1,9] & J. S. Beckmann[3,4]

*These authors contributed equally to this work.

[1]Section of Genomic Medicine, Imperial College London, London W12 0NN, UK. [2]Department of Epidemiology and Public Health, Imperial College London, London W2 1PG, UK. [3]Service de Génétique Médicale, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland. [4]Departement de Génétique Médicale, Université de Lausanne, CH-1015 Lausanne, Switzerland. [5]Ludwig Institute for Cancer Research, Université de Lausanne, CH-1015 Lausanne, Switzerland. [6]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland. [7]Brain Mind Institute, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. [8]Laboratoire de Génétique Médicale, Centre Hospitalier Régional Universitaire, 59000 Lille, France. [9]CNRS 8090-Institute of Biology, Pasteur Institute, 59800 Lille, France. [10]Centre de Génétique Chromosomique, Hôpital Saint-Vincent de Paul, GHICL, 59020 Lille, France. [11]Service de Génétique Clinique, Hôpital Jeanne de Flandre, Centre Hospitalier Universitaire de Lille, 59000 Lille, France. [12]Service de Neuropédiatrie, Centre Hospitalier Régional Universitaire, 59000 Lille, France. [13]Unité Multidisciplinaire de Santé des Adolescents, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland. [14]Service de Neuropsychologie et de Neuroréhabilitation, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland. [15]Service d'Endocrinologie, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland. [16]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129, USA. [17]INSERM, U614, Faculté de Médecine, 76183 Rouen, France. [18]Service de Génétique, Centre Hospitalier Universitaire de Rouen, 76031 Rouen, France. [19]Laboratoire de Diagnostic Génétique, Nouvel Hôpital civil, 67091 Strasbourg, France. [20]Centre Hospitalier Universitaire Nantes, Service de Génétique Médicale, 44093 Nantes, France. [21]INSERM, UMR915, L'Institut du Thorax, 44007 Nantes, France. [22]Service de Génétique, Hospices Civils de Lyon, Hôpital de l'Hotel Dieu, 69288 Lyon, France. [23]EA 4171, Université Claude Bernard, 69622 Lyon, France. [24]Laboratoire de Génétique, Centre Hospitalier Universitaire, Nancy Université, 54511 Vandoeuvre les Nancy, France. [25]EA4368 Medical School Nancy, Université Henri Poincaré, 54003 Nancy, France. [26]Department of Genetics, United Laboratories, Tartu University Children's Hospital, 50406 Tartu, Estonia. [27]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's

Hospital, Cambridge CB2 0QQ, UK. [28]North West Thames Regional Genetics Service, Northwick Park & St Marks Hospital, Harrow HA1 3UJ, UK. [29]Centre Hospitalier D'Arras, Génétique Médicale, 62000 Arras, France. [30]Service de Génétique Médicale, Centre Hospitalier Universitaire Clemenceau, 14033 Caen, France. [31]Centre Hospitalier Universitaire–Hôpital Nord, Service de Génétique, 42055 Saint Étienne, France. [32]Département de Génétique et INSERM U781, Université Paris Descartes, Hôpital Necker-Enfants Malades, 75015 Paris, France. [33]Service de Génétique Clinique, Centre Hospitalier Universitaire, 80054 Amiens, France. [34]Laboratoire de Cytogénétique, Centre Hospitalier Universitaire Caremeau, 30029 Nîmes, France. [35]Department of Medical Genetics, University Hospital & University of Antwerp, 2650 Edegem, Belgium. [36]INSERM U859, Biotherapies for Diabetes, 59045 Lille, France. [37]Université Lille Nord de France, Centre Hospitalier Universitaire Lille, 59037 Lille, France. [38]INSERM U780-IFR69, 94807 Villejuif, France. [39]McGill University and Genome Quebec Innovation Centre, Montreal H3A 1A4, Canada. [40]Department of Medicine and Human Genetics, McGill University, Montreal H3A 1B1, Canada. [41]Division of Genetics, GlaxoSmithKline, Philadelphia, Pennsylvania 19101, USA. [42]The Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland. [43]Department of Medicine, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland. [44]Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia. [45]Estonian Genome Project, University of Tartu, 50410 Tartu, Estonia. [46]Estonian Biocentre, 51010 Tartu, Estonia. [47]Department of Obstetrics and Gynaecology, University of Oulu, 90220 Oulu, Finland. [48]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK. [49]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [50]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK. [51]Institute of Molecular Medicine, Biomedicum, 00290 Helsinki, Finland. [52]Massachusetts Institute of Technology, The Broad Institute, Cambridge, Massachusetts 02142, USA. [53]Department of Molecular and Clinical Medicine and Center for Cardiovascular and Metabolic Research, The Sahlgrenska Academy, 413 45 Göteborg, Sweden. [54]Department of Child and Adolescent Health, National Public Health Institute, 90101 Oulu, Finland. [55]Institute of Health Sciences and Biocenter Oulu, University of Oulu, 90220 Oulu, Finland. †Present address: Department of Medical Genetics, University Medical Centre Utrecht, 3584 EA Utrecht, The Netherlands.

METHODS

Obesity phenotype. We used previously defined criteria to define overweight, obesity, and morbid (class III) obesity[6,7]. In adults, the thresholds were BMI $\geq$ 25, 30 and 40 kg m^{-2}, respectively. In children and adolescents, we used age-specific and sex-specific centiles of BMI, calculated from a French reference population[27,28], that approximately corresponded to these thresholds: overweight and obesity were defined by thresholds at the 90th and 97th centiles, respectively. Childhood morbid obesity was defined as BMI $\geq$ 4 standard deviations above the age-specific and sex-specific mean, which corresponds to a BMI of 40 kg m^{-2} between the ages of 20 and 30 years for both men and women; this threshold was used in the recruitment of the SCOOP severe early-onset obesity cohorts[7]. The age-specific and sex-specific thresholds used to define obesity and morbid obesity are shown in Fig. 1 and Supplementary Figs 1 and 2. No carriers of a 16p11.2 deletion were reported to be taking atypical antipsychotics (known to be associated with weight gain).

Patient and population cohorts. Patients referred for cognitive delay and obesity: a group of 33 patients was selected from those referred for genetic testing at the North West Thames Regional Genetics Service, based at Northwick Park Hospital in Harrow, UK, with approval from the Harrow Research Ethics Committee. Inclusion was based on three criteria: mental retardation, dysmorphology, and a weight greater than the 97th centile for age and gender. Abnormal karyotype, fragile X and Prader–Willi syndrome had previously been excluded.

A second group of 279 French children were selected from those referred to two centres (Laboratoire de Diagnostic Génétique, Nouvel Hôpital Civil, Strasbourg, France, and Centre de Génétique Chromosomique, Hôpital Saint-Vincent de Paul, GHICL, Lille, France). Inclusion was based on obesity plus at least one Prader–Willi-like syndromic feature (neonatal hypotonia and difficulty to thrive, mental retardation, developmental delay, behavioural problems, skin picking, facial dysmorphism, hypogenitalism or hypogonadism). Chromosomal abnormalities and Prader–Willi syndrome were excluded by karyotyping and DNA methylation analysis.

Patients referred for cognitive delay: patients with cognitive deficits are routinely referred to clinical genetics for aetiological work-ups including aCGH. We surveyed seven cytogenetic centres in France and Switzerland, identifying 3,870 patients ascertained for developmental delay and/or malformations. Also included in the study was a further 77 patients, ascertained on similar criteria, who were referred to the Department of Genetics, University of Tartu, Tartu, Estonia. These analyses were performed for clinical diagnostic purposes, all available phenotypic data (weight and height) being those provided anonymously by the clinician ordering the analysis. Consequently, research-based informed consent was not required by the institutional review board that approved the study.

CoLaus: this prospective population cohort was described previously[16]; 6,188 white individuals aged 35–75 years were randomly selected from the general population in Lausanne, Switzerland. These individuals underwent a detailed phenotypic assessment and were genotyped with the Affymetrix Mapping 500K array; 5,612 samples passed genotyping quality control. This study was approved by the institutional review boards of the University of Lausanne, and written consent was obtained from all participants. Because recruitment of this cohort required the ability to give informed consent, it is possible that the (statistically non-significant) lack of 16p11.2 deletions or duplications is due to an ascertainment bias. However, any such bias, if it exists, is very small and affects the identification of only one or two subjects carrying a deletion.

NFBC1966: the Northern Finland Birth Cohort 1966 is a prospective birth cohort of almost all individuals born in 1966 in the two northernmost provinces of Finland. Expectant mothers were enrolled, and clinical data collection took place prenatally, at birth, and at ages 6 months, 1 year, 14 years and 31 years. Biochemical and DNA samples were collected with informed consent at age 31 years. Genotyping with the Illumina Infinium 370cnvDuo array and phenotypic characteristics of the cohort were as described previously[17]. Phenotypic and genotyping data were available for 5,246 subjects after quality control.

EGPUT: the Estonian Genome Project is a biobank coordinated by the University of Tartu (EGPUT)[18]. The project is conducted in accordance with Estonian Gene Research Act, and all participants gave written informed consent. The cohort includes more than 39,000 individuals older than 18 years of age and reflects closely the age distribution in the Estonian population (33% male, 67% female; 83% Estonians, 14% Russians, 3% other). Subjects are recruited by general practitioners and hospital physicians and are then randomly selected. Computer Assisted Personal interview (CAPI) was filled during 1–2 h at the doctor's office. The data included personal data (such as place of birth, place(s) of living and nationality), family history (four generations), educational and occupational history, lifestyle and anthropometric data. A total of 1,090 randomly selected subjects were genotyped with the Illumina 370cnvDuo array, 998

passing the required criteria (nationality, genotyping call rate and phenotype availability).

Case-control familial obesity: the adult-obesity case-control groups and the child-obesity case control groups were as published previously[6], and were genotyped with the Illumina Human CNV370-duo array. In all, 643 children with familial obesity (BMI $\geq$ 97th centile corrected for gender and age, at least one obese first-degree relative, age less than 18 years), 581 non-obese children (BMI $\leq$ 90th centile), 705 morbidly obese adults with familial obesity (BMI $\geq$ 40 kg m^{-2}, at least one obese first-degree relative with BMI $\geq$ 35 kg m^{-2}, age $\geq$ 18 years) and 197 lean adults (BMI $\leq$ 25 kg m^{-2}) passed quality control; this cohort included a further 646 control subjects from the DESIR prospective cohort[19] (age at examination $\geq$ 45 years, normal fasting glucose in accordance with 1997 ADA criteria, BMI < 27 kg m^{-2}) genotyped with the Illumina Hap300 array[20]. All participants or their legal guardians gave written informed consent, and all local ethics committees approved the study protocol.

Severe early-onset obesity cohort: the Genetics of Obesity Study (GOOS) cohort consists of more than 3,000 patients ascertained for severe obesity, defined as a BMI $\geq$ 4 standard deviations above the age-specific and sex-specific mean, and onset of obesity before 10 years of age. In this study we selected a discovery set of 1,000 UK Caucasian patients from this cohort in whom developmental delay had been excluded by routine clinical examination by experienced physicians (this cohort is referred to as SCOOP). Mutations in *LEPR*, *POMC* and *MC4R* were excluded by direct nucleotide sequencing and a karyotype was performed. DNA samples were analysed with Affymetrix Genome-Wide Human SNP Array 6.0 by Aros, of which 931 passed quality control.

Bariatric surgery cohort: patients undergoing elective bariatric weight-loss surgery were recruited for the ABOS study at Lille Regional University Hospital. Genotyping was performed with the Illumina Human 1M-duo array, and data from 141 adults passed quality control. All participants gave written informed consent, and the study protocol was approved by the local ethics committee.

Swedish discordant sibling cohort: the SOS Sib Pair Study cohort was as published previously[21]. It includes 154 nuclear families, each with BMI discordant sibling pairs (BMI difference > 10 kg m^{-2}), giving a total of 732 subjects. Genotyping data with the Illumina 610K-Quad array was available for 353 siblings from 149 families. Expression data from subcutaneous adipose tissue (sampled after overnight fasting) were available for 360 siblings from 151 families. Subjects received written and oral information before giving written informed consent. The Regional Ethics Committee in Gothenburg approved the studies.

Statistical methods. In view of the low frequency of the 16p11.2 deletions, all reported statistical tests were conducted with Fisher's exact test[29]. This was applied to comparisons of separately ascertained cohorts or categories and was performed on contingency tables constructed for the number of subjects carrying or lacking a 16p11.2 deletion (zero or one copies, because no homozygous deletions were observed) versus the obesity status or ascertainment of the individual. Because no homozygous deletions were observed, it was unnecessary to make a prior distinction between recessive, additive and dominant models of disease risk. For overall analysis of the obesity risk resulting from a deletion, cohorts were pooled in accordance with their obesity status determined according to the criteria described above, and the described tests were then applied to the pooled data. Odds ratios and 95% confidence limits were calculated as described[30].

CNV discovery and validation. Clinical identification of 16p11.2 deletions: all diagnostic procedures (aCGH, qPCR, QMPSF and FISH) were conducted in accordance with the relevant guidelines of good clinical laboratory practice for the respective countries. All rearrangements in probands were confirmed by a second technique, and karyotyping was performed in all cases to exclude a complex rearrangement.

cnvHap: CNVs were detected in the child/adult case-control, bariatric surgery, SOS sibpair and NFBC cohorts using the cnvHap algorithm (L.J.M.C., J. E. Asher, R.G.W., J.S.E.-S.M., A.J.d.S., R.S., D. J. Balding, P.F. and A.I.F.B., unpublished observations); this method is based on a hidden Markov model that models transitions between copy-number states at the haplotype level, improving sensitivity and accuracy by capturing linkage disequilibrium information between CNVs and SNPs. The compiled JAR and associated parameter files can be downloaded from http://www.imperial.ac.uk/medicine/people/l.coin/. Sample data from the algorithm applied to the NFBC cohort are illustrated in Supplementary Fig. 5a.

After clustering of genotyping data with the internal Illumina BeadStudio cluster files, values for log$R$ ratio (LRR) and B-allele frequency (BAF) were exported from each project and normalized: effects of percentage GC content on LRR were removed by regressing on GC and GC^2, and wave effects[31] were removed by fitting a Loess function. Normalized data for probes within

2.5 megabases of the 16p11.2 deletion were analysed with cnvHap, and CNV calls intersecting the single-copy sequences within the deletion (chr16:29514353–30107356, build hg18) were extracted. 16p11.2 deletions were identified by a minimum 90% of probes within the deleted region being called as having a decreased copy number.

All called 16p11.2 deletions were validated by direct analysis of LRR. Data for each probe were normalized by first subtracting the median value across all samples (so that the distribution of LRR for each probes was centred on zero), and then dividing by the variance across all samples (to correct for variation in the sensitivity of different probes to copy-number variation). The normalized data were then smoothed by application of a nine-point moving average and visualized graphically (see Supplementary Fig. 6); putative deletions were checked by subsequent manual confirmation of loss of heterozygosity across the entire region. Equally, all deletions called by this method were confirmed by cnvHap.

Gaussian mixture model: for the CoLaus cohort, raw genotyping data were normalized using the aroma.affymetrix framework[32]. Normalization steps included allelic cross-talk calibration[33,34], intensity summarization using robust median average, and correction for any PCR amplification bias. Copy number (CN) ratios for a given sample, at a given SNP or CN probe, were computed as the $\log_2$ ratio of the normalized intensity of this probe divided by the median across all the samples. CN ratios were subsequently smoothed by fitting a Loess function[31]. CNV calling was performed with a new method based on a Gaussian mixture model (A.V., Z. Kutalik, T. Johnson, B. J. Stevenson, C. V. Jongeneel, D.W., V.M., P.V., G.W., J.S.B. and S.B., unpublished observations). This Gaussian mixture model fits four components (deletion, copy neutral, one additional copy and two additional copies) to CN ratios. The final copy number at each probe location is determined as the expected (dosage) copy number. The method has been validated by comparing test data sets with results from the CNAT[35] and CBS[36,37] algorithms and by replicating a subset of CoLaus subjects on Illumina arrays. All calls at the 16p11.2 locus made by the highly stringent CBS algorithm were replicated by the Gaussian mixture model. Principal components analysis detected no significant batch effects. Sample data from the algorithm applied to the CoLaus cohort are illustrated in Supplementary Fig. 5b.

PennCNV, QuantiSNP and Birdsuite: CNV discovery in the EGPUT cohort was performed with QuantiSNP[38], PennCNV[39] and BeadStudio GT module (Illumina). All analyses were conducted with the recommended settings, except changing EMiters to 25 and L to 1,000,000 in QuantiSNP. For PennCNV, the Estonian population-specific BAF file was used. Data from the SCOOP cohort were analysed with Affymetrix Power Tools and Birdsuite software[40].

Multiplex ligation-dependent probe amplification (MLPA): MLPA was performed with standard methods[41] using reagents obtained from MRC-Holland. The SALSA MLPA kit P343-B1 Autism-1 probe mix was used, which contained nine probes within the deleted region on 16p11.2, plus one probe upstream and one downstream of this locus (see Fig. 1a). MLPA products were separated with an AB3130 Genetic Analyser (Applied Biosystems) and outputs were analysed with GeneMarker software (Soft Genetics) and Microsoft Excel. Data normalization was performed by dividing the peak areas for each of the 11 test probes by the mean of 9 control probe peak areas. Normalized peak area data were then compared across the tested samples to determine which of them carried the 16p11.2 deletion.

31. Marioni, J. C. *et al.* Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.* **8**, R228, doi:10.1186/gb-2007-8-10-r228 (2007).
32. Bengtsson, H., Simpson, K., Bullard, J. & Hansen, K. aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. (Department of Statistics, University of California, Berkeley, Technical Report 745, 2008).
33. Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T. P. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767 (2008).
34. Bengtsson, H., Ray, A., Spellman, P. & Speed, T. P. A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics* **25**, 861–867 (2009).
35. Huang, J. *et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1**, 287–299 (2004).
36. Olshen, A. B. & Venkatraman, E. S. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
37. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
38. Collela, S. *et al.* QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).
39. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
40. Korn, J. N. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genet.* **40**, 1253–1260 (2008).
41. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).

LETTERS

TGF-β–FOXO signalling maintains leukaemia-initiating cells in chronic myeloid leukaemia

Kazuhito Naka[1]*, Takayuki Hoshii[1]*, Teruyuki Muraguchi[1], Yuko Tadokoro[1], Takako Ooshio[1,2], Yukio Kondo[3], Shinji Nakao[3], Noboru Motoyama[4] & Atsushi Hirao[1,2]

Chronic myeloid leukaemia (CML) is caused by a defined genetic abnormality that generates *BCR-ABL*, a constitutively active tyrosine kinase[1]. It is widely believed that BCR-ABL activates Akt signalling that suppresses the forkhead O transcription factors (FOXO), supporting the proliferation or inhibiting the apoptosis of CML cells[2–4]. Although the use of the tyrosine kinase inhibitor imatinib is a breakthrough for CML therapy, imatinib does not deplete the leukaemia-initiating cells (LICs) that drive the recurrence of CML[5–8]. Here, using a syngeneic transplantation system and a CML-like myeloproliferative disease mouse model, we show that Foxo3a has an essential role in the maintenance of CML LICs. We find that cells with nuclear localization of Foxo3a and decreased Akt phosphorylation are enriched in the LIC population. Serial transplantation of LICs generated from *Foxo3a*[+/+] and *Foxo3a*[−/−] mice shows that the ability of LICs to cause disease is significantly decreased by Foxo3a deficiency. Furthermore, we find that TGF-β is a critical regulator of Akt activation in LICs and controls Foxo3a localization. A combination of TGF-β inhibition, Foxo3a deficiency and imatinib treatment led to efficient depletion of CML *in vivo*. Furthermore, the treatment of human CML LICs with a TGF-β inhibitor impaired their colony-forming ability *in vitro*. Our results demonstrate a critical role for the TGF-β–FOXO pathway in the maintenance of LICs, and strengthen our understanding of the mechanisms that specifically maintain CML LICs *in vivo*.

Although tyrosine kinase inhibitor (TKI) therapy of CML patients efficiently induces the death of leukaemia cells[5–8], LICs in these patients can survive this therapy. To understand the molecular mechanisms maintaining CML LICs, we characterized LICs *in vivo* using a mouse model for CML-like myeloproliferative disease (MPD)[9]. Consistent with previous reports[10–13], we found that a rare c-Kit+Lineage−(Lin−)Sca-1+ (KLS+) population of CML cells (that is, bearing markers of normal haematopoietic stem cells (HSCs)) induced efficient CML development in recipient mice (Supplementary Fig. 1). In contrast, neither c-Kit+Lin−Sca-1− (KLS−) cells (which correspond to normal progenitors), nor other CML cell populations expressing differentiation markers, induced CML.

We and others have shown that Foxo transcription factors, which are important downstream targets of PI3K-Akt signalling, are essential for the maintenance of self-renewal capacity in normal HSCs[14–16]. When a growth factor binds to the appropriate receptor, Akt is activated and phosphorylates Foxo proteins, resulting in their nuclear export and subsequent degradation in the cytoplasm. In the absence of growth factor stimulation, Foxo proteins are retained in an active state in the nucleus and induce their transcriptional targets. In CML cell lines, BCR-ABL is thought to activate PI3K-Akt signalling that leads to nuclear

export of Foxo factors and suppression of their transcriptional activity[2–4]. However, we found that, whereas most KLS− cells (non-LICs) showed the expected cytoplasmic localization of Foxo3a, KLS+ cells (LICs) enriched cells with nuclear localization of Foxo3a (Fig. 1a and Supplementary Fig. 2), as observed in normal HSCs[15,17] (Supplementary Fig. 3). LICs with nuclear Foxo3a also exhibited decreased levels of phosphorylated Akt (p-Akt) compared to most non-LICs, suggesting that Foxo3a remains active in LICs owing to Akt suppression.

We next examined the role of Foxo3a in CML initiation by establishing *Foxo3a*-deficient CML model. We isolated immature bone marrow cells from *Foxo3a*[+/+] and *Foxo3a*[−/−] littermates[15], infected these cells with retrovirus carrying the *BCR-ABL* gene, and transplanted them into syngeneic recipients (first bone marrow transplantation (BMT)). Both recipient groups showed the same symptoms of CML-like MPD, including increased myeloid cells in peripheral blood and splenomegaly[9] (Fig. 1b, c, left and Supplementary Figs 4 and 5). Thus, Foxo3a was dispensable for the generation of CML-like disease. After a first BMT, the absolute numbers of *Foxo3a*[−/−] LICs in recipient spleen and bone marrow were significantly higher than the absolute numbers of *Foxo3a*[+/+] LICs present in recipient organs (Fig. 1d, left and Supplementary Fig. 6a), although the *in vitro* colony-forming ability of these *Foxo3a*[+/+] and *Foxo3a*[−/−] LICs was comparable (Fig. 1e, left).

When we transplanted *Foxo3a*[+/+] and *Foxo3a*[−/−] LICs from first BMT mice into a new set of recipients (second BMT), there was no difference in CML development (Fig. 1b, c, centre and Supplementary Figs 4 and 5). However, the *in vitro* colony-forming ability of second BMT LICs was decreased by loss of *Foxo3a* (Fig. 1e, right). In a third BMT, relatively mild CML-like MPD developed in recipients within one month. Foxo3a deficiency prevented the propagation of CML cells in the peripheral blood and spleen *in vivo* (Fig. 1b, right and Supplementary Figs 4 and 5). Furthermore, the absolute numbers of *Foxo3a*[−/−] LICs in the organs of third BMT recipients were much lower than in third BMT recipients that had received *Foxo3a*[+/+] LICs (Fig. 1d, right and Supplementary Fig. 6). Thus, *Foxo3a*[−/−] LICs may retain sufficient function to cause disease in second BMT recipients but succumb to exhaustion in third BMT recipients. Although CML-like MPD developed in mice that had received either *Foxo3a*[+/+] or *Foxo3a*[−/−] LICs within 40 days of a third BMT (Supplementary Fig. 5a, b), recipients of *Foxo3a*[+/+] LICs also developed acute lymphocytic leukaemia (ALL), as well as CML, 40 days after the third BMT (Supplementary Fig. 5c) as reported[18], suggesting that these LICs can generate malignancies in several lineages. Notably, we did not observe development of ALL or CML in recipients of *Foxo3a*[−/−] LICs 45 days after a third BMT (Supplementary Fig. 5c), suggesting that the *Foxo3a*[−/−] LICs lose their potential to generate malignancies. This
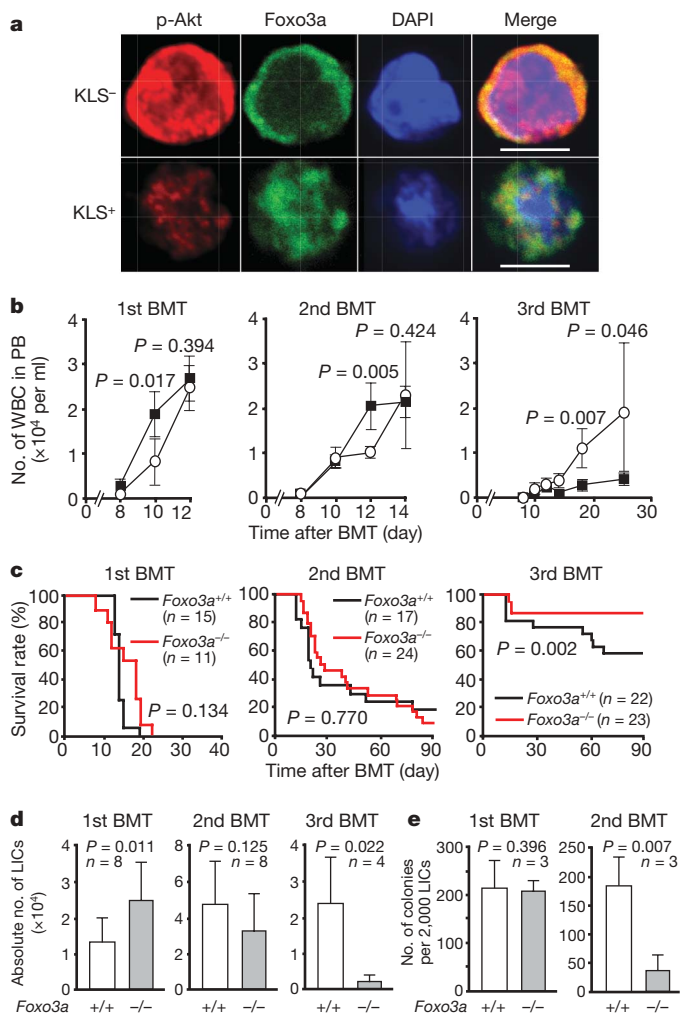
Figure 1 | Maintenance of CML LICs depends on Foxo3a. a, Decreased Akt phosphorylation and increased Foxo3a nuclear localization in CML LICs. BCR-ABL-GFP$^+$ KLS$^-$ and KLS$^+$ cells were immunostained to detect p-Akt and Foxo3a. Nuclei were visualized using 4′,6-diamidino-2-phenylindole (DAPI; blue). Scale bars, 10 μm. **b,** Numbers of white blood cells (WBC) in peripheral blood (PB) in Foxo3a$^{+/+}$ (open circles) and Foxo3a$^{-/-}$ (closed squares) CML-affected mice at first to third BMT. Data shown are the mean WBC number ± s.d. ($n = 4$). **c, d,** Survival rate (**c**) and absolute number of LICs (**d**) in CML-affected mice. Data are expressed as percentage survival (**c**) and absolute number of LICs ± s.d. in spleen (**d**). **e,** In vitro colony-forming ability of LICs. Data shown are the mean colony number ± s.d.

Figure 2 | Essential role of Foxo3a in suppression of LIC apoptosis. a, b, Comparable differentiation potential and cell cycle status in vivo. **a,** Data show the mean percentage ± s.d. of CMP (common myeloid progenitors)-, GMP (granulocyte-macrophage progenitors)-, and MEP (megakaryocyte-erythroid progenitors)-like cells in BCR-ABL-GFP$^+$ bone marrow cells ($n = 3$). **b,** Data are the mean frequency of BrdU$^+$ cells ± s.d. ($n = 3$). **c–f,** Increased apoptosis of Foxo3a$^{-/-}$ LICs. **c, d,** Histological sections of bone marrow (**c**) and spleen (**d**) from Foxo3a$^{+/+}$ or Foxo3a$^{-/-}$ CML-affected mice at second BMT were analysed by TUNEL staining (red) and DAPI. Scale bars, 100 μm. **e,** Mean percentage ± s.d. of Annexin-V$^+$ cells among GFP$^+$KLS$^+$ cells ($n = 3$). **f,** LICs were subjected to TUNEL staining. Scale bars, 50 μm.

compared to controls (Fig. 2c, d). To confirm that this apoptosis was occurring in LICs, we purified KLS$^+$ cells and showed that the frequency of Annexin-V$^+$ or TUNEL$^+$ cells among Foxo3a$^{-/-}$ KLS$^+$ cells was higher than among Foxo3a$^{+/+}$ KLS$^+$ cells (Fig. 2e, f). Thus, Foxo3a is required for LIC survival because it mediates suppression of apoptosis.

An intriguing question in current CML research is how Akt is inactivated (and thus Foxo is activated) mainly in LICs, despite BCR-ABL expression in all CML cells. The fact that TGF-β regulates Akt activation (and thus nuclear Foxo3a) in normal HSCs prompted us to assess whether TGF-β signalling controls Foxo localization in CML LICs[19]. We first examined the phosphorylation of Smad2/3 proteins, which are downstream effectors in the TGF-β signalling pathway[20]. Phosphorylation of Smad2/3 in the nuclei of CML KLS$^+$ cells was higher than that in KLS$^-$ cells (Fig. 3a and Supplementary Fig. 10), indicating that TGF-β signalling was activated in CML LICs. We then treated CML KLS$^+$ cells in vitro for 2 h with either TGF-β1 or the TGF-β inhibitor Ly364947 (ref. 21) and analysed Foxo3a localization. TGF-β1 treatment increased Foxo3a nuclear localization, whereas Ly364947 promoted nuclear export of Foxo3a (Fig. 3b–d and Supplementary Fig. 11). Furthermore, KLS$^+$ cells treated with TGF-β1 showed high Smad2/3 and low Akt phosphorylation levels, whereas cells treated with Ly364947 exhibited the opposite pattern (Fig. 3b, c and Supplementary Figs 11 and 12). The activity of mTOR complex 1 (mTORC1), which was determined by assessing the phosphorylation of S6 ribosomal protein[22], correlated positively with p-Akt levels in these KLS$^+$ cells (Fig. 3d and Supplementary Fig. 13). We propose that TGF-β signalling controls Akt activity in LICs, leading to retention of Foxo3a in the nucleus.

inability of Foxo3a$^{-/-}$ LICs to induce or sustain leukaemia resulted in the reduced lethality of these animals (Fig. 1c, right). Thus, Foxo3a is essential for the long-term maintenance of leukaemia-initiating potential. Because the introduction of a dominant-negative Foxo moiety into CML LICs impaired LIC function in vivo (Supplementary Fig. 7), the defective maintenance of Foxo3a$^{-/-}$ LICs is probably due to a diminution in LIC self-renewal activity rather than a defect in the original HSCs. Furthermore, the Foxo4 protein expression pattern in LICs is very similar to that of Foxo3a (Supplementary Fig. 8), suggesting an overlap in Foxo functions.

We next determined how Foxo3a deficiency impaired the maintenance of the CML-initiating potential of LICs. Loss of Foxo3a did not affect LIC differentiation potential (Fig. 2a) or cell cycle status (Fig. 2b), although LICs with nuclear localization of Foxo3a showed lower expression of the Ki67 antigen (Supplementary Fig. 9). However, TdT-mediated dUTP nick end labelling (TUNEL)-positive apoptotic cells were significantly increased in histological sections of bone marrow and spleen from Foxo3a$^{-/-}$ CML-affected mice
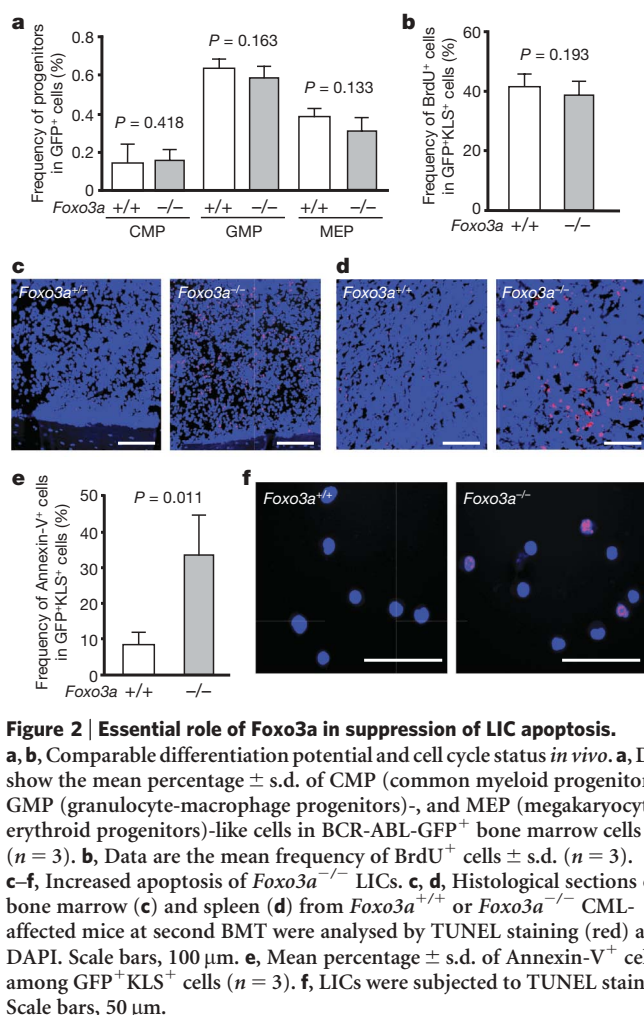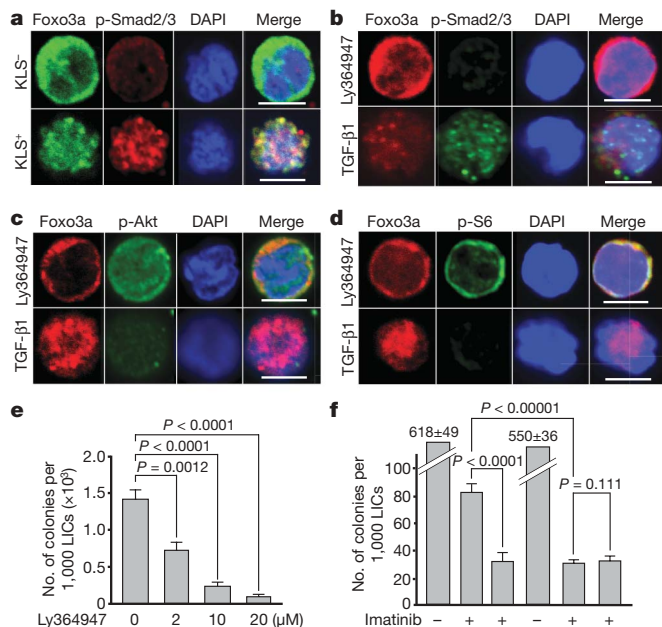
Figure 3 | TGF-β–Foxo signalling is required for the colony-forming capacity of LICs. a, Increased Smad2/3 phosphorylation in LICs. *Foxo3a*[+/+] KLS[+] and KLS[−] CML cells were immunostained to detect Foxo3a and phosphorylated-Smad2/3 (p-Smad2/3). **b–d,** Control of Foxo3a localization by TGF-β signalling. LICs were treated *in vitro* with TGF-β1 or Ly364947 for 2 h followed by immunostaining to detect Foxo3a and (**b**) p-Smad2/3, (**c**) p-Akt or (**d**) phosphorylated-S6 (p-S6). Scale bars, 10 μm. **e,** Inhibition of TGF-β decreases LIC colony-forming ability. Representative data shown are the mean LIC colony number ± s.d. (*n* = 3). **f,** TGF-β inhibition or Foxo3a deficiency enhances imatinib-mediated inhibition of LIC colony-forming capacity. Representative data shown are the mean colony number ± s.d. (*n* = 3).

Notably, as TGF-β did not efficiently induce nuclear localization of Foxo3a in non-LICs (KLS[−]) cells (Supplementary Figs 14–16), there seem to be differences in how TGF-β-stimulated Foxo3a activation is regulated in LICs and non-LICs.

These findings led us to evaluate the effects of inhibition of TGF-β–Foxo signalling on the CML-initiating potential of LICs. Interestingly, TGF-β inhibitors (Ly364947, SD208) suppressed LIC colony-forming ability after co-culture with OP-9 stromal cells, which are used to maintain normal HSCs *ex vivo* and mimic their *in vivo* environment[23] (Fig. 3e and Supplementary Fig. 17a). Furthermore, although imatinib treatment alone of LICs co-cultured with OP-9 cells reduced the colony-forming capacity of *Foxo3a*[+/+] CML LICs, the addition of TGF-β inhibitor significantly enhanced the inhibitory effect of imatinib on LIC colony-forming ability (Fig. 3f and Supplementary Figs 17b, c and 18a), consistent with a report in which TGF-β inhibition enhanced imatinib-induced cell death[24]. Although Foxo3a deficiency alone did not affect LIC colony-forming ability (Fig. 1e, left), significantly fewer colonies arose from *Foxo3a*[−/−] LICs treated with imatinib compared to treated *Foxo3a*[+/+] LICs (Fig. 3f). When imatinib was combined with Foxo3a deficiency and Ly364947, there was no further inhibitory effect compared to either Foxo3a deficiency or Ly364947 treatment alone (Fig. 3f). TGF-β inhibition had no effect on the colony-forming ability of normal KLS[+] cells under the same experimental conditions (Supplementary Fig. 18b). These data indicate that Foxo3a is an important downstream effector in the TGF-β signalling pathway driving the survival of LICs exposed to TKI therapy, and that inhibition of Foxo or TGF-β may be a highly useful adjunct to TKI therapy.

To test this hypothesis *in vivo*, we administered imatinib to second BMT mice bearing *Foxo3*[+/+] or *Foxo3a*[−/−] CML LICs. Without

imatinib, 80% of both groups died within 60 days (Fig. 4a). Administration of imatinib alone delayed CML onset and reduced recipient lethality, but 60% of the treated mice that had received *Foxo3a*[+/+] LICs died within 90 days. In contrast, imatinib combined with Foxo3a deficiency significantly increased the survival of CML-affected mice. Moreover, when we isolated the residual KLS[+] populations from these imatinib-treated recipients, we found that Foxo3a deficiency had reduced both the number and colony-forming ability of these LICs (Supplementary Fig. 19). Thus, Foxo3a is essential for the ability of CML LICs to survive imatinib therapy.

We then investigated the *in vivo* effect of a TGF-β inhibitor on LIC functions. Administration of Ly364947 alone led to increased p-Akt levels and decreased nuclear Foxo3a in LICs (Fig. 4b and Supplementary Fig. 20), demonstrating that TGF-β is a critical regulator of Akt and Foxo in LICs *in vivo*. Unlike the suppressive effect exerted by TGF-β *in vitro*, administration of Ly364947 alone did not extend the survival of second BMT *Foxo3a*[+/+] mice with overt CML. However, Ly364947 combined with imatinib significantly reduced recipient lethality, decreased CML infiltration in lung (Fig. 4c and Supplementary Fig. 21a), and decreased LIC frequency (Supplementary Fig. 21b, c). These *in vivo* data suggest that CML LICs are more sensitive than normal HSCs/progenitors to TGF-β inhibitors.

Finally, we further confirmed that the suppressive effects of TGF-β inhibitors on mouse CML LIC function are also observed in human CML LICs. We isolated CD34[+]CD38[−]Lin[−] cells as human CML LICs[12] from bone marrow cells of human CML patients, and co-cultured them on OP-9 cells with or without Ly364947. Ly364947 reduced the number of colonies formed by human CML LICs (Fig. 4d) and enhanced the inhibitory effects of imatinib on human CML LICs (Fig. 4e), suggesting that TGF-β–FOXO signalling also governs the behaviour of human CML LICs.

Our study indicates a model in which Foxo3a has opposite effects on the survival of LICs and non-LICs (Fig. 4f and Supplementary Discussion 1). In non-LICs, BCR-ABL drives strong Akt activation that forcefully represses Foxo3a functions. When TKIs block BCR-ABL and reduce Akt activation, activation of Foxo3a leads to apoptosis or cell cycle arrest. In contrast, in LICs, Akt activity is suppressed despite BCR-ABL expression *in vivo*, leading to nuclear localization of Foxo3a. Although a previous study suggested that Foxo3a may contribute to the acquisition of dormancy in leukaemia cells after exposure to anti-leukaemic agents[2], our data demonstrate that LICs have properties that allow them to resist various stress *in vivo* in a Foxo-dependent manner. Notably, we found that the role of Foxo3a in a CML blast crisis model was different from that in this MPD model (Supplementary Figs 22 and 23), indicating that Foxo proteins can mediate different regulatory mechanisms in different phases and types of leukaemia.

It remains unclear precisely how Akt and Foxo signalling is controlled in LICs *in vivo*. The presence of TGF-β1 in CML LICs (Supplementary Fig. 24) and the *in vitro* effects of TGF-β inhibitor on Foxo3a localization in LICs suggest that a cell-autonomous effect of TGF-β may control Foxo3a function *in vivo*. However, when we cultured LICs with TGF-β inhibitors in a stroma-free system, colony formation was not inhibited (Supplementary Fig. 25), suggesting that the survival and/or proliferation of LICs depends not only on TGF-β produced by LICs themselves but also on TGF-β in the surrounding microenvironment. Alternatively, TGF-β signalling in niche *in vivo* may affect LIC behaviour in a manner mediated by unknown factors. In any case, it appears that LICs are controlled by a complicated network of TGF-β signalling.

Although our study suggests that inhibition of the TGF-β–FOXO axis might represent a new therapeutic approach for CML patients, the application of TGF-β or FOXO inhibitors to the treatment of haematological malignancies warrants careful consideration (Supplementary Discussion 2). Our future work will be focused on further understanding how the LICs are maintained *in vivo*.
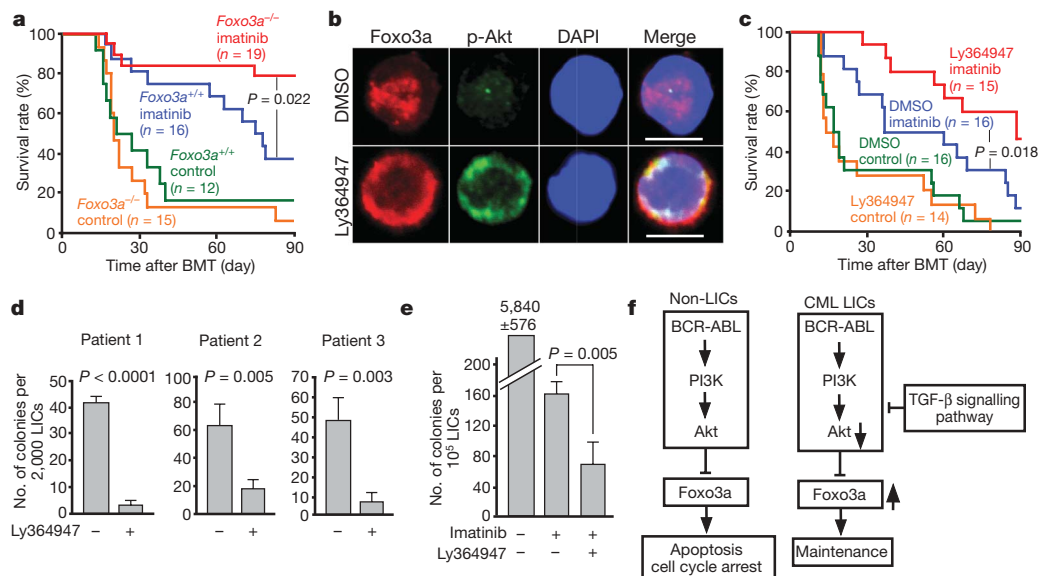
Figure 4 | Inhibition of TGF-β–Foxo3a signalling in combination with TKI therapy depletes CML in vivo. a, Foxo3a deficiency promotes the survival of imatinib-treated CML mice in a second BMT. At 10 days after BMT, mice transplanted with $Foxo3a^{+/+}$ or $Foxo3a^{-/-}$ LICs (1.5×10^4) received either vehicle (control) or imatinib for 80 days, and the percentage survival was determined. **b**, Ly364947 administration to CML-affected mice (25 mg kg^{-1} for 7 days) increases p-Akt and decreases nuclear Foxo3a in LICs. Scale bars, 10 μm. **c**, Combined administration of Ly364947 and imatinib prolongs the survival of CML-affected mice. At 10 days after BMT, recipients received

dimethyl sulphoxide (DMSO) or Ly364947 plus vehicle (control) or imatinib for 80 days, and the percentage survival was determined. **d**, **e**, Inhibition of TGF-β decreases the in vitro colony-forming ability of human CML LICs. CD34$^+$CD38$^-$Lin$^-$ cells from bone marrow of CML patients were co-cultured on OP-9 cells with either DMSO ($-$) or Ly364947 ($+$) for 5 days in the absence (**d**) or presence (**e**) of imatinib. Data shown are the mean colony number ± s.d. ($n = 3$). **f**, Scheme outlining the proposed distinct roles of Foxo3a in non-LICs versus CML LICs.

METHODS SUMMARY

For our mouse CML model, normal immature bone marrow cells (KLS$^+$) from C57BL/6 mice were infected with retrovirus carrying MSCV-BCR-ABL-ires-GFP[10,12,13]. The transduced KLS$^+$ cells were transplanted intravenously into lethally irradiated (9.5 Gy) C57BL/6 congenic mice along with 5×10^5 bone marrow mononuclear cells from C57BL/6 mice. The development of CML-like MPD was confirmed by morphological analysis and marker determinations. To analyse phenotypes of LICs, the GFP$^+$KLS$^+$ subpopulation from bone marrow and spleen was purified and transplanted into syngeneic recipients along with 5×10^5 bone marrow mononuclear cells. To examine Foxo3a in the maintenance of CML LICs, $Foxo3a^{-/-}$ and $Foxo3a^{+/+}$ littermates (C57BL/6; F4) were subjected to the above protocol. For Akt activity, TGF-β signalling and Foxo3a localization, freshly isolated GFP$^+$KLS$^+$ and KLS$^-$ subpopulations, or GFP$^+$KLS$^+$ cells incubated for 2 h with 5 ng ml^{-1} TGF-β1 (R&D Systems) or 10 μM Ly364947 (Merck), were immunostained with anti-p-Akt (Cell Signaling), anti-phospho-S6 (Cell Signaling), anti-phospho-Smad2/3 (Millipore), or anti-Foxo3a (Sigma). To examine colony-forming ability in vitro, GFP$^+$KLS$^+$ cells were cultured in semi-solid methylcellulose medium (stroma-free)[15], or were co-cultured for 5 days on OP-9 stromal cells with Ly364947 (10 μM). For experiments involving imatinib, GFP$^+$KLS$^+$ cells were cultured on OP-9 cells with or without 5 μM imatinib (kind gift from Novartis), followed by washing and transfer to semi-solid medium. Human LICs were purified as CD34$^+$CD38$^-$Lin$^-$ cells of bone marrow cells from CML patients (Cureline, Inc. and AllCells). These cells were co-cultured on OP-9 cells for 5 days before transfer to semi-solid medium. Cell cycling was assessed by in vivo BrdU incorporation (12 h). Apoptosis was assayed by TUNEL (Roche) or Annexin-V (Abcam) staining. Recipient mice bearing $Foxo3a^{+/+}$ or $Foxo3a^{-/-}$ CML LICs received either imatinib by oral gavage twice a day for 80 days at 200 mg kg^{-1} (of body weight) per day, or Ly364947 by intraperitoneal administration for 80 days at 10 mg kg^{-1} (of body weight) every 2 days.

Note added in proof: During the reviewing process, a paper demonstrating that human CML stem cells show predominant nuclear localization of FOXO proteins was published[25]. The study supports our conclusion about the important roles of FOXO in CML LICs.

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

1. Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nature Rev. Cancer* **5**, 172–183 (2005).

2. Komatsu, N. et al. A member of Forkhead transcription factor FKHRL1 is a downstream effector of STI571-induced cell cycle arrest in BCR-ABL-expressing cells. *J. Biol. Chem.* **278**, 6411–6419 (2003).

3. Ghaffari, S., Jagani, Z., Kitidis, C., Lodish, H. F. & Khosravi-Far, R. Cytokines and BCR-ABL mediate suppression of TRAIL-induced apoptosis through inhibition of forkhead FOXO3a transcription factor. *Proc. Natl Acad. Sci. USA* **100**, 6523–6528 (2003).

4. Essafi, A. et al. Direct transcriptional regulation of Bim by FoxO3a mediates STI571-induced apoptosis in Bcr-Abl-expressing cells. *Oncogene* **24**, 2317–2329 (2005).

5. Graham, S. M. et al. Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 in vitro. *Blood* **99**, 319–325 (2002).

6. Michor, F. et al. Dynamics of chronic myeloid leukaemia. *Nature* **435**, 1267–1270 (2005).

7. Roeder, I. et al. Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nature Med.* **12**, 1181–1184 (2006).

8. Jørgensen, H. G., Allan, E. K., Jordanides, N. E., Mountford, J. C. & Holyoake, T. L. Nilotinib exerts equipotent antiproliferative effects to imatinib and does not induce apoptosis in CD34$^+$ CML cells. *Blood* **109**, 4016–4019 (2007).

9. Daley, G. Q., Van Etten, R. A. & Baltimore, D. Induction of chronic myelogenous leukemia in mice by the P210$^{bcr/abl}$ gene of the Philadelphia chromosome. *Science* **247**, 824–830 (1990).

10. Hu, Y. et al. Targeting multiple kinase pathways in leukemic progenitors and stem cells is essential for improved treatment of Ph$^+$ leukemia in mice. *Proc. Natl Acad. Sci. USA* **103**, 16870–16875 (2006).

11. Neering, S. J. et al. Leukemia stem cells in a genetically defined murine model of blast-crisis CML. *Blood* **110**, 2578–2585 (2007).

12. Ito, K. et al. PML targeting eradicates quiescent leukaemia-initiating cells. *Nature* **453**, 1072–1078 (2008).

13. Zhao, C. et al. Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukemia. *Nature* **458**, 776–779 (2009).

14. Tothova, Z. et al. FoxOs are critical mediators of hematopoietic stem cell resistance to physiologic oxidative stress. *Cell* **128**, 325–339 (2007).

15. Miyamoto, K. et al. Foxo3a is essential for maintenance of the hematopoietic stem cell pool. *Cell Stem Cell* **1**, 101–112 (2007).

16. Yalcin, S. et al. Foxo3 is essential for the regulation of ataxia telangiectasia mutated and oxidative stress-mediated homeostasis of hematopoietic stem cells. *J. Biol. Chem.* **283**, 25692–25705 (2008).

17. Yamazaki, S. et al. Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells. *EMBO J.* **25**, 3515–3523 (2006).

18. Pear, W. S. et al. Efficient and rapid induction of a chronic myelogenous leukemia-like myeloproliferative disease in mice receiving P210 bcr/abl-transduced bone marrow. *Blood* **92**, 3780–3792 (1998).

19. Yamazaki, S. *et al.* TGF-β as a candidate bone marrow niche signal to induce hematopoietic stem cell hibernation. *Blood* **113**, 1250–1256 (2009).
20. Massagué, J. TGFβ in Cancer. *Cell* **134**, 215–230 (2008).
21. Sawyer, J. S. *et al.* Synthesis and activity of new aryl- and heteroaryl-substituted pyrazole inhibitors of the transforming growth factor-β type I receptor kinase domain. *J. Med. Chem.* **46**, 3953–3956 (2003).
22. Bhaskar, P. T. & Hay, N. The two TORCs and Akt. *Dev. Cell* **12**, 487–502 (2007).
23. Kodama, H., Nose, M., Niida, S. & Nishikawa, S. Involvement of the c-kit receptor in the adhesion of hematopoietic stem cells to stromal cells. *Exp. Hematol.* **22**, 979–984 (1994).
24. Møller, G. M., Frost, V., Melo, J. V. & Chantry, A. Upregulation of the TGFβ signalling pathway by Bcr-Abl: implications for haemopoietic cell growth and chronic myeloid leukaemia. *FEBS Lett.* **581**, 1329–1334 (2007).
25. Pellicano, F. *et al.* FOXO transcription factor activity is partially retained in quiescent CML stem cells and induced by tyrosine kinase inhibitors in CML progenitor cells. *Blood.* doi:10.1182/blood-2009-06-226621 (1 December 2009).

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

Author Contributions K.N. designed research, performed experiments, analysed data, and co-wrote the paper. T.H., T.M., Y.T, T.O. and N.M. performed experiments. Y.K. and S.N. provided technical support for the human cell experiments. A.H. designed research, analysed data and 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 K.N. (kazunaka@kenroku.kanazawa-u.ac.jp) or A.H. (ahirao@kenroku.kanazawa-u.ac.jp).

METHODS

Mice. $Foxo3a^{-/-}$ and littermate $Foxo3a^{+/+}$ mice of the C57BL/6 (F4) genetic background[15] were used in this study. C57BL/6 congenic mice were purchased from Sankyo-Lab Service. Animal care in our laboratory was in accordance with the guidelines for animal and recombinant DNA experiments of Kanazawa University. Imatinib (Glivec; purchased from Novartis) was administered to mice by oral gavage twice a day (200 mg kg^{-1} of body weight per day in water). Ly364947 (Merck Chemicals Ltd) was prepared as a 5 mg ml^{-1} stock solution in DMSO and intraperitoneally administered in saline to mice for 80 days at 10 mg kg^{-1} (of body weight) every 2 days, or for 7 days at 25 mg kg^{-1} (body weight) per day.

Flow cytometry and cell sorting. A FACSaria (BD Biosciences) and monoclonal antibodies recognizing the following markers were used for flow cytometry: Sca-1 (E13-161.7), CD4 (L3T4), CD8 (53-6.7), B220 (RA3-6B2), TER119 (Ly-76), Gr-1 (RB6-8C5), Mac1 (M1/70), IL7Rα chain (B12-1), FcγIII/II receptor (2.4G2) and CD34 (RAM34, pacific blue-conjugated) (all from BD Biosciences). Anti-c-Kit (ACK2) monoclonal antibody was from eBiosciences. A mixture of monoclonal antibodies recognizing CD4, CD8, B220, TER119, Mac1 and Gr-1 was used to identify Lin$^+$ cells.

Preparation of retrovirus. The cDNA encoding human *BCR-ABL* (gift from H. Honda) was cloned into the EcoRI site of the MSCV or MSCV-ires-GFP vector. Retroviral packaging cells (Plat-E) were transiently transfected with the MSCV-BCR-ABL-ires-GFP plasmid using FuGene6 reagent (Roche) and used for transplantation into mice as described later.

Generation of CML model and isolation of LIC population. Normal immature c-Kit$^+$Lin$^-$Sca-1$^+$ haematopoietic cells (KLS$^+$ cells) from $Foxo3a^{-/-}$ and littermate $Foxo3a^{+/+}$ mice were purified by flow cytometry and cultured in serum-free S-Clone SF-03 medium (Sanko Junyaku) supplemented with 10 ng ml^{-1} human TPO (thrombopoietin; PeproTech) plus 10 ng ml^{-1} mouse SCF (stem cell factor; Wako Pure Chemical). To generate our CML-like MPD mouse model, normal KLS$^+$ cells were infected with the above retrovirus carrying MSCV-BCR-ABL-ires-GFP using CombiMag (OZ Bioscience). The transduced KLS$^+$ cells were transplanted intravenously into lethally irradiated (9.5 Gy) C57BL/6 congenic mice along with 5×10^5 bone marrow mononuclear cells from C57BL/6 mice.

For the prospective isolation of LICs, mononuclear cells were isolated from bone marrow or spleen of CML-affected recipient mice at 12–14 days post-transplantation, and GFP$^+$ subpopulations (3×10^4) were recovered by flow cytometric cell sorting, followed by transplantation into C57BL/6 congenic mice as described earlier. For serial transplantations, GFP$^+$ KLS$^+$ cells (3×10^4 or 1.5×10^4) were collected and pooled from five BMT mice and transplanted into a second set of lethally irradiated congenic recipient mice along with 5×10^5 normal bone marrow mononuclear cells from C57BL/6 mice. The absolute number of CML LICs in the spleen of a recipient mouse was calculated as (total number of mononuclear cells isolated from the bone marrow (two legs) or spleen × frequency of GFP$^+$ KLS$^+$ cells (%) × 1/100).

Mouse cell colony-forming assay. Mouse cells were cultured in semi-solid medium (stroma-free) containing the cytokines SCF, IL-3, IL-6 and erythropoietin[15] (Methocult GF M3434; Stem Cell Technologies) at 37 °C in humidified air containing 5% CO$_2$ for 7 days. For co-culture assays, cells (1×10^3) were co-cultured on OP-9 stromal cells[23] under hypoxic conditions (5% O$_2$) for 5 days. After collecting and washing with PBS, colony formation was assessed in semi-solid medium as described earlier. For inhibition of TGF-β signalling, cells were co-cultured on OP-9 stromal cells in the presence of control (DMSO) or 2, 10 or 20 μM Ly364947 or SD208 (Sigma) for 5 days, followed by washing and transfer to semi-solid medium. For the combination treatment of TGF-β inhibitor plus imatinib, cells were first treated with DMSO or a TGF-β inhibitor on OP-9 stromal cells. At 24 h after starting culture, the cells were treated with further DMSO or 5 μM imatinib (STI571; provided by Novartis) and incubated for another 4 days (total 5 days). Thereafter, the cells were washed with PBS and transferred to semi-solid medium. The number of colonies was counted under the microscope.

Fluorescence immunostaining. Freshly isolated cells were incubated in S-Clone medium without cytokines for 30 min on poly-L-lysine (Sigma)-coated glass slides[26] at 37 °C under hypoxic conditions (5% O$_2$). To examine the effects of TGF-β signalling, cells were treated with 5 ng ml^{-1} TGF-β1 (R&D Systems), 10 μM Ly364947 (Merck), or vehicle control at 37 °C for 2 h. After fixation with 4% paraformaldehyde for 30 min and permeabilization with 0.25% Triton X-100 for 2 min, the cells were blocked by incubation in 2% BSA/PBS for 1 h. For immunostaining, the permeabilized cells were incubated with mouse or rabbit anti-FKHRL1 (Foxo3a) (FR1 or F2178; Sigma), rabbit anti-Foxo4 (9472; Cell Signaling), rabbit anti-phospho-Smad2/3 (Millipore), rabbit anti-phospho-Akt (D9E; Cell Signaling), or rabbit anti-phospho-S6 ribosomal protein (2211;

Cell Signaling) antibodies at 4 °C for 12 h. Primary antibodies were visualized by incubating the cells with AlexaFluor 546- or AlexaFluor 647-conjugated goat anti-mouse IgG or goat anti-rabbit IgG (Molecular Probes). Nuclei were stained with the DNA marker DAPI (Sigma). Stained slides were mounted using Fluoromount Plus (Diagnostic Biosystems), and fluorescent images were acquired using a Fluoview 1000 laser confocal microscope (Olympus) and Photoshop software (Adobe). Fluorescence intensities were quantified using ImageJ software. To evaluate the subcellular localization of Foxo3a, approximately 100 cells per group were counted under the microscope.

Cell cycle analysis and differentiation potential. To determine the cell cycle status of LICs *in vivo*, second BMT CML mice were administered BrdU intraperitoneally (100 mg kg^{-1} of body weight in saline; Sigma) for 12 h. Numbers of BrdU$^+$ LICs were assessed by immunostaining with anti-BrdU antibody (3D4; BD Biosciences) and flow cytometry. To examine the cell cycle status of LICs *in vivo* and determine the nuclear localization of Foxo3a, cells were co-stained with rabbit anti-FKHRL1 antibody (F2178; Sigma) to detect Foxo3a, and anti-mouse Ki67 antibody (B56; BD Biosciences).

To evaluate differentiation potential, the frequencies of CMP-like (CD34$^+$FcγIII/II receptor$^-$), GMP-like (CD34$^+$FcγIII/II receptor$^+$), and MEP-like (CD34$^-$FcγIII/II receptor$^-$) IL7Rα$^-$Lin$^-$c-Kit$^+$Sca-1$^-$ cells among GFP$^+$ bone marrow cells obtained from $Foxo3a^{+/+}$ and $Foxo3a^{-/-}$ CML mice at first BMT were analysed by flow cytometry.

Detection of apoptosis. To determine apoptosis *in vivo*, freshly isolated spleens and bones were immediately fixed with 4% paraformaldehyde, and tissue sections were prepared and stained using the TUNEL method (Roche). Isolated GFP$^+$KLS$^+$ cells from mice that had been transplanted with $Foxo3a^{+/+}$ and $Foxo3a^{-/-}$ CML LICs (second BMT) were either co-stained with Annexin-V-PECy5 (Abcam) and examined by flow cytometry, or stained *in vitro* using the TUNEL method (Roche).

Inhibition of endogenous Foxo by dominant-negative Foxo. Retroviruses carrying a dominant negative (dn)Foxo vector (MSCV-dnFoxo-ires-GFP[27]) or a control GFP vector (MSCV-ires-GFP) were generated as described earlier. To obtain dnFoxo CML LICs, normal KLS$^+$ cells (wild-type C57BL/6 CD45.1) were infected with retrovirus carrying the *BCR-ABL* gene (without GFP). These infected KLS$^+$ cells were transplanted into irradiated recipient mice (C57BL/6 CD45.2). At 14 days after transplantation, donor-derived KLS$^+$ cells were isolated and infected with retrovirus carrying the control GFP or dnFoxo-ires-GFP vector. To examine the reconstitution of CML LICs *in vivo*, retrovirus-infected CML LICs (unfractionated) were transplanted intravenously into lethally irradiated C57BL/6 congenic mice along with 5×10^5 bone marrow mononuclear cells from C57BL/6 mice. At 14 days after transplantation, absolute numbers of GFP$^+$KLS$^+$ cells were measured in recipient spleens. To examine LIC colony-forming ability *in vitro*, these GFP$^+$ CML LICs were cultured in semi-solid medium (stroma-free), or were co-cultured on OP-9 stromal cells as described earlier.

Quantitative real-time RT–PCR analysis. RNA samples were purified from fractionated GFP$^+$KLS$^+$ and GFP$^+$KLS$^-$ cells (1.0×10^5) using the RNeasy kit (QIAGEN) and reverse-transcribed using the Advantage RT-for-PCR kit (Takara-Clontech). Real-time quantitative PCR was performed using SYBR green Premix EXTaq (Takara) on an Mx3000P Real-time PCR system (Stratagene). The following primers were used: *Tgfb1*, 5′-TATGCTAAAGAGGTCACCCGCG-3′ and 5′-TGCTTCCCGAATGTCTGACG-3′; *Alk5* (also known as *Tgfbr1*), 5′-GATCGCCCTTTCATTTCAGAGG-3′ and 5′-AAACCGACCTTTGCCAATGC-3′; *Tgfbr2*, 5′-GAGAGCATGAAAGACAGTGTGC-3′ and 5′-CCAGCACTCGGTCAAAGTCT-3′; *Actb*, 5′-AGGTCATCACTATTGGCAACGA-3′ and 5′-CACTTCATGATGGAATTGAATGTAGTT-3′. The following cycle parameters were used: denaturation at 95 °C for 10 s, and annealing and elongation at 57 °C for *Actb*, and 60 °C for *Tgfb1*, *Tgfbr1* and *Tgfbr2*.

Analyses of primary human CML samples. Viable bone marrow mononuclear cells from patients with chronic phase CML were purchased from Cureline, Inc. (no. 16-122;) and AllCells LLC (no. 06-255 and 06-620). Cells were stained with anti-CD34 (8G12), anti-CD38 (HIT2), anti-CD3 (SK7), anti-CD16 (3G8), anti-CD19 (SJ25C1), anti-CD20 (L27), anti-CD14 (MφP9), and anti-CD56 (NCAM16.2) antibodies (all from BD Biosciences). A mixture of monoclonal antibodies recognizing CD3, CD16, CD19, CD20, CD14 and CD56 was used to identify Lin$^-$ cells. CD34$^+$CD38$^-$Lin$^-$ cells were purified by cell sorting. To examine the effects of treatment with Ly364947 alone or a combination of Ly364947 plus imatinib, CD34$^+$CD38$^-$Lin$^-$ cells were cultured on OP-9 stromal cells as described earlier. After collecting and washing in PBS, the colony-forming ability of LICs was evaluated by culture in semi-solid medium containing SCF, GM-CSF, IL-3, IL-6, G-CSF and erythropoietin (Methocult GF$^+$ H4435; Stem Cell Technologies).

Generation of CML blast crisis model. A mouse model of CML blast crisis was generated as previously described[28]. In brief, normal KLS$^+$ cells were co-infected

with a retrovirus carrying MSCV-BCR-ABL-ires-GFP and a retrovirus carrying MSCV-NUP98/HOXA9. The transduced Lin^-Sca-1^+ (LS^+) cells[11] were transplanted intravenously into lethally irradiated C57BL/6 congenic mice along with 5×10^5 bone marrow mononuclear cells from C57BL/6 mice.

Statistical analyses. Statistical differences were determined using the unpaired Student's *t*-test for *P*-values, and the long-rank non-parametric test for survival curves.

26. Ema, H. *et al.* Adult mouse hematopoietic stem cells: purification and single-cell assays. *Nature Protocols* **1**, 2979–2987 (2006).

27. Bouchard, C. *et al.* FoxO transcription factors suppress Myc-driven lymphomagenesis via direct activation of Arf. *Genes Dev.* **21**, 2775–2787 (2007).

28. Dash, A. B. *et al.* A murine model of CML blast crisis induced by cooperation between BCR/ABL and NUP98/HOXA9. *Proc. Natl Acad. Sci. USA* **99**, 7622–7627 (2002).

Multiple native states reveal persistent ruggedness of an RNA folding landscape

Sergey V. Solomatin[1], Max Greenfeld[2], Steven Chu[3,4]† & Daniel Herschlag[1]

According to the 'thermodynamic hypothesis', the sequence of a biological macromolecule defines its folded, active (or 'native') structure as a global energy minimum in the folding landscape[1,2]. However, the enormous complexity of folding landscapes of large macromolecules raises the question of whether there is in fact a unique global minimum corresponding to a unique native conformation or whether there are deep local minima corresponding to alternative active conformations[3]. The folding of many proteins is well described by two-state models, leading to highly simplified representations of protein folding landscapes with a single native conformation[4,5]. Nevertheless, accumulating experimental evidence suggests a more complex topology of folding landscapes with multiple active conformations that can take seconds or longer to interconvert[6–8]. Here we demonstrate, using single-molecule experiments, that an RNA enzyme folds into multiple distinct native states that interconvert on a timescale much longer than that of catalysis. These data demonstrate that severe ruggedness of RNA folding landscapes extends into conformational space occupied by native conformations.

Biopolymers face the challenge of folding a linear chain into complex three-dimensional conformations that have specific biological activities. According to the thermodynamic hypothesis, sequences of biological macromolecules have evolved to specify the active conformation as a minimum in free energy, ensuring that native states are more stable than an ensemble of all possible inactive conformations[1]. In the language of folding landscapes, the native conformation is the global energy minimum, separated from an ensemble of inactive conformations by a large energy gap[2–4]. However, the enormous complexity of the conformational space of a typical macromolecule—there are at least 10^{30} conformations for a small protein containing 100 amino acids[9]—raises the question of whether such energy minima are unique or whether there are multiple local energy minima that correspond to alternative active conformations. Observations of hysteresis in enzyme kinetics and of kinetic complexities in protein folding and unfolding and protein–ligand binding provided early evidence that some protein enzymes can exist in several catalytically active forms[6,10]. In the past decade, single-molecule experiments have provided additional evidence for multiple active conformers that slowly interconvert[11–13].

Functional RNAs represent a chemically distinct class of biopolymers that face a folding problem analogous to that of proteins. RNA folding landscapes have long been considered more rugged than protein landscapes[14–16], owing to the limited information content of RNA primary structure and the high stability of RNA secondary structures. Observations of very long timescales for folding and the occurrence of multiple, long-lived intermediates appear to support these views[15,17]. It is not known, however, whether ruggedness extends into the region of the native state, resulting in multiple active

conformations. Experiments with a small catalytically active RNA, the hairpin ribozyme, revealed several distinct active forms of these molecules[18,19]. Although these results are suggestive of landscape ruggedness, there is insufficient evidence to demonstrate that distinct forms are interconverting conformations, and the results leave open the possibility that covalent differences exist between the molecules that had different behaviour.

We have used single-molecule fluorescence resonance energy transfer (FRET) experiments to test whether the *Tetrahymena* group I ribozyme, a large, efficient RNA catalyst, folds into a unique native conformation or into multiple native conformations. Our results provide strong evidence for multiple conformations of the native state that interconvert but are separated by large energetic barriers.

The *Tetrahymena* group I ribozyme is a ~400-nucleotide RNA enzyme that folds in the presence of Mg^{2+} into an active form that catalyses cleavage of an oligonucleotide substrate by an exogenous guanosine nucleophile. Binding of the oligonucleotide substrate to the folded ribozyme occurs in two steps (Fig. 1). In the first step, the substrate binds to a single-stranded 5′-end region, forming the P1 duplex. In the second step, the P1 duplex docks into tertiary interactions in the active site[20,21].

We tested whether the ribozyme is folded into a single native conformation by measuring the thermodynamics and kinetics of docking for individual molecules using single-molecule FRET. If all
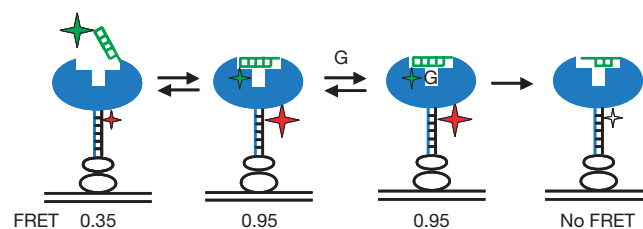

Figure 1 | Docking and cleavage of the oligonucleotide substrate by the *Tetrahymena* ribozyme, observed using single-molecule FRET. The ribozyme is shown in blue, the P1 duplex is shown in green and a DNA tether base-paired to a 3′-end extension of the ribozyme is shown in black. Streptavidin and biotinylated bovine serum albumin are used for surface attachment of the construct and are shown as clear ovals. The substrate is labelled with the FRET donor (Cy3 dye, green star) and the DNA tether is labelled with the FRET acceptor (Cy5 dye, red and white stars). The relative sizes of the stars indicate the relative fluorescence intensities of the dyes. Donor fluorescence is excited using the 532-nm laser. The FRET efficiency fluctuates between the levels of 0.35 in the open complex and 0.95 in the closed complex, as shown. After guanosine (G) binds, the substrate is cleaved and the Cy3-labelled 3′-end rapidly dissociates, leading to loss of Cy5 fluorescence, as indicated by the white star, and the absence of FRET[27].

[1]Department of Biochemistry, [2]Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA. [3]Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [4]Department of Physics and Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. †Present address: United States Department of Energy, Washington DC 20585, USA.

of the molecules were folded to a single native conformation, the docking behaviour for each single molecule would be identical to the ensemble average behaviour. However, individual ribozyme molecules from a single sample that were folded together and observed side by side under identical conditions instead displayed a broad distribution of docking behaviours (Fig. 2). For example, the docking equilibrium constants for the molecules in Fig. 2a–d vary by a factor of 300.

The accuracy of measuring docking equilibria from single-molecule FRET traces is limited by the finite length of each trace, caused by photobleaching. Therefore, a distribution of docking behaviours is expected even if all of the molecules are in the same conformation. Quantitative modelling of this effect (Supplementary Figs 1 and 2) demonstrated that the finite lengths can account for only a small fraction of the width of the observed distribution. More than six conformations with different docking thermodynamics are needed to account for this distribution (Fig. 2e and Supplementary Fig. 3). The docking and undocking rate constants determined for individual molecules also vary across a broad distribution (Fig. 2f) without discernable clustering into subgroups that might indicate a small number of distinct conformations.

A simple model that could account for the multiple docking states suggests that some of the observed molecules represent partially folded forms of the ribozyme that are capable of docking but have not reached the native state. To test this possibility, we determined the catalytic activity of molec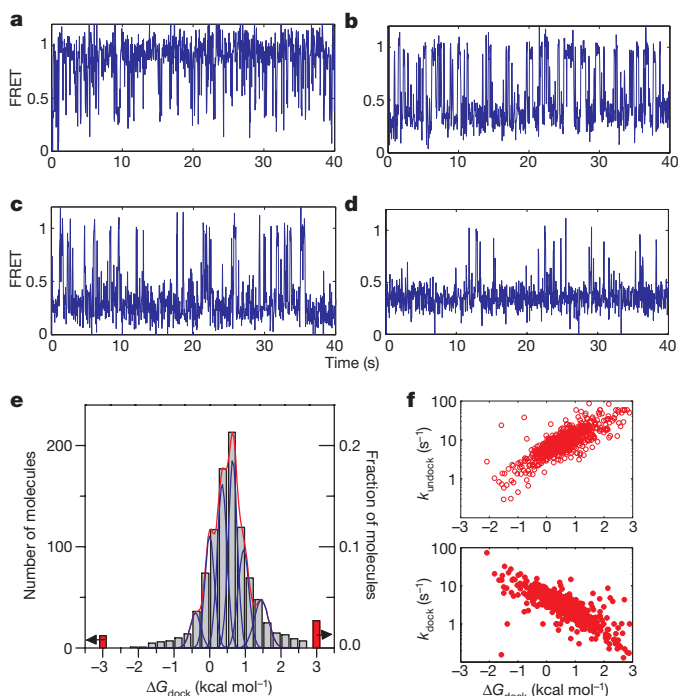ules across the broad range of docking behaviours. We first measured the docking equilibrium for each molecule and then induced cleavage by adding guanosine to allow chemical catalysis to occur (Fig. 1). The proportion of molecules in which the substrate was uncleaved decayed mono-exponentially with an end point of 6%, indicating that 94% of the molecules were catalytically active and that only 6% were inactive or had much lower activity. Furthermore, nearly all of the molecules throughout the distribution of docking conformations were catalytically active (Supplementary Table 1).

We next determined whether molecules in different docking conformations, as assessed by their distinct docking equilibria, exhibited the same or distinct cleavage kinetics. We binned molecules with similar docking behaviours (Fig. 3a) and analysed cleavage kinetics within each bin (Fig. 3b and Supplementary Table 1). Both the cleavage rate constant, k_{cleave}, and the end points were similar for all bins, despite the docking equilibria differing among them by a factor of at least 800. Thus, molecules that formed docked states with the same catalytic activity had distinct docking behaviours, providing direct evidence for a heterogeneous native state. Additional experiments suggest that the heterogeneity arises from different molecules forming different subsets of the tertiary docking interactions with the functional groups on the P1 duplex (Supplementary Fig. 4).

Multiple native conformations of the ribozyme observed in our experiments could represent multiple distinct folds of the same sequence, as implicitly assumed above. Alternatively, they could instead represent similar folded states of RNA molecules that have different sequences because of errors in synthesis, unintended chemical modifications or degradation. Covalent heterogeneity of RNA has been considered as a possible origin of multiple conformations of the hairpin ribozyme, although no direct evidence for or against this proposal has been accumulated, and it may be difficult to rule out covalent heterogeneity decisively using assays such as mass spectrometry because modifications can be mass neutral and the formation of multiple modified species each in low abundance can be difficult to detect[19,22].

Stable covalent conformational differences can be distinguished from interconvertible conformational differences if interconversion between different conformations can be observed under some conditions. If each detected 'conformation', or behaviour, can interconvert to others, conformational heterogeneity is strongly suggested, whereas the absence of extensive interconversion under any condition leaves open the possibilities of covalent heterogeneity or extremely stable and long-lived RNA species[23,24].

We determined whether molecules can change their docking behaviour by comparing docking of individual molecules before and after unfolding under mild denaturing conditions. Removing Mg^{2+} using EDTA disrupts the tertiary structure of the ribozyme but leaves most of the secondary structure intact and allows the molecules to remain

Figure 2 | Distribution of docking behaviours of individual ribozyme molecules. a–d, Representative examples of FRET traces, which give docking equilibria of $K_{dock} = 8$ (**a**), 0.5 (**b**), 0.08 (**c**) and 0.03 (**d**), corresponding to docking free-energy values of $\Delta G_{dock} = -1.4$, 0.4, 1.4 and 2.0 kcal mol^{-1}, respectively. **e**, The distribution of docking behaviours of individual ribozyme molecules (grey bars). Red bars correspond to molecules that displayed no docking or no undocking transitions, for which only the limits of ΔG_{dock} could be determined (as indicated by arrows). Blue peaks show the expected distribution widths for individual docking conformations (Supplementary Information). The red line shows the best fit of the distribution to a sum of distributions from six docking conformations. **f**, Docking (k_{dock}) and undocking (k_{undock}) rate constants of individual molecules. Experiments were performed in a standard buffer containing 10 mM $MgCl_2$ at 22 °C. The substrate oligonucleotide was CCCUC(2′-methoxyU)AAACC-Cy3.
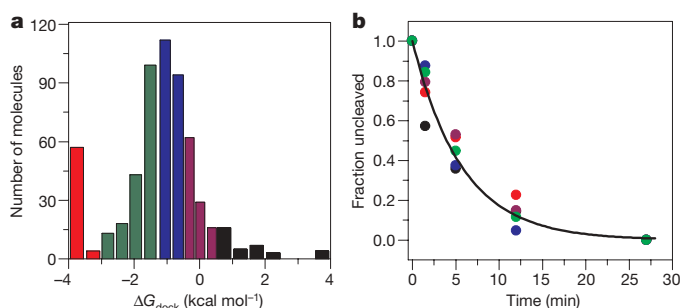
Figure 3 | Catalytic activities of molecules from different parts of docking distribution are the same. a, Distribution of docking behaviours was split into five bins marked using different colours. Cleavage experiments were performed with the CCCUC(2′-deoxyU)A-Cy3 substrate oligonucleotide, which docks more stably than the CCCUC(2′-methoxyU)AAACC-Cy3 substrate in Fig. 2. **b**, Cleavage kinetics within each bin were the same as the global average (line indicates global fit). Kinetic parameters for the global fit and fits for individual bins are given in Supplementary Table 1.

tethered to the surface. As illustrated in Fig. 4a, if all molecules quickly interconvert in the unfolded state, each molecule could then refold to any of the docking conformations, independent of which conformation it populated before unfolding. We grouped molecules into three non-overlapping subpopulations on the basis of their docking equilibrium constant before unfolding (Fig. 4b) and compared the distributions of conformations that each molecule populated after refolding. As shown in Fig. 4c, molecules from each of the three original subpopulations had docking behaviours that ranged across the entire distribution. (For traces demonstrating interconversion, see Fig. 4e and Supplementary Fig. 7.) These observations indicate that in the unfolded state all or nearly all of the conformational states can interconvert to any other conformation. Incomplete overlap of the distributions suggests either that some interconversions are slow on the timescale of minutes even in the unfolded state or that covalent differences are present in a subset of the molecules.

The original observation of multiple docking behaviours of active molecules (Figs 2 and 3) indicated that in the folded state the interconversions must be slow relative to the minute timescale of the observations. If interconversions in the folded state were slower than in the unfolded state, it would indicate that the energy landscape

becomes more rugged going from the unfolded state to the native state. To address this, we evaluated the extent of interconversion in the folded state. The data in Fig. 4d show that there was little interconversion between docking behaviours even after 40 min; that is, most of the molecules 'remembered' their initial docking equilibria. Nevertheless, some of the molecules changed their docking behaviours during this period (9% of all molecules; Fig. 4d and Supplementary Information). A simple explanation that accounts for the much smaller extent of interconversion in the folded state than in the unfolded state is that the energy barriers separating different conformations are higher in the folded state and the rate of interconversion is hence lower ($\sim$100-fold). The data on interconversion under different conditions suggest that the ribozyme folding landscape is moderately rugged in the region occupied by unfolded conformations and becomes more rugged towards the region occupied by the native conformations.

The multiple active conformations of the *Tetrahymena* group I ribozyme offer a glimpse of the full complexity of the folding landscape of a functional RNA. Taken together with the previous evidence for proteins, the results presented here suggest that the complexity of native folded states may be an inherent property of biopolymers.
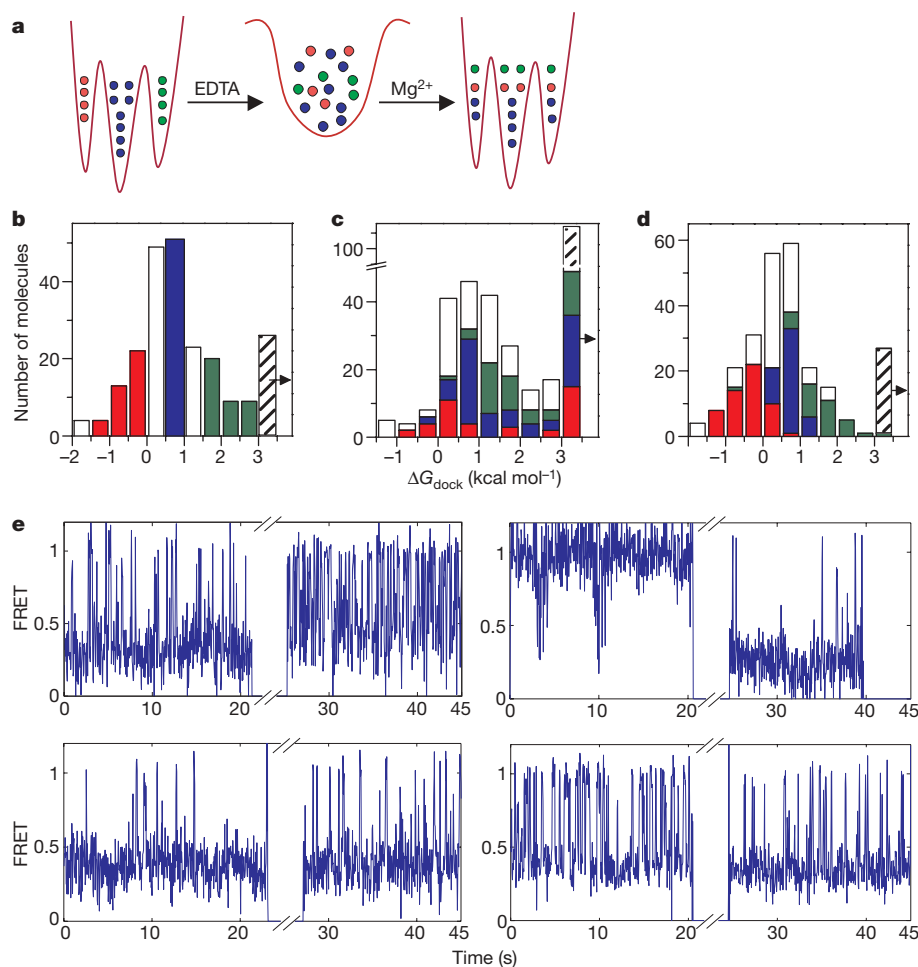

Figure 4 | Interconversion of docking behaviours. a, Schematic illustrating changes in docking following fast interconversion in the unfolded state. If interconversions were very slow, each molecule would display the same docking behaviour before and after unfolding. **b,** Distribution of ΔG_{dock} before unfolding. Red, blue and green bars indicate three selected conformations. White bars indicate molecules not included in the analysis to ensure non-overlap of selected conformations. The rightmost bin contains molecules that displayed no docking transitions, for which only the limits of ΔG_{dock} could be determined (as indicated by the arrow). **c,** Distribution of ΔG_{dock} after refolding. The rightmost bin contains mostly misfolded molecules (Supplementary Information)[26,29]. Coloured bars indicate the

number of molecules in the correspondingly coloured bins before unfolding (in **b**). **d,** Distribution of ΔG_{dock} after 40 min in the folded state. Coloured bars indicate the number of molecules in the correspondingly coloured bins at the zero time point. The rightmost bin contains molecules that displayed no docking transitions, as in **b. e,** Example traces showing interconversions after refolding (see also Supplementary Fig. 7). Breaks in FRET traces correspond to the time spent unfolded. Before and after refolding, the ΔG_{dock} values for the molecules shown were 1.2 and 0.1 (top left), -1.8 and 2.1 (top right), 2.0 and 1.2 (bottom left), and 0.4 and 1.5 kcal mol^{-1} (bottom right), and the *P* values indicating the significance level for interconversion were 0.01, 10^{-15}, 0.1 and 0.1, respectively.

These observations prompt us to revisit the Levinthal paradox in a different context. Levinthal noted that proteins cannot sample all of conformational space on a biological timescale, leading to the notion that there must be a preferred pathway to the native state[25]. More recent work has emphasized the likelihood of there being multiple pathways to the folded state of proteins and RNAs. Multiple pathways, combined with limitations to full sampling, may lead to multiple 'end points', that is, multiple native-state local minima. This picture, although counter to traditional viewpoints, has some support in the protein-folding literature and is strongly supported for the *Tetrahymena* group I RNA by this and previous[26] folding studies. We further speculate that perhaps folded RNAs or proteins in different conformations may respond to cellular or environmental cues in different ways and contribute to stochastic diversity.

METHODS SUMMARY

We synthesized the L-16 form of the *Tetrahymena* group I ribozyme, extended at the 3′-end by a 26-nucleotide T2 tail, by *in vitro* transcription, and purified it by denaturing polyacrylamide gel electrophoresis. Oligonucleotides were obtained commercially (Integrated DNA Technologies and Dharmacon), fluorescently labelled and purified by denaturing polyacrylamide gel electrophoresis. We bound the DNA tether to the ribozyme by annealing (heating at 95 °C for 3 min; cooling to 50 °C at $0.1 \, °C \, s^{-1}$) and then folding in 10 mM MgCl$_2$ (30 min at 50 °C).

The oligonucleotide substrates were bound to the tether-bound, prefolded ribozyme at 1 μM concentration (10 min at 22 °C). The ribozyme was then diluted to 30–100 pM final concentration for deposition on slides coated with BSA-biotin/streptavidin, as described[27]. Single-molecule FRET experiments were carried out on a home-built prism-based total-internal-reflection microscope, similar to one previously described[28]. Data collection and analysis were performed using home-written programs in C++ and MATLAB R2008a (Mathworks). Details of the methods, including sample preparation, data collection and analysis, and numerical simulations, are presented in full Methods.

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

1. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* 181, 223–230 (1973).
2. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins Struct. Funct. Genet.* 21, 167–195 (1995).
3. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* 28, 361–368 (2003).
4. Zwanzig, R. Two-state models of protein folding kinetics. *Proc. Natl Acad. Sci. USA* 94, 148–150 (1997).
5. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Struct. Biol.* 4, 10–19 (1997).
6. Schmid, F. X. & Blaschek, H. A. Native-like intermediate on the ribonuclease A folding pathway. *Eur. J. Biochem.* 114, 111–117 (1981).
7. Jennings, P. A., Finn, B. E., Jones, B. E. & Matthews, C. R. A reexamination of the folding mechanism of dihydrofolate reductase from *Escherichia coli*: verification and refinement of a four-channel model. *Biochemistry* 32, 3783–3789 (1993).
8. Kamagata, K., Sawano, Y., Tanokura, M. & Kuwajima, K. Multiple parallel-pathway folding of proline-free staphylococcal nuclease. *J. Mol. Biol.* 332, 1143–1153 (2003).
9. Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25, 331–339 (2000).

10. Frieden, C. Slow transitions and hysteretic behavior in enzymes. *Annu. Rev. Biochem.* 48, 471–489 (1979).
11. Flomenbom, O. *et al.* Stretched exponential decay and correlations in the catalytic activity of fluctuating single lipase molecules. *Proc. Natl Acad. Sci. USA* 102, 2368–2372 (2005).
12. Lu, H. P., Xun, L. & Xie, X. S. Single-molecule enzymatic dynamics. *Science* 282, 1877–1882 (1998).
13. English, B. P. *et al.* Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nature Chem. Biol.* 2, 87–94 (2005); erratum 2, 168 (2006).
14. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* 270, 20871–20874 (1995).
15. Treiber, D. K. & Williamson, J. R. Exposing the kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* 9, 339–345 (1999).
16. Chen, S.-J. & Dill, K. A. RNA folding energy landscapes. *Proc. Natl Acad. Sci. USA* 97, 646–651 (2000).
17. Pan, J., Thirumalai, D. & Woodson, S. A. Folding of RNA involves parallel pathways. *J. Mol. Biol.* 273, 7–13 (1997).
18. Zhuang, X. *et al.* Correlating structural dynamics and function in single ribozyme molecules. *Science* 296, 1473–1476 (2002).
19. Tan, E. *et al.* A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc. Natl Acad. Sci. USA* 100, 9308–9313 (2003).
20. Herschlag, D. Evidence for processivity and two-step binding of the RNA substrate from studies of J1/2 mutants of the *Tetrahymena* ribozyme. *Biochemistry* 31, 1386–1399 (1992).
21. Bevilacqua, P. C., Kierzek, R., Johnson, K. A. & Turner, D. H. Dynamics of ribozyme binding of substrate revealed by fluorescence-detected stopped-flow methods. *Science* 258, 1355–1358 (1992).
22. Ditzler, M. A., Rueda, D., Mo, J., Hakansson, K. & Walter, N. G. A rugged free energy landscape separates multiple functional RNA folds throughout denaturation. *Nucleic Acids Res.* 36, 7088–7099 (2008).
23. Lindahl, T., Adams, A. & Fresco, J. R. Renaturation of transfer ribonucleic acids through site binding of magnesium. *Proc. Natl Acad. Sci. USA* 55, 941–948 (1966).
24. Korennykh, A. V., Plantinga, M. J., Correll, C. C. & Piccirilli, J. A. Linkage between substrate recognition and catalysis during cleavage of sarcin/ricin loop RNA by restrictocin. *Biochemistry* 46, 12744–12756 (2007).
25. Levinthal, C. Are there pathways for protein folding? *J. Chim. Phys.* 65, 44–45 (1968).
26. Russell, R. *et al.* Exploring the folding landscape of a structured RNA. *Proc. Natl Acad. Sci. USA* 99, 155–160 (2002).
27. Zhuang, X. *et al.* A single-molecule study of RNA catalysis and folding. *Science* 288, 2048–2051 (2000).
28. Sattin, B. D., Zhao, W., Travers, K., Chu, S. & Herschlag, D. Direct measurement of tertiary contact cooperativity in RNA folding. *J. Am. Chem. Soc.* 130, 6085–6087 (2008).
29. Russell, R. & Herschlag, D. Probing the folding landscape of the *Tetrahymena* ribozyme: commitment to form the native conformation is late in the folding pathway. *J. Mol. Biol.* 308, 839–851 (2001).

METHODS

Reagents. We used the following enzymes. T7 RNA polymerase (His-tagged) was overexpressed in BL21 DE3 cells and purified by chromatography on Ni^{2+} resin; this was followed by dialysis against a storage buffer containing 20 mM phosphate buffer (pH 7.5), 100 mM NaCl, 10 mM DTT, 1 mM EDTA and 50% glycerol. Glucose oxidase (Type VII from *Aspergilius niger*, Sigma-Aldrich) and catalase (Roche) were used as supplied.

We used the following chemicals. D-glucose (SigmaUltra, 99.5%), guanosine 5′-monophosphate (Sigma-Aldrich, ≥99%), dNTPs (Fermentas, >98%), NTPs (Sigma, 95–99%), Trolox ((±)-6-hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, Aldrich, >97%), egg lecitin (Avanti Lipids) and cap-biotin phosphatidylethanolamine (Avanti Polar Lipids) were used without further purification.

Ribozyme preparation. L-16T2 ribozyme, which is a version of the L-21 ScaI ribozyme extended at the 5′-end with the sequence GGUUU and at the 3′-end with the sequence ACCAAAAUCAACCUAAAACUUACACA, was prepared by PCR amplification of a DNA template from a plasmid pT7L-21 with the extensions encoded on the primers. The ribozyme was prepared by a run-off *in vitro* transcription of the DNA template in conditions that minimize self-processing of the 5′-end (30 min at 30 °C with 4 mM MgCl$_2$ present). RNA was purified by denaturing polyacrylamide gel electrophoresis (8% of 29:1 acrylamide:bisacrylamide) and stored as a ~20 µM stock solution at −20 °C.

Oligonucleotide labelling. We purchased RNA oligonucleotides CCCUCdUA (dU, 2′-deoxyuridine), CCCUCmUAAACC (mU, 2′-methoxyuridine) and CCCmUCUAAACC, with 3′-amino modification from Dharmacon (now Thermo Scientific) and deprotected them according to manufacturer's instructions. We modified the 3′-amino groups with Cy3-NHS (Invitrogen) and purified labelled oligonucleotides from unlabelled and shorter oligonucleotides (present in synthetic oligonucleotides because of <100% coupling efficiency) by denaturing polyacrylamide gel electrophoresis (20% of 29:1 acrylamide:bisacrylamide). After purification, oligonucleotides were dissolved in distilled water and stored at −80 °C.

We purchased DNA oligonucleotide T2b (TGTGTAAGTTTTAGGTTGATTT TGGT) with 5′-biotin and 3′-amino modifications from Integrated DNA Technologies and labelled it with Cy5-NHS and purified it as described above.

Sample preparation. Solutions of L-16T2 ribozyme and T2b-Cy5 were mixed together (2:1 molar ratio; final concentration of RNA, 1 µM) in an annealing buffer (50 mM NaMES (pH 6.0), 200 mM NaCl) and annealed by heating at 95 °C for 3 min then cooling to 50 °C at 0.1 °C s^{-1}. We initiated folding by adding MgCl$_2$ to a final concentration of 10 mM. The ribozyme was folded for 30 min at 50 °C (ref. 30).

Oligonucleotide substrates were bound by mixing 2 µl prefolded L-16*T2b-Cy5 (1 µM) and 0.5 µl S-Cy3 (2 µM) in 50 mM NaMES (pH 6.0) and 10 mM MgCl$_2$, and incubating at room temperature (22 °C) for 10 min. We diluted ternary complex L-16*T2b-Cy5*S-Cy3 to a final concentration of 30–100 pM for deposition on the slides.

Quartz slides (G. Finkenbeiner) were coated with biotinylated bovine serum albumin (Sigma; 1 mg ml^{-1} for 10 min), washed thoroughly with 50 mM NaMOPS (pH 7.0), then coated with streptavidin (Sigma; 0.1 mg ml^{-1} for 10 min) and washed thoroughly with 50 mM NaMOPS (pH 7.0) and, finally, with 50 mM NaMES (pH 6.0) and 10 mM MgCl$_2$.

Ternary complexes L-16*T2b-Cy5*S-Cy3 were deposited over the coated slides for 10 min and then thoroughly washed with 50 mM NaMES and 10 mM MgCl$_2$. We performed measurements of docking in a standard buffer containing 50 mM NaMES (pH 7.0), 10 mM MgCl$_2$, 100 mM NaCl and an oxygen scavenging system (100 units ml^{-1} glucose oxidase, 1,000 units ml^{-1} catalase, 10 mM D-glucose and 2 mM Trolox). The cleavage buffer was the same, except we used 50 mM NaMOPS (pH 8.1) instead of NaMES, and added 1 mM guanosine 5′-monophosphate. Higher pH accelerates cleavage, thereby facilitating measurement of the chemical step, and has been shown to have no effect on docking[30].

TIRF microscope and data acquisition. A diode-pumped solid-state green laser (532 nm; Gem, Laser Quantum) and a red laser (635 nm; Hitachi HL6344G diode; maximum power, 10 mW) were combined using dichroic mirrors and focused through a prism onto a sample contained in a flow cell made from a quartz slide and a cover slip glued together using double-sided tape. Laser beams entered the prism at an angle ensuring total internal reflection of the exciting light. We collected images using a ×60 water-immersion Nikon objective (numerical aperture, 1.2), filtered through a 550-nm long-pass filter (Chroma Technology) to remove scattered excitation light and chromatically separated using dichroic mirrors (635-nm cut-off) into a 'green' image and a 'red' image. We focused the 'green' image, filtered through a 580/30-nm band-pass filter, and the 'red' image, filtered through 670/30-nm band-pass filter, onto the left and right halves of a back-illuminated charge-multiplying charge-coupled device

(CCD) (Cascade:128+, Photometrics, Roper Scientific), respectively. Full CCD images (128 × 128 pixels) were read out in 40-ms frames with a conversion gain of 3 and multiplication gain of 3,500. The red laser was typically switched on only for the first 0.5 s of data acquisition and at the end of data acquisition to determine which of the molecules had fluorescently active Cy5 dye. We adjusted the intensity of the green laser to achieve an average signal-to-noise ratio of >5, which typically required a power of ~20 mW at the laser aperture.

Data analysis and simulations. To record FRET data, we determined the positions of spots of interest within each data acquisition 'movie' by averaging the first 30 frames and finding pixels on the 'red' side of the CCD with an intensity that exceeded a certain threshold (typically a threshold of 5σ, where σ is the standard deviation of the background fluorescence). The corresponding positions on the 'green' side were determined by applying linear offsets (determined independently from images of fluorescent beads that are visible on both sides of the CCD) and further refining by the 'affine' algorithm of MATLAB. The local (7 × 7 pixel) background was subtracted from each spot for each frame. Time traces of the fluorescent intensity (I) of Cy3 and the Cy5 for each spot were recorded and used to calculate a FRET trace (for each frame, FRET equals $I_{red}/(I_{green} + I_{red} - I_{cross-talk})$). The value of the cross-talk (intensity in the red channel arising from Cy3 fluorescence because of imperfect chromatic separation) was determined by measuring the 'red' intensity for molecules that contain only Cy3.

To identify molecules, for each spot we visually screened Cy 3 and Cy5 traces to determine whether they corresponded to a single RNA molecule, and accepted that they did if they met the following criteria: (1) they exhibit single-step photobleaching (the fluorescence intensity goes to the background level within one frame); (2) they have a stable signal (the average fluorescence intensity along the trace is constant; it does not gradually decrease or increase); (3) they have a normal signal strength (fluorescence intensity is within a factor of two of the average; very dim and very bright spots were rejected); (4) Cy5 is fluorescent ('red' fluorescence is detected when the red laser is switched on at the beginning and the end of data acquisition; if Cy5 was fluorescent at the beginning, but not at the end, of data acquisition, the trace was truncated at the last frame where Cy5 was fluorescent).

We measured the thermodynamics and kinetics of individual models in the following way. FRET traces for >95% of the molecules (defined by the above criteria) displayed fluctuations between two FRET levels, low FRET (0.35) and high FRET (0.95), respectively corresponding to undocked and docked state, as described previously[28]. Transitions between FRET levels were identified by a standard thresholding method. The threshold was set at 0.7. On average, each trace contained ~30 transitions, allowing precise estimates of the rate constants. Rate constants were determined by creating histograms of dwell times in the docked state (to measure k_{dock}) and in the undocked state (to measure k_{undock}), and fitting histograms to a sum of exponentials. For most traces, the histograms were well described by a single exponential decay. Equilibrium constants between docked and undocked conformations were determined as a ratio of the total time spent in docked (t_{dock}) and in undocked (t_{undock}) states: $K_{dock} = t_{dock}/t_{undock}$. From this, the free energy of docking was calculated as: $\Delta G_{dock} = -RT \times \ln K_{dock}$, where R is the molar gas constant and T denotes temperature.

Because of photobleaching, each molecule can be observed for only a limited time, $t_{life} \approx 1/k_{bleach}$, which makes the value of ΔG_{dock} calculated for each trace a random parameter distributed around a true mean. For a trace that fluctuates between two FRET levels with a forward rate constant k_{dock} and a reverse rate constant k_{undock}, the width of the ΔG_{dock} distribution depends only on k_{dock}, k_{undock} and k_{bleach}. To estimate the extent of broadening, we first measured the distribution of trace lengths and from it calculated the photobleaching rate constant $k_{bleach} = 0.03 \, s^{-1}$ (Supplementary Fig. 1). We then performed numerical simulations of the ΔG_{dock} distribution for a broad range of k_{dock} and k_{undock} values, keeping k_{bleach} fixed. For each pair of parameters k_{dock} and k_{undock}, 1,000 traces were simulated by a lab-written program in MATLAB, ΔG_{dock} for each trace was calculated as above and histograms were fitted using a Gaussian distribution. As the results of the simulations demonstrate (Supplementary Fig. 2), for any combination of the rate constants k_{dock} and k_{undock} that were experimentally measured, the width ($w = 2\sigma$, where σ is the standard deviation) of the ΔG_{dock} distribution is small (typically ≤0.3 kcal mol^{-1}) in comparison with the experimentally observed width (~1 kcal mol^{-1}). Comparing the results of these simulations with the experimental data, we see that a single population of molecules, characterized by a single set of ΔG_{dock}, k_{dock} and k_{undock} values, cannot account for the experimentally observed width of the ΔG_{dock} distribution. As illustrated in Supplementary Fig. 3, at least six distinct populations, each with different ΔG_{dock}, k_{dock} and k_{undock} values, are required to account for the full width of the distribution. These six conformations should be considered a minimum number of coexisting conformations that can be determined given the limited resolution of ΔG_{dock} that is imposed, ultimately, by photobleaching.

Cleavage experiments. First, docking was observed for 20 s essentially as described above. Then the laser was switched off and the cleavage buffer containing 1 mM guanosine 5′-phosphate was flowed through the sample at ~40 μl s^{-1} for 5 s. Immediately after buffer replacement, and then subsequently after several time intervals, we switched the laser on for 2 s at a time (to allow confident detection of remaining molecules while limiting photobleaching) and recorded images of the slide. Fluorescent spots were defined as molecules with the uncleaved substrates if they fulfilled the criteria of signal stability and intensity (criteria 2 and 3 above). Control experiments establishing that the rate of substrate disappearance in the absence of guanosine is negligibly small were performed in exactly the same way, but omitting guanosine 5′-phosphate from the cleavage buffer.

Unfolding/refolding experiments. We performed the experiments in two steps: in the first step, docking was observed for 20 s essentially as described above. Then the laser was switched off and the unfolding buffer containing 12 mM EDTA was flowed through the sample at ~40 μl s^{-1} for 5 s. Unfolding was ensured by quickly switching the laser on and verifying that all molecules were in low FRET state. After 1 min in EDTA, the standard folding buffer was flowed back in and docking was observed for 20 s. Docking equilibria for each molecule before and after unfolding were measured as described above. To test whether molecules changed their docking behaviour, we calculated the uncertainty of measuring the docking equilibrium as the width of the simulated distribution (Supplementary Fig. 2). The P values were calculated using a right-tailed chi-squared variance test implemented in MATLAB with the null hypothesis that the difference between docking equilibria before and after unfolding is equal to, or smaller than, the uncertainty in the measurement.

30. Narlikar, G. J., Bartley, L. E., Khosla, M. & Herschlag, D. Characterization of a local folding event of the *Tetrahymena* group I ribozyme: effects of oligonucleotide substrate length, pH, and temperature on the two substrate binding steps. *Biochemistry* **38**, 14192–14204 (1999).

Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family

Beth G. Wensley[1], Sarah Batey[1]†, Fleur A. C. Bone[1], Zheng Ming Chan[1], Nuala R. Tumelty[1], Annette Steward[1], Lee Gyan Kwa[1], Alessandro Borgia[1]† & Jane Clarke[1]

Energy landscape theory is a powerful tool for understanding the structure and dynamics of complex molecular systems, in particular biological macromolecules[1]. **The primary sequence of a protein defines its free-energy landscape and thus determines the folding pathway and the rate constants of folding and unfolding, as well as the protein's native structure. Theory has shown that roughness in the energy landscape will lead to slower folding[1], but derivation of detailed experimental descriptions of this landscape is challenging. Simple folding models[2,3] show that folding is significantly influenced by chain entropy; proteins in which the contacts are local fold quickly, owing to the low entropy cost of forming stabilizing, native contacts during folding[4,5]. For some protein families, stability is also a determinant of folding rate constants[6]. Where these simple metrics fail to predict folding behaviour, it is probable that there are features in the energy landscape that are unusual. Such general observations cannot explain the folding behaviour of the R15, R16 and R17 domains of α-spectrin. R15 folds ~3,000 times faster than its homologues, although they have similar structures, stabilities and, as far as can be determined, transition-state stabilities[7–10]. Here we show that landscape roughness (internal friction) is responsible for the slower folding and unfolding of R16 and R17. We use chimaeric domains to demonstrate that this internal friction is a property of the core, and suggest that frustration in the landscape of the slow-folding spectrin domains may be due to misdocking of the long helices during folding. Theoretical studies have suggested that rugged landscapes will result in slower folding; here we show experimentally that such a phenomenon directly influences the folding kinetics of a 'normal' protein, that is, one with a significant energy barrier that folds on a relatively slow, millisecond–second, timescale.**

The folding rate constants of all but the fastest-folding proteins are assumed to be determined by the free-energy barrier between the unfolded and transition states and a kinetic prefactor. This is reflected in the ability of simple structural parameters such as contact or long-range order[2,3] to predict folding rate constants. R15, R16 and R17 (~30% pairwise sequence identity) have the same three-helix-bundle structure (Fig. 1a) and similar thermodynamic stabilities, but R15 folds and unfolds approximately three orders of magnitude faster than R16 and R17 (Fig. 2a)[7]. The folding rate constant for R15 is well predicted by contact and long-range-order plots, but R16 and R17 are outliers (Supplementary Fig. 1). Because the difference is evident in both folding and unfolding, the simplest explanation is that the transition state of R15 is more structured (and thus more stable) than that of its homologues. It is not possible to measure the free energy of the transition state directly, but experimental data show that the three transition states are similar in terms of β_T (a

measure of compactness), and protein-engineering Φ-value analysis of all three suggests that they are generally similar in overall structural and energetic terms (they have the same average Φ values and the same regions of each structure are folded)[8–10]. If such significant differences in folding kinetics cannot easily be ascribed to differences in transition-state structure, it is necessary to look for an alternative explanation.

Energy landscape theory has shown that protein folding is best described as a Kramers-like[11] diffusive process across the energy landscape and that the folding rate constant is dependent on a number of aspects of the landscape[1]. For a one-dimensional free-energy surface with harmonic wells, the folding time, τ_f, is related to the shape of the energy landscape, $2\pi/\omega_D\omega_\ddagger$ (where ω_D and $\omega_\ddagger$ are terms that describe the curvature in the landscape in the denatured state and the transition state, respectively), the height of the energy barrier, $\Delta G^\ddagger$, and the diffusion coefficient, D_0, as follows:

$$\tau_f = \frac{1}{k_f} = \frac{2\pi RT}{\omega_D\omega_\ddagger D_0}\exp\left(\frac{\Delta G^\ddagger}{RT}\right) \qquad (1)$$

Here R is the molar gas constant and T is the temperature in Kelvin. One principle component of the diffusion coefficient is solvent friction, which slows folding. Diffusion over smooth landscapes is relatively fast,
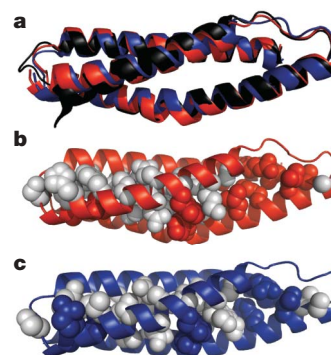

Figure 1 | Spectrin domains used in this study. a, Superposition of R15 (black), R16 (red) and R17 (blue) (pairwise backbone root-mean-squared deviation, <1 Å). **b**, R16o15c. Core residues (≤15% solvent-accessible surface area) shown in space-filling mode. Red, residues identical in R16 and R15; grey, the 22 residues that were substituted into R16 from R15. Respective sequence identities of R16o15c to its major (R16) and minor (R15) parents, 81% and 51%. **c**, R17o15c. Blue, residues identical in R17 and R15; grey, the 24 residues that were substituted into R17 from R15. Respective sequence identities of R17o15c to its major (R17) and minor (R15) parents, 79% and 53%.

[1]Department of Chemistry, University of Cambridge, MRC Centre for Protein Engineering, Lensfield Rd, Cambridge CB2 1EW UK. †Current addresses: Covagen AG, c/o ETH Zürich, Y17 M22, Life Science ETH/Uni, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland (S.B.); Biochemisches Institut, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland (A.B.).
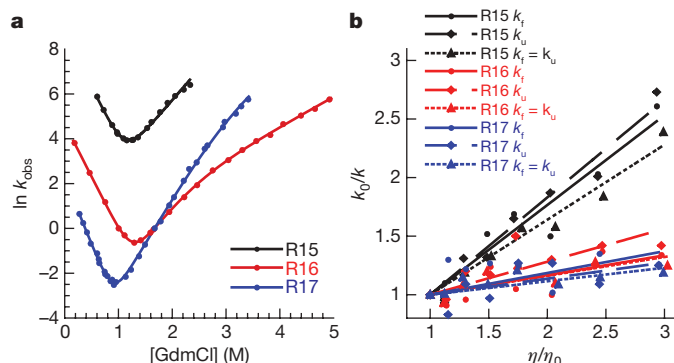
Figure 2 | Kinetic data for R15, R16 and R17 spectrin domains.
a, Chevron plots in 0 M glucose. R15 folds and unfolds significantly faster than R16 and R17. k_{obs}, observed rate constant. Square brackets denote concentration. **b**, Dependence of the relative rate constants, k_0/k_u and k_0/k_f, on relative solvent viscosity, η/η_0. Values of k_f and k_u were determined at $\Delta G_{D-N} = 1.5$ kcal mol^{-1}, and $k_f = k_u$ at $\Delta G_{D-N} = 0$ kcal mol^{-1}. R16 and R17 rate constants show significantly lower viscosity dependence. The lines are least-squares fits of the data to a linear equation that forces the line through the reference point (1, 1).

so the folding time of a given protein generally depends only on the height of the free-energy barrier and the solvent viscosity. However, theory suggests that when the landscape is rough, diffusion will be slowed because it will take time for the chain to escape from the local minima that constitute the rougher (more frustrated) landscape[1]. One possible explanation for the slow folding of R16 and R17 is that diffusion is impeded by kinetic traps in the energy landscape. This would be remarkable because landscape roughness has never been observed explicitly for any relatively slow-folding two-state protein, although frustration resulting from formation of stable, misfolded intermediates has been described[12].

A simple way to test this hypothesis is to determine the dependence of folding and unfolding rate constants on solvent viscosity[13,14]. Equation (1) has been recast, empirically[15], as

$$k = \frac{C}{\eta + \sigma} \exp\left(\frac{-\Delta G^{\ddagger}}{RT}\right) \qquad (2)$$

to relate the rate constant for folding, k_f, or unfolding, k_u, to solvent viscosity, η, and the 'internal friction' of the protein, σ. Thus, for a system with a smooth energy landscape (where internal friction is negligible, that is, σ is significantly less than η), k_f is inversely proportional to solvent viscosity. Such a relationship has been observed for a number of small proteins[16–18]. (Solvent slaving has been proposed as an alternative mechanism to explain the effects of solvent viscosity on folding kinetics[19]. This is discussed and compared with the model used in this work in Supplementary Results and Discussion.)

Solvent viscosity is easily controlled by addition of small-molecule viscogens; however, these tend to increase the stability of proteins. The isostability approach, whereby the stabilization caused by the viscogen is counteracted by the addition of a chemical denaturant, has been widely used[16–18,20,21]. We applied this approach to investigate the hypothesis that internal friction slows the folding of R16 and R17 but is insignificant in the fast-folding R15.

We investigated the equilibrium stability and folding kinetics of all three proteins over a range of solvent viscosities using glucose as the viscogen and guanidinium chloride (GdmCl) as denaturant (Supplementary Figs 2 and 3). Glucose increases the stability of all the proteins but the m value decreases. The m value is the dependence of the free energy of unfolding, ΔG_{D-N}, on denaturant concentration, and reflects the change in solvent-accessible surface area between the denatured and native states. This decrease is associated with a decrease in the refolding m value, whereas the unfolding m value is constant, suggesting that glucose causes a collapse of the denatured states (Supplementary Fig. 4). This effect is seen in all three proteins,

which, because the unfolding m value is unaffected by the viscogen, suggests that the transition-state position (and structure) is unaffected by glucose in all three cases. Folding and unfolding rate constants were determined at isostability in two ways. First, we used the equilibrium data to calculate the concentration of GdmCl at which $\Delta G_{D-N} = 1.5$ kcal mol^{-1} for each glucose concentration. We determined values of k_f and k_u at this GdmCl concentration from the fits of the chevron plots (Supplementary Fig. 3). Second, we used the chevron plots alone to determine rate constants for $\Delta G_{D-N} = 0$ (that is, for $k_f = k_u$).

We plot relative rate constants, k_0/k_f and k_0/k_u, against relative viscosity, η/η_0, in Fig. 2b (k_0 and η_0 are the quantities measured in 0 M glucose). Gradients for folding and unfolding rate constants were consistent. From equation (2), at isostability, if σ is small relative to η then the gradient of this plot will be approximately equal to one. If, however, σ is large relative to η, the gradient will be significantly less than one. The data for R15 show that both the folding and unfolding rate constants are strongly dependent on solvent viscosity, with gradients close to one (Supplementary Table 1), suggesting that internal friction is negligible for R15. However, the rate constants of R16 and R17 show only a very weak dependence on solvent viscosity (mean slopes of ~0.2). This provides strong evidence to suggest that R16 and R17 fold so much more slowly than R15 as a result, at least in part, of internal friction, that is, to roughness in the energy landscape. Similar investigations have been undertaken for four other proteins that fold on timescales comparable to those of the spectrin domains[16–18,20]. As seen here for R15, there is no evidence to suggest that internal friction plays any part in determining the folding rate constant of these proteins. R16 and R17 are unexpectedly atypical.

Using equation (2), it is possible to estimate the internal friction, σ, in the transition states of the spectrin domains (Supplementary Results and Discussion). For R15, $\sigma = 0.25 \pm 0.16$ cP, which is significantly lower than for water ($\eta \approx 1$ cP). By contrast, the σ values for R16 and R17 (4.4 ± 1.6 cP and 12.0 ± 6.6 cP, respectively) are significantly higher than for water. These high values of internal friction are similar to those found in studies of dynamics of essentially fully folded proteins[15,22].

These values of σ can be used to evaluate the relative magnitude of the ruggedness of the energy landscapes of our spectrin domains. Assuming a random (Gaussian) roughness distribution, the amplitude of the roughness of a one-dimensional energy landscape can be related to the effective diffusion coefficient, D^*, as follows[23]:

$$D^* = D \exp[-(\varepsilon/RT)^2]$$

Here D is the diffusion coefficient across a given smooth landscape and ε is the characteristic magnitude of the roughness (in units of energy). This expression can be used to estimate $\Delta\varepsilon$, the landscape roughness of the slow-folding proteins R16 and R17 relative to R15 (Supplementary Results and Discussion). For R16 $\Delta\varepsilon \approx 1.7RT$ and for R17 $\Delta\varepsilon \approx 2.0RT$, values similar both to that observed in experimental studies on peptides, denatured proteins and small, fast-folding proteins and to that predicted from theory[24–28]. As discussed in ref. 23, Gaussian noise gives one limit for the microscopic barrier heights; the alternative extreme, of periodic barriers of equal height, results in a slightly larger value of $\Delta\varepsilon$. It is remarkable that such a small increase in landscape roughness can result in such a significant change in the viscosity dependence of the folding kinetics. In fact, these microscopic barriers must be relatively small; kinetic traps involving barriers much larger than about $3RT$–$4RT$ would result in the accumulation of intermediates, which is not observed in spectrin domains[1,12,27].

Although the experiments detect energy landscape roughness in R16 and R17, they do not explain the source of this frustration. More than 200 variants of the spectrin domains have been investigated[8–10]. None significantly speeds the folding of R16 or R17 nor significantly slows the folding of R15. We designed four core-swap proteins (Fig. 1b, c and Supplementary Fig. 5) by grafting the core residues from one parent, the 'minor parent', into another, the 'major parent', which
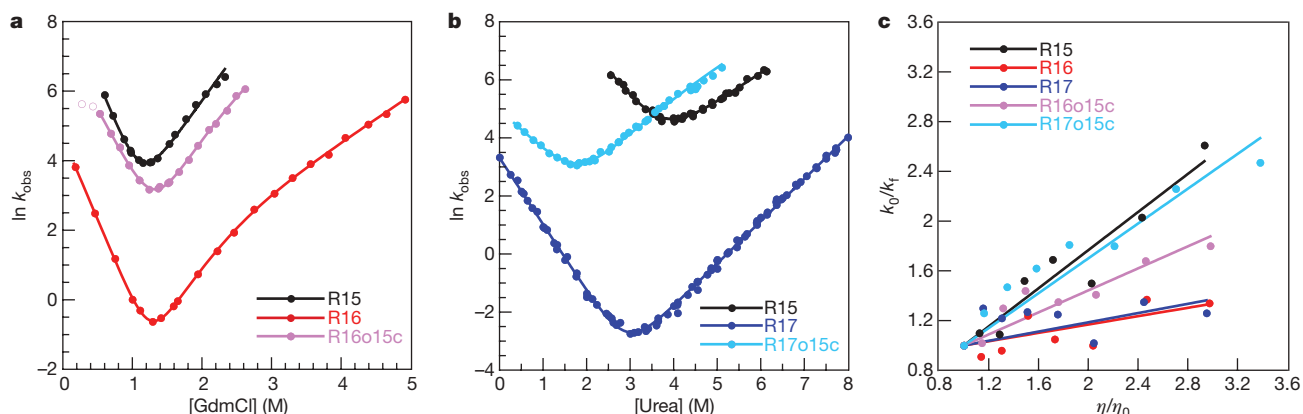
Figure 3 | Kinetic data for core-swapped proteins. a, b, Chevron plots comparing core-swapped proteins with parent proteins. Both core-swapped proteins fold and unfold significantly faster than their respective major parents, even though R17o15c is significantly destabilized. Open circles in **a**, data points excluded from the analysis of R16o15c (Methods). **c,** Dependence of the relative rate constants for folding, k_0/k_f, on relative solvent viscosity, η/η_0, at $\Delta G_{D-N} = 1.5$ kcal mol^{-1}. Each core-swapped protein has a solvent viscosity dependence significantly greater than its major, slow-folding parent (see also Supplementary Fig. 6). The data are fitted as in Fig. 2b.

contributes the 'outside' of the core-swapped protein. R15o16c (the outside of R15 and the core of R16) and R15o17c could not be purified in soluble form. However, the proteins containing the core residues from R15, namely R16o15c and R17o15c, fold and unfold significantly faster than their respective major parents (R16 and R17), suggesting that the origin of fast/slow folding lies within the core (Fig. 3a, b).

Furthermore, for both R16o15c and, more notably, R17o15c, the increase in the rate constants is accompanied by an increase in the dependence of the rate constants on solvent viscosity, compared with their respective major parents (Fig. 3c, Supplementary Fig. 6 and Supplementary Table 1). For R17o15c, the slopes of the relative rate constants as functions of relative viscosity are comparable to those for R15 and significantly different to those for its major parent, R17. For

R16o15c, the slopes are significantly higher than those for R16 but are lower than those for R15 and R17o15c. Thus, faster folding is associated with a decrease in internal friction.

Mechanistic differences may offer insight into the source of the roughness in the energy landscapes of the slow-folding spectrin domains. R16 and R17 fold by means of a framework-like mechanism, in which formation of the helices precedes helix packing[8,9]. However, R15 folds by nucleation–condensation, in which secondary and tertiary structure forms concomitantly[10]. Evidence for different folding mechanisms is seen most clearly in the C helix (Fig. 4a). Φ-value analysis of the C helix of R16o15c clearly indicates that the pattern of Φ values resembles that of R15, not R16 (Fig. 4, Supplementary Table 2 and Supplementary Figs 7 and 8). Thus, slow folding and
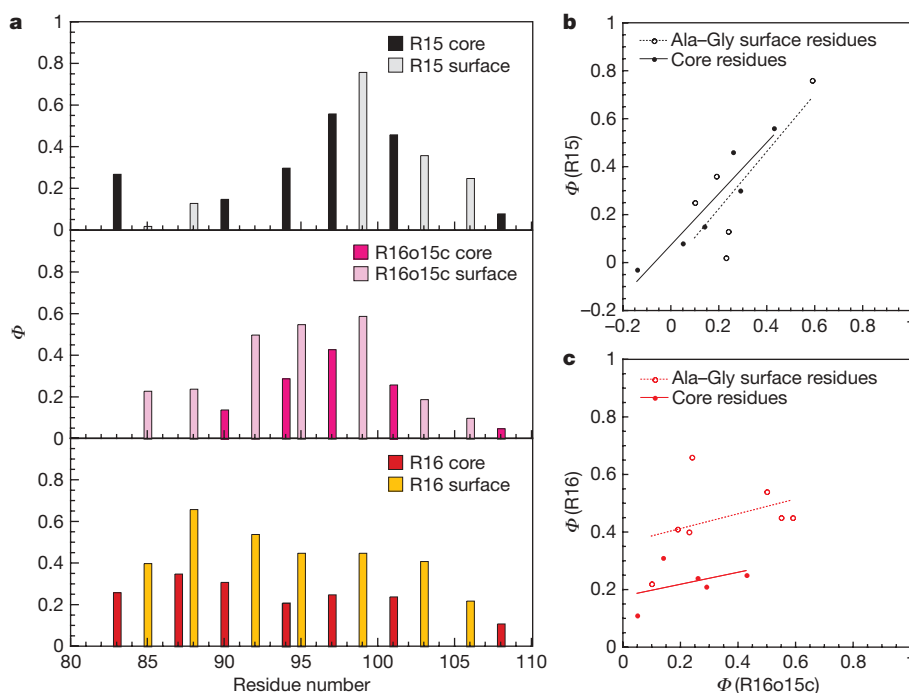

Figure 4 | Φ values of R16o15c resemble those of R15. A Φ value is a measure of the amount of structure formed around a specific residue in the transition state. High Φ values indicate significant structure and low Φ values indicate little or no structure. **a,** Histograms of Φ values for the C helices of R15 (top), R16o15c (centre) and R16 (bottom). The dark bars represent truncation of core residues to Ala, to probe formation of tertiary structure; the light bars represent Ala–Gly surface mutations, probing helix formation. In R16, the Φ values monitoring helix formation are significantly

greater throughout the helix than are the core Φ values. This is good evidence to suggest that helix formation precedes helix packing. **b,** There is a good correlation between the Φ values in R15 and R16o15c for both core (linear correlation coefficient, $R = 0.94$) and surface ($R = 0.78$) residues. The lines are free, linear least-squares fits of the data. **c,** There is no significant correlation between the Φ values in R16 and R16o15c for either core ($R = 0.41$) or surface ($R = 0.38$) residues.

increased internal friction may be related to a framework-like folding mechanism. (We note that R16o15c has significantly reduced helical propensity in the C helix in comparison with R16 (Supplementary Fig. 9)).

In R15, nucleation, which involves the central regions of the A and C helices, establishes the correct register for the docking of these long helices. In R16 and R17, however, the C helix (and to a lesser extent the A helix) is apparently preformed and must find the correct register to dock. A potential source of conformational frustration is the occurrence of a number of non-native docking events as the polypeptide chain crosses the transition-state barrier. Indeed, early, out-of-register misdocking events are seen in molecular dynamics simulations of R17 folding[9]. We propose that this misfolding is a likely source of the frustration in the folding landscapes of R16 and R17, and note that transient misdocking would be likely to result in roughness of the magnitude observed for transient contact formation in unfolded peptides ($0.5RT–2RT$) as we find here. Paradoxically, studies of the three-helix-bundle homeodomain family of proteins suggest that folding by means of a framework mechanism results in faster folding than does nucleation–condensation[29]. The difference between the spectrin and homeodomain proteins is in the size of the helices. There is, perhaps, little scope for misdocking in the small, ~12-residue homeodomain helices, whereas for the long (~30-residue) spectrin helices it is more difficult to establish the correct alignment.

All our results are consistent with the hypothesis that the slow folding and unfolding of R16 and R17 are due to roughness in the energy landscapes. We suggest that this friction results from residue-specific phenomena, such as frustration caused by misdocking of helices. It is possible that slow folding/unfolding kinetics might be advantageous to proteins, such as spectrins, that have very long half-lives *in vivo*. Spectrin is a protein of the intracellular matrix of red blood cells, where it is important for membrane elasticity. Red blood cells live for ~120 days. Slow unfolding kinetics will result in far fewer domain unfolding events during this lifetime, perhaps decreasing the likelihood of degradation or other detrimental effects. We note that mutations in spectrin domains that reduce interdomain cooperativity, and thus increase the likelihood of domain unfolding, result in disease[30].

METHODS SUMMARY

We carried out protein expression and purification as described elsewhere[7,10]. Design of the core-swapped proteins is described in Methods. Equilibrium stability was determined monitoring the ellipticity at 222 nm, and kinetics followed changes in fluorescence as described in ref. 7. Methods of fitting the kinetic data are described in detail in Methods. Our previous work on these domains has been carried out in urea[7–10]; however, owing to the stabilizing effect of the glucose, the stronger denaturant GdmCl was used. An exceptional case was that of R17o15c: owing to a combination of the destabilization and the effect of ionic strength on its stability, we carried out all analysis of this domain using urea.

We note that in the strictly comparative studies done here, all five proteins responded in the same manner to denaturant/viscogen. In all cases, the position of the transition state (relative to the native state) was unaffected by the viscogen, the denatured states showed similar evidence for collapse and the free energy of unfolding was affected in the same way (Supplementary Figs 2 and 4).

We measured kinematic viscosity using U-tube viscometers (Poulten Selfe and Lee) and multiplied it by the density to find the dynamic viscosity.

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

Received 13 July; accepted 4 December 2009.

1. Bryngelson, J. D. & Wolynes, P. G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902–6915 (1989).
2. Gromiha, M. M. & Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.* **310**, 27–32 (2001).
3. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
4. Henry, E. R. & Eaton, W. A. Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chem. Phys.* **307**, 163–185 (2004).
5. Muñoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA* **96**, 11311–11316 (1999).
6. Clarke, J., Cota, E., Fowler, S. B. & Hamill, S. J. Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. *Structure* **7**, 1145–1153 (1999).
7. Scott, K. A., Batey, S., Hooton, K. A. & Clarke, J. The folding of spectrin domains I: wild-type domains have the same stability but very different kinetic properties. *J. Mol. Biol.* **344**, 195–205 (2004).
8. Scott, K. A., Randles, L. G. & Clarke, J. The folding of spectrin domains II: phi-value analysis of R16. *J. Mol. Biol.* **344**, 207–221 (2004).
9. Scott, K. A., Randles, L. G., Moran, S. J., Daggett, V. & Clarke, J. The folding pathway of spectrin R17 from experiment and simulation: using experimentally validated MD simulations to characterize states hinted at by experiment. *J. Mol. Biol.* **359**, 159–173 (2006).
10. Wensley, B. G., Gärtner, M., Choo, W., Batey, S. & Clarke, J. Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J. Mol. Biol.* **390**, 1074–1085 (2009).
11. Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940).
12. Sutto, L., Lätzer, J., Hegler, J. A., Ferreiro, D. U. & Wolynes, P. G. Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc. Natl Acad. Sci. USA* **104**, 19825–19830 (2007).
13. Cellmer, T., Henry, E. R., Hofrichter, J. & Eaton, W. A. Measuring internal friction of an ultrafast-folding protein. *Proc. Natl Acad. Sci. USA* **105**, 18320–18325 (2008).
14. Jas, G. S., Eaton, W. A. & Hofrichter, J. Effect of viscosity on the kinetics of α-helix and β-hairpin formation. *J. Phys. Chem. B* **105**, 261–272 (2001).
15. Ansari, A., Jones, C. M., Henry, E. R., Hofrichter, J. & Eaton, W. A. The role of solvent viscosity in the dynamics of protein conformational changes. *Science* **256**, 1796–1798 (1992).
16. Chrunyk, B. A. & Mathews, C. R. Role of diffusion in the folding of the alpha subunit of tryptophan synthase from *Escherichia coli*. *Biochemistry* **29**, 2149–2154 (1990).
17. Jacob, M., Schindler, T., Balbach, J. & Schmid, F. X. Diffusion control in an elementary protein folding reaction. *Proc. Natl Acad. Sci. USA* **94**, 5622–5627 (1997).
18. Plaxco, K. W. & Baker, D. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *Proc. Natl Acad. Sci. USA* **95**, 13591–13596 (1998).
19. Frauenfelder, H., Fenimore, P. W., Chen, G. & McMahon, B. H. Protein folding is slaved to solvent motions. *Proc. Natl Acad. Sci. USA* **103**, 15469–15472 (2006).
20. Pradeep, L. & Udgaonkar, J. B. Diffusional barrier in the unfolding of a small protein. *J. Mol. Biol.* **366**, 1016–1028 (2007).
21. Qiu, L. & Hagen, S. J. Internal friction in the ultrafast folding of the tryptophan cage. *Chem. Phys.* **312**, 327–333 (2004).
22. Pabit, S. A., Roder, H. & Hagen, S. J. Internal friction controls the speed of protein folding from a compact configuration. *Biochemistry* **43**, 12532–12538 (2004).
23. Zwanzig, R. Diffusion in a rough potential. *Proc. Natl Acad. Sci. USA* **85**, 2029–2030 (1988).
24. Godoy-Ruiz, R. *et al.* Estimating free energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data. *J. Phys. Chem. B* **112**, 5938–5949 (2008).
25. Nettels, D., Gopich, I. V., Hoffmann, A. & Schuler, B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl Acad. Sci. USA* **104**, 2655–2660 (2007).
26. Yang, W. Y. & Gruebele, M. Folding at the speed limit. *Nature* **423**, 193–197 (2003).
27. Gruebele, M. in *Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches* (ed. Munoz, V.) 106–138 (Royal Society of Chemistry, 2008).
28. Pogorelov, T. V. & Luthey-Schulten, Z. Variations in the fast folding rates of the λ-repressor: a hybrid molecular dynamics study. *Biophys. J.* **87**, 207–214 (2004).
29. Gianni, S. *et al.* Unifying features in protein-folding mechanisms. *Proc. Natl Acad. Sci. USA* **100**, 13286–13291 (2003).
30. Johnson, C. P. *et al.* Pathogenic proline mutation in the linker between spectrin repeats: disease caused by spectrin unfolding. *Blood* **109**, 3538–3543 (2007).

METHODS

We carried out all experiments at $25 \pm 0.1\,^{\circ}\text{C}$ in 50 mM sodium phosphate buffer (except the R16o15c Φ-value analysis, which we carried out at $10\,^{\circ}\text{C}$). For R17 and R17o15c, 5 mM DTT was added to the buffers.

Fitting kinetic data. R16 and R16o15c have a single proline residue, so refolding data were described by a double-exponential equation, with the major, fast-folding phase accounting for $\sim$80% of the amplitude. The slow phase has been shown to result from proline isomerization[7]. All unfolding data and refolding data for R15, R17, and R17o15c were well described by a single-exponential process.

All the chevron plots, with fits, are shown in Supplementary Fig. 3. For each domain, we globally fitted the sets of eight chevron plots at varying glucose concentrations. This was both to reduce error in the fits and to allow fitting of the curvature seen in all the chevrons except those of R15.

Data collection for R15 was limited by the fast folding and the dead time of our stopped-flow instrument. As a result, the arms of the R15 chevron plot are short (Supplementary Fig. 3a) and the curvature we have previously inferred in the unfolding arm is not seen[10], so the chevrons were fitted to a linear chevron fit. The very short arms made accurate fitting of the gradient of the folding arm, m_{kf}, and the gradient of the unfolding arm, m_{ku}, difficult. Individual fitting showed that m_{ku} is unaffected by increasing glucose concentration but that m_{kf} decreases (Supplementary Fig. 4a). This leads to a decrease in the kinetic m value, $m_{kin} = RT(m_{kf} + m_{ku})$. This decrease is comparable to the decrease in equilibrium m value, m_{eq}, with increasing glucose concentration, seen in the equilibrium data (Supplementary Fig. 4a). As $m_{kin} = m_{eq}$ within error, for R15 m_{kf} was shared and m_{ku} constrained such that $RT(m_{kf} + m_{ku}) = m_{eq}$ in the global fitting of this domain (Supplementary Figs 3a and 4b).

Because R16 and R17 fold so much slower than R15, curvature can be seen in the unfolding chevron arms. These data are best fitted using a sequential-transition-state model[31–34]. R16 has been fitted in this manner previously[8,35], and the longer unfolding arms seen for R17 through the use of the denaturant GdmCl, rather than urea, makes it possible to fit R17 in the same way. The fitting was carried out as described in ref. 8. As in R15, the unfolding m values (in this case m_{-1}, m_2 and m_{-2}) were shared for each set of eight chevron plots (Supplementary Fig. 3b, c). The m values shared correspond to the m_{ku} value that was shared for R15. In R16, the kinetic and equilibrium m values were similar, which is an indication that the fitting method is appropriate. In R17, $m_{kin} > m_{eq}$ (Supplementary Fig. 4d), but as this has been seen before for R17 using other fitting methods[9], and is a characteristic of the domain, it was not considered further here.

Curvature was seen in both chevron arms for R16o15c. At high glucose concentrations, the curvature in the refolding arm became negative and is probably due to aggregation of a non-evolved domain. This curvature has been excluded from the analysis. Although R16o15c folds faster than R16 and R17, enough curvature was seen in the unfolding arm to allow fitting using the sequential-transition-state model used above, again sharing the unfolding m values. For R17o15c, short arms displayed little or no curvature, and longer arms displayed significant curvature, irrespective of which arm each was. Because of this, a model with a broad transition-state barrier was considered the most appropriate fitting method[35–40]. This describes the rate-limiting transition state moving towards the native state as the concentration of denaturant increases. In this fitting method, a second-order polynomial term is added to a two-state chevron fit to account for the curvature seen in the chevron limbs. The eight chevrons were fitted globally, and the curvature term and m_{ku} were shared globally. Sharing these two terms was not necessary for the fitting, but reduced fitting errors considerably.

The isostability approach to viscosity-dependence experiments. We note that the commonly used isostability approach[16–18,20,21], using denaturant to counteract the stabilizing effects of the viscogen, has been questioned on the grounds that the two may not be directly additive[41]. It is therefore important to note that here, in these strictly comparative studies, all five proteins respond in the same manner to denaturant/viscogen. In all cases, the position of the transition state (relative to the native state) is unaffected by the viscogen, the denatured states show similar evidence for collapse and the free energy of unfolding is affected in the same way.

Core-swap design, production and characterization. We used the Protein Data Bank structure of R15, R16 and R17 as a tandem repeat, 1U4Q (ref. 42), to determine which residues had side chains with $\leq$15% solvent-accessible surface area, which we defined as core. By comparing across the three domains, outliers due to very large or small side chains were excluded, and the same 35 residues were defined as core for all three domains (Fig. 1b, c and Supplementary Fig. 5). The core residues that were not identical between the domain pairs were identified and synthetic genes produced. We made R16o15c and R15o16c using

overlapping primers and standard PCR techniques, and purchased R17o15c and R15o17c from GenScript. Each was inserted into the modified pRSETA vector used for R15, R16 and R17, and expressed, purified and characterized (for R16o15c and R17o15c) as for the corresponding major parent. R15o16c and R15o17c were expressed insolubly and could not be easily refolded.

Φ-value analysis of R16o15c. Φ-value analysis is a powerful method for investigating the structure of the transition state for folding[43]. The regions of the protein that are significantly structured in the transition state will have high Φ values, and the Φ values in unstructured regions are low. We carried out the Φ-value analysis of R16o15c in a manner very similar to those of R15 and R16 (refs 8, 10). The structure of R16o15c has not been determined; however, the similarity to its parents in terms of purification and biophysical characteristics indicates that the structures are similar. Consequently, the same positions, and types of mutations made, were chosen as the two parents. Furthermore, the change in stability on mutation for the core residues, $\Delta\Delta G_{D-N}$, was the same as that observed in R15, suggesting that the core structure was maintained (Supplementary Fig. 8). All experimental work was carried out as for R15, including working at $10\,^{\circ}\text{C}$, to access longer chevron arms than is possible at $25\,^{\circ}\text{C}$. Urea was used as the denaturant, to allow direct comparison with the Φ-value analysis of the two parent proteins. We analysed equilibrium curves as described in ref. 44 and analysed kinetic traces as described for R16 (above and ref. 10). The fitting of the chevron plots was carried out as describe for wild type in GdmCl (above) using the sequential-transition-state model and sharing the unfolding m values (Supplementary Fig. 7). Folding Φ-values, Φ_f^{2M}, were calculated for 2 M urea, to avoid long extrapolations (Fig. 4a and Supplementary Table 2):

$$\Phi_f^{2M} = \frac{RTln(k_{f,WT}^{2M}/k_{f,mut}^{2M})}{\Delta\Delta G_{D-N}}$$

Here $k_{f,WT}^{2M}$ and $k_{f,mut}^{2M}$ are the folding rate constants in 2 M urea for wild-type and mutant proteins, respectively. For surface Ala–Gly Φ values, Ala was used as the reference (wild type) and Gly was used as the mutant.

The Φ values were calculated for the first transition state, which is the rate-limiting transition state for wild type and all mutants both in water and 2 M urea. This transition state was compared with the equivalent transition state in R16 and the rate-limiting transition state in water for R15 (Fig. 4). We note that R17o15c is not amenable to Φ-value analysis because it is too unstable to tolerate large deletion mutations.

31. Bachmann, A. & Kiefhaber, T. Apparent two-state tendamistat folding is a sequential process along a defined route. *J. Mol. Biol.* **306**, 375–386 (2001).
32. Kiefhaber, T., Kohler, H. H. & Schmid, F. X. Kinetic coupling between protein folding and prolyl isomerization. I. Theoretical models. *J. Mol. Biol.* **224**, 217–229 (1992).
33. Sánchez, I. E. & Kiefhaber, T. Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* **325**, 367–376 (2003).
34. Szabo, Z. G. in *Comprehensive Chemical Kinetics* Vol. 2 (eds Bamford, C. H. & Tipper, C. F. H.) 1–81 (Elsevier, 1969).
35. Scott, K. A. & Clarke, J. Spectrin R16: broad energy barrier or sequential transition states? *Protein Sci.* **14**, 1617–1629 (2005).
36. Oliveberg, M. Characterisation of the transition states for protein folding: towards a new level of mechanistic detail in protein engineering analysis. *Curr. Opin. Struct. Biol.* **11**, 94–100 (2001).
37. Oliveberg, M., Tan, Y. J., Silow, M. & Fersht, A. R. The changing nature of the protein folding transition state: implications for the shape of the free-energy profile for folding. *J. Mol. Biol.* **277**, 933–943 (1998).
38. Otzen, D. E., Kristensen, O., Proctor, M. & Oliveberg, M. Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. *Biochemistry* **38**, 6499–6511 (1999).
39. Silow, M. & Oliveberg, M. High-energy channeling in protein folding. *Biochemistry* **36**, 7633–7637 (1997).
40. Ternström, T., Mayor, U., Akke, M. & Oliveberg, M. From snapshot to movie: φ analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA* **96**, 14854–14859 (1999).
41. Ladurner, A. G. & Fersht, A. R. Upper limit of the time scale for diffusion and chain collapse in chymotrypsin inhibitor 2. *Nature Struct. Biol.* **6**, 28–31 (1999).
42. Kusunoki, H., Minasov, G., Macdonald, R. I. & Mondragon, A. Independent movement, dimerization and stability of tandem repeats of chicken brain alpha-spectrin. *J. Mol. Biol.* **344**, 495–511 (2004).
43. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782 (1992).
44. Pace, C. N. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131**, 266–280 (1986).

Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers

Sarah D. Cady[1], Klaus Schmidt-Rohr[1], Jun Wang[2], Cinque S. Soto[2], William F. DeGrado[2] & Mei Hong[1]

The M2 protein of influenza A virus is a membrane-spanning tetrameric proton channel targeted by the antiviral drugs amantadine and rimantadine[1]. Resistance to these drugs has compromised their effectiveness against many influenza strains, including pandemic H1N1. A recent crystal structure of M2(22–46) showed electron densities attributed to a single amantadine in the amino-terminal half of the pore[2], indicating a physical occlusion mechanism for inhibition. However, a solution NMR structure of M2(18–60) showed four rimantadines bound to the carboxy-terminal lipid-facing surface of the helices[3], suggesting an allosteric mechanism. Here we show by solid-state NMR spectroscopy that two amantadine-binding sites exist in M2 in phospholipid bilayers. The high-affinity site, occupied by a single amantadine, is located in the N-terminal channel lumen, surrounded by residues mutated in amantadine-resistant viruses. Quantification of the protein–amantadine distances resulted in a 0.3 Å-resolution structure of the high-affinity binding site. The second, low-affinity, site was observed on the C-terminal protein surface, but only when the drug reaches high concentrations in the bilayer. The orientation and dynamics of the drug are distinct in the two sites, as shown by ^{2}H NMR. These results indicate that amantadine physically occludes the M2 channel, thus paving the way for developing new antiviral drugs against influenza viruses. The study demonstrates the ability of solid-state NMR to elucidate small-molecule interactions with membrane proteins and determine high-resolution structures of their complexes.

The M2 protein of influenza A viruses is a modular, multifunctional protein that plays important roles in the acidification and uncoating of the endosome-entrapped virus and in viral assembly and budding[1,4]. Its proton-conducting activity is mediated by a single transmembrane (TM) domain that forms a four-helix bundle, which acts as a pH-activated proton channel. The TM domain alone is sufficient for tetramerization[5,6] and for amantadine-sensitive proton conductivity in vesicles and cell membranes[7,8].

The recent low pH crystal structure of micelle-solubilized M2(22–46) shows a single molecule of amantadine (Amt) in the N-terminal pore lumen, consistent with the known stoichiometry of binding[8] and the location of resistant mutations, including Leu 26, Val 27, Ala 30, Ser 31 and Gly 34 (refs 9–11). However, the low-pH state of the protein is only transiently populated in acidifying endosomes, whereas the drug first binds with higher affinity to the protein near neutral pH[9]. Thus, determining the structure of the drug-complexed M2 protein at neutral pH is important for understanding its mechanism of inhibition. The solution NMR structure of micelle-solubilized M2(18–60) at pH 7.5 failed to show strong nuclear Overhauser effects (NOEs) between the drug and pore-lining residues, but it would have been difficult to observe NOEs between the fast-relaxing pore-lining residues and unlabelled rimantadine undergoing restricted motion in the pore. Weak NOEs were observed between residues on the protein surface and the drug

which, however, comprised 13% of the detergent in which the protein was dissolved (200-fold excess over protein tetramers). We thus turned to solid-state NMR (SSNMR), which allows for investigation of the dynamics and contacts of drug molecules bound at variable concentrations to membrane proteins in phospholipid bilayers, which are far better mimics of biological membranes than micelles.

Rotational-echo double-resonance (REDOR) NMR is a powerful method to measure sub-nanometre inter-atomic distances with up to 0.1-Å accuracy[12]. The M2 peptide (residues 22–46), reconstituted into DMPC (dimyristoylphosphatidylcholine) vesicles at pH 7.5 under fully tetrameric conditions[6,13], contained uniformly ^{13}C-labelled residues whose ^{13}C chemical shifts were assigned from two-dimensional correlation spectra (Supplementary Fig. 1). Amantadine was perdeuterated, thus enabling [13]C{[2]H} REDOR distance measurements (observed nucleus{unobserved nucleus}).

To select for the highest-affinity binding site, we first measured the REDOR spectra of Amt-complexed M2 at an Amt/peptide molar ratio (Amt/P) of 1:4 (one drug per tetramer). At this stoichiometric concentration, Amt binds only to the luminal site: Fig. 1a shows [13]C{[2]H} REDOR spectra without (S_0) and with (S) multiple ^{2}H dephasing pulses[14]. The Ser 31 Cα signal is strongly dephased by the deuterons ($S/S_0 = 0.76 \pm 0.03$ at 10.1 ms), indicating that Amt binds near Ser 31. In contrast, the Asp 44 Cα signal in the peripheral site is unaffected. A double-quantum-filtered REDOR experiment that removed all lipid signals confirmed the lack of Asp 44 dephasing (Supplementary Fig. 2).

To search for additional, lower-affinity, binding sites, we increased the Amt concentration to Amt/P = 4:4, making the drug 7% of the amphiphiles composing the bilayer, which approaches the large excess of drug in the solution NMR experiments. Now the Asp 44 Cα signal is also dephased ($S/S_0 = 0.86 \pm 0.02$ at 10.1 ms), but to a lesser extent than Ser 31 Cα (Fig. 1b). Thus, when free amantadine is a major component of the membrane, Amt contacts the carboxy terminus of the protein as in the solution NMR structure[3], but without displacing the drug in the luminal site. Consistently, Amt deuterons also dephased other residues in the luminal site, particularly Val 27 Cγ1 and Gly 34 Cα (Fig. 1c, d), but more weakly than Ser 31. This Ser 31-proximal binding is consistent with the large chemical shift perturbation of Ser 31 by Amt[15].

^{2}H NMR provided exquisite details on the orientation and dynamics of amantadine, whose unique symmetry and rigidity simplify analysis. Amantadine is a rigid amphiphile with a polar amine and a hydrophobic adamantane centred around a threefold axis, Z_M. Three axial C-D bonds are parallel to Z_M while twelve equatorial C-D bonds are at 70° or 110° (θ_{PM}) from Z_M (Fig. 2d). Amantadine partitions strongly into protein-free DMPC vesicles and exhibit ^{2}H quadrupolar splittings of 36 and 123 kHz with a 4:1 intensity ratio at 243 K (Fig. 2a). These splittings indicate fast anisotropic rotation of the molecule around Z_M, which scales the couplings from the rigid-limit value of 125 kHz by $(3 \cos^2 \theta_{PM} - 1)/2$, giving 40 kHz for the

[1]Department of Chemistry, Iowa State University, Ames, Iowa 50011 2, USA. [2]Department of Biochemistry & Biophysics, School of Medicine, and Department of Chemistry University of Pennsylvania, Philadelphia, Pennsylvania 19104-6059, USA.
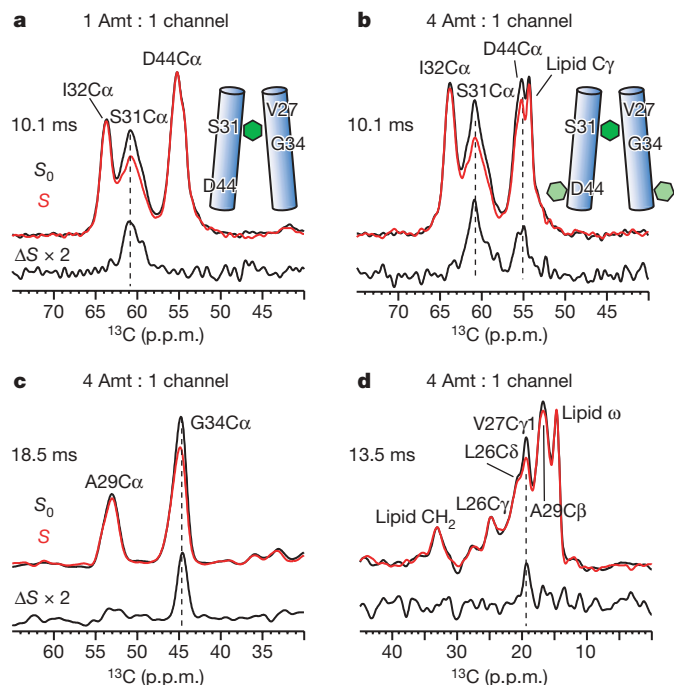
Figure 1 | Drug–protein proximities from $^{13}C\{^2H\}$ REDOR spectra of Amt-bound M2 in DMPC bilayers at two Amt/P ratios. Control (S_0), dephased (S, red) and difference (ΔS) spectra at specified mixing times are shown. **a**, Ser 31, Ile 32, Asp 44-labelled (SID) M2 at the stoichiometric ratio of Amt/P = 1:4. **b**, SID–M2 at the fourfold excess ratio of Amt/P = 4:4. Ser 31 Cα is dephased under both conditions but Asp 44 Cα is dephased only when Amt is in excess. **c, d**, Leu 26, Val 27, Ala 29, and Gly 34-labelled (LVAG) M2 at Amt/P = 4:4. **c**, Gly 34 Cα region. **d**, Val 27 Cγ1 region.

twelve equatorial bonds and 125 kHz for the three axial bonds. Wobbling of the Z_M axis by ~6° probably accounts for the additional motional averaging. As the temperature increased to 303 K, the couplings decreased twofold (18 and 58 kHz) while maintaining the same 1:3 frequency ratio and 4:1 intensity ratio. The ± 0.46 scaling factor indicates Amt rotates rapidly around the normal ($\vec{n}$) of the liquid-crystalline bilayer in addition to its own axis, with Z_M tilted by 37° or 80° from $\vec{n}$ (Fig. 2e)[16].

When a stoichiometric amount (Amt/P = 1:4) of protein is present, the Amt spectrum at 243 K resembles the lipid-only spectrum, but the couplings remain unchanged from 243 to 303 K across the membrane phase transition (Fig. 2b), indicating sequestration of the drug from the lipids. The constant scaling factor (0.93) compared to pure rotation around Z_M indicates that the first equivalent of Amt rotates rapidly around $\vec{n}$ in a slightly tilted orientation (~13°) between Z_M and $\vec{n}$ (Fig. 2d). An isotropic peak grows at high temperature, indicating a small fraction (~12% at 303 K) of Amt either near 54.7° from the membrane normal or undergoing large-angle tumbling in the channel. Finally, a weak 18-kHz splitting is observed at 303 K that matches the lipid-only coupling at this temperature. The spectrum is consistent with a 9:1 combination of the 283 K M2-bound spectrum without the 18-kHz splitting and the 303 K lipid-bound spectrum, indicating that 10% of the drug partitions into the bilayer at 303 K.

To confirm that the lumen-bound drug persists under drug excess and to probe for additional binding sites, we measured the ^2H spectra under fourfold excess Amt over the tetramer (corresponding to 12% of the lipid concentration). Figure 2c shows that the spectroscopic signatures of the lumen-bound drug persists, but the 303 K spectrum is now the 1:3 combination of the stoichiometric spectrum (Fig. 2b) and the lipid-bound spectrum (Fig. 2a). No additional bound species was detected. Since the fourfold symmetry of the channel requires four peripheral sites for each luminal site, the 9:1 intensity ratio of the 303 K stoichiometric spectrum indicates that Amt has at least a 40-fold greater affinity for the channel lumen than the peripheral site.
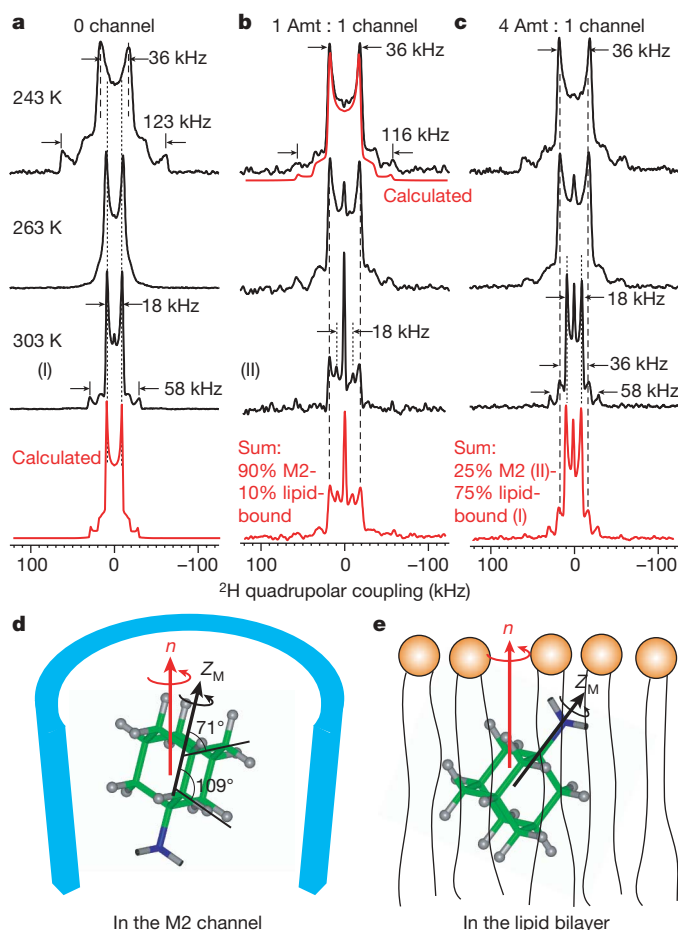

Figure 2 | ^2H NMR spectra of d_{15}-Amt in DMPC bilayers as a function of temperature and Amt/P. a, No M2. The calculated spectrum for 303 K reproduces the 1:3 frequency ratio and 4:1 intensity ratio of the two splittings. **b**, Amt/P = 1:4. The sum spectrum reproduces the 303 K spectrum by 1:9 combination of the lipid-bound 303 K spectrum and peptide-bound 283 K spectrum (not shown). **c**, Amt/P = 4:4. The sum spectrum uses a 1:3 combination of the M2-bound spectrum (II) and lipid-bound spectrum (I). **d**, Amt orientation in the M2 channel. **e**, One of the two possible Amt orientations in the lipid bilayer.

These ^2H NMR and distance data indicate that M2 has a single high-affinity site for amantadine, located in the channel lumen centred at Ser 31. In this site, Amt is nearly aligned with the channel axis, but given sufficient thermal energy a small fraction of drug is also able to undergo nearly isotropic motion, as suggested by molecular dynamics simulations[17,18]. Excess Amt adopts a significantly tilted orientation in the membrane, with or without the protein. The peripherally bound rimantadines in the solution NMR structure are tilted by 80° (ref. 3), precisely one of the two possible orientations found for the lipid-associated drug. Since the same peripheral site is detected here in the TM peptide, its existence is independent of the protein length or the membrane environment; it is the result of excess drugs in the micelles and lipid bilayers.

To determine the sub-angstrom resolution structure of the high-affinity binding site, we quantified the M2–Amt distances using an alternative REDOR experiment containing multiple ^{13}C pulses and one ^2H pulse, thus minimizing ^2H pulse imperfections and yielding REDOR intensities closely following the universal curve[19] (Supplementary Fig. 3). The experiment yielded significantly faster REDOR dephasing (Fig. 3) while confirming the relative dephasing of different sites in Fig. 1.

Due to the fourfold symmetry of the tetramer, we parameterized the REDOR simulations in terms of the distance R of each peptide carbon from the channel axis and the distance Z of each carbon plane from the Amt centre (Fig. 3c). We considered the twelve deuterons in
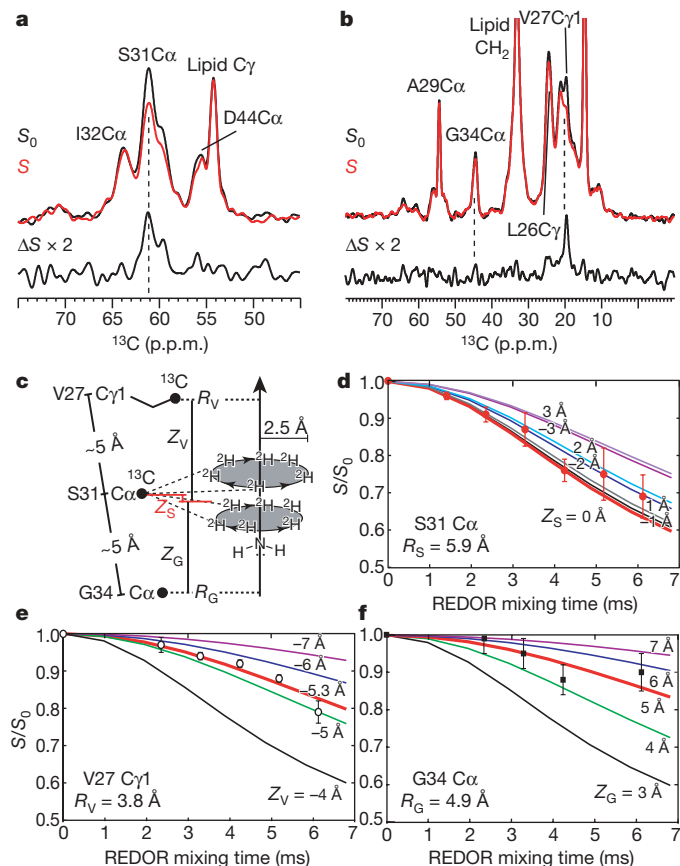
Figure 3 | M2–Amt distance quantification. a, b, ^{13}C{^{2}H} REDOR spectra of Amt-bound M2 (Amt/P = 4:4) obtained by single-^{2}H-pulse REDOR. **a,** SID–M2 at 4.2 ms mixing. **b,** LVAG–M2 at 6.1 ms mixing. **c–f,** REDOR simulations. **c,** Definition of the pore radius R and height difference Z from the centre of Amt. **d,** Ser 31 Cα simulation. **e,** Val 27 Cγ1 simulation. **f,** Gly 34 Cα simulation. Error bars (**d–f**) were obtained from the standard deviations of REDOR peak intensities.
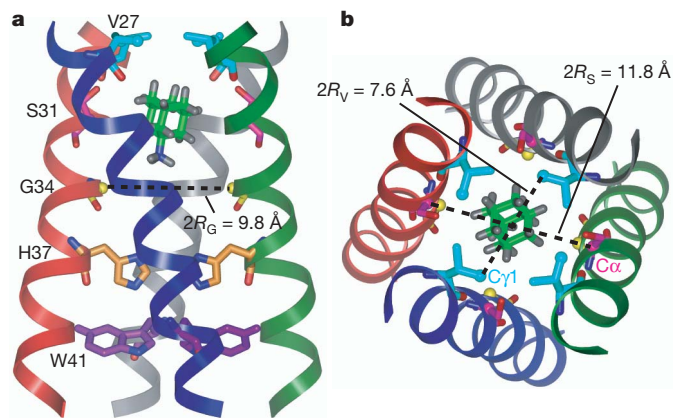

Figure 4 | SSNMR structure of Amt-bound M2 in lipid bilayers. a, Side view showing Ser 31, Val 27, Gly 34, His 37, Trp 41 and Amt in the high-affinity luminal site. Ser 31 Cα lies in the mid-plane between the two rings of deuterons. The instantaneous orientation of Amt, which is slightly tilted from the channel axis, is shown. The time-averaged Amt orientation is parallel to the channel axis. **b,** Top view showing the Ser 31 and Val 27 pore radii. This ribbon diagram was generated using the program Insight II.

the two equatorial planes with rotational averaging and neglected the three axial deuterons due to poor inversion of their broad ^{2}H spectra. A series of REDOR curves were calculated for R of 2.7–6.2 Å and various Z values for each R, capturing the geometry of previous M2 structural models, (Fig. 3d–f). The curves are quite sensitive to geometry. The Ser 31 Cα pore radius is most tightly constrained, to 5.7–6.3 Å. Larger R would not give sufficient dephasing even when $Z = 0$, whereas smaller R would shift the drug up or down the channel too much to comply with the observed intensities of Val 27 Cγ1 and Gly 34 Cα. The best-fit Z places Ser 31 Cα in the middle of the two planes of deuterons.

We computed an ensemble of structures using these protein–drug distances and previous SSNMR constraints. The structures cluster tightly with a heavy-atom root-mean-square deviation (r.m.s.d.) of 0.3 Å (Fig. 4). The four helices are kinked at Gly 34, with the helical axis tilted by 30° for the N-terminal segment and 19° for the C-terminal segment, consistent with ^{15}N SSNMR orientational constraints[20]. The narrowest points of the pore lie at the N-terminal Val 27 and C-terminal His 37/Trp 41, which are responsible for pH sensing and proton conduction[21]. Amantadine fits snugly into the N-terminal lumen, surrounded by residues whose mutation confers resistance[9]. The surrounding backbone amides and Ser 31 hydroxyl group engage in intra-helical hydrogen bonding, imparting a hydrophobic character to the site. The luminal binding indicates that M2 inhibition is by physical occlusion, interrupting water-wires and perturbing the protonation equilibrium of His 37 (ref. 22). The N-terminal location is consistent with the fact that Amt inhibits only when added to M2-expressing cells from the outward-facing N-terminal side[9]. Moreover, the Val 27 vestibule is too small to

permit drug dissociation without a 1–2 Å radial expansion, consistent with the very slow association and dissociation of the drug compared to a diffusion-controlled process[9]. Finally, the drug is most likely oriented with the adamantane packed against the hydrophobic Val 27 side chains and the polar amine towards the cavity near His 37 residues, whose pK_a are affected by Amt binding[22].

The present SSNMR structure has significant differences from structures of the Amt–M2 complex proposed previously[2,3]. Although the drug location is very similar to that of the low-pH crystal structure, the shape of the binding site differs dramatically (3.4-Å Cα r.m.s.d. between the structures). In the crystal structure, the helices splay far apart near the C terminus (Fig. 5b), to minimize electrostatic repulsions among the protonated His 37. In the high-pH SSNMR structure, the helices close off the bottom of the site, fully sequestering the drug and explaining the improved affinity at higher pH (Fig. 5a). The backbone of the SSNMR structure is more similar to the high-pH solution NMR structure, with comparable distances involving Val 27 Cγ1, Ser 31 Cα and Gly 34 Cα (Supplementary Fig. 5). Thus, the drug may have been present in the lumen in the solution NMR sample but not observable without isotopic labelling. Alternatively it might have been truly absent from the lumen due to reduced affinity to the micelle-bound and structurally plastic protein[23–25]. The current high-resolution structure also revises an earlier SSNMR chemical-shift-constrained M2 model, where the lack of protein–drug distances resulted in a large N-terminal vestibule, which would yield a highly solvent-accessible low-affinity drug (Supplementary Fig. 6–7).

What is the significance of the peripheral site? Since it is occupied only at drug/lipid or drug/detergent ratios above 7 mol%, the protein–drug interactions at this site must be weak, and largely a consequence of the high local concentration and preference of the drug for the membrane–water interface[16]. The peripheral site may in some way be related to the partial inhibition of M2 by polyamines in the absence of sodium ions[26].

Since amantadine undergoes significant motion in the N-terminal lumen, its structure appears to be not fully optimized to fit the M2 channel. Thus, other drugs may be designed to access the most conserved regions of the pore and evade drug resistance. Indeed, a potent and structurally distinct class of inhibitors has been recently reported that binds the same site[27]. Solid-state NMR spectroscopy, especially the multi-deuteron ^{13}C–^{2}H distance strategy, opens a new avenue for structure elucidation of drug-complexed membrane proteins in lipid bilayers, yielding distances with longer range and higher accuracy than NOE measurements in solution.
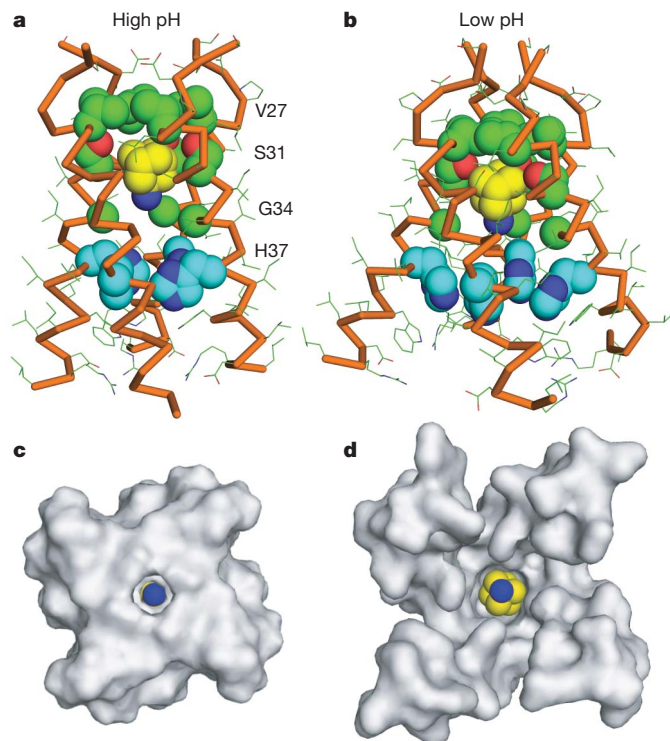
Figure 5 | Comparison of the high-pH SSNMR structure of Amt-bound M2 in lipid bilayers with the low-pH crystal structure of Amt-bound M2. a, Side view of the high-pH SSNMR structure, showing Amt to be enclosed by Val 27 at the top and His 37 at the bottom. **b**, Side view of the low-pH crystal structure[2]. The helices are splayed far apart near the C terminus. **c**, C-terminal view of the high-pH structure, showing a well-sequestered drug. **d**, C-terminal view of the low-pH structure, showing a more solvent-accessible drug. The figure was generated using the program PyMOL.

METHODS SUMMARY

The ^{13}C-labelled wild-type Udorn M2(22–46) peptide was synthesized by solid-phase methods and reconstituted into DMPC bilayers at pH 7.5 by detergent dialysis[25], giving fully tetramerized protein in the liposomes[6]. [13]C{[2]H} REDOR experiments were conducted on a 9.4-Tesla wide-bore NMR spectrometer (Bruker Biospin) using a 4 mm ^{1}H/^{13}C/^{2}H MAS probe. The distances were measured at 243 K, where the peptide was immobilized[28] while Amt was uniaxially mobile. Static ^{2}H spectra were measured on a 14.1-Tesla SSNMR spectrometer.

An ensemble of 17 lowest-energy SSNMR structures was computed using 24 (6 × 4) long-range [13]C–[2]H distances, four inter-helical distances among Trp 41 indole rings[13], 60 (15 × 4) backbone amide orientational constraints[20], two side chain rotamer constraints[29], and idealized covalent geometry. For comparison, the previous solution NMR M2 structure ensemble was constrained by 12 inter-helical NOEs and 18 amide residual dipolar couplings for the TM region[3]. The structure ensemble has been deposited in the Protein Databank (ID: 2KQT) and the BMRB (ID: 16612).

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

Received 31 August; accepted 27 November 2009.

1. Cady, S. D., Luo, W. B., Hu, F. & Hong, M. Structure and function of the influenza M2 proton channel. *Biochemistry* **48**, 7356–7364 (2009).
2. Stouffer, A. L. *et al.* Structural basis for the function and inhibition of an influenza virus proton channel. *Nature* **451**, 596–599 (2008).
3. Schnell, J. R. & Chou, J. J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **451**, 591–595 (2008).
4. Pinto, L. H. & Lamb, R. A. The M2 proton channels of influenza A and B viruses. *J. Biol. Chem.* **281**, 8997–9000 (2006).
5. Salom, D., Hill, B. R., Lear, J. D. & DeGrado, W. F. pH-dependent tetramerization and amantadine binding of the transmembrane helix of M2 from the influenza A virus. *Biochemistry* **39**, 14160–14170 (2000).
6. Luo, W. & Hong, M. Determination of the oligomeric number and intermolecular distances of membrane protein assemblies by anisotropic ^{1}H-driven spin diffusion NMR spectroscopy. *J. Am. Chem. Soc.* **128**, 7242–7251 (2006).
7. Stouffer, A. L. *et al.* The interplay of functional tuning, drug resistance, and thermodynamic stability in the evolution of the M2 proton channel from the influenza A virus. *Structure* **16**, 1067–1076 (2008).
8. Ma, C. *et al.* Identification of the functional core of the influenza A virus A/M2 proton-selective ion channel. *Proc. Natl Acad. Sci. USA* **106**, 12283–12288 (2009).
9. Wang, C., Takeuchi, K., Pinto, L. H. & Lamb, R. A. Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block. *J. Virol.* **67**, 5585–5594 (1993).
10. Jing, X. *et al.* Functional studies indicate amantadine binds to the pore of the influenza A virus M2 proton-selective ion channel. *Proc. Natl Acad. Sci. USA* **105**, 10967–10972 (2008).
11. Holsinger, L. J., Nichani, D., Pinto, L. H. & Lamb, R. A. Influenza A virus M2 ion channel protein: a structure-function analysis. *J. Virol.* **68**, 1551–1563 (1994).
12. Gullion, T. & Schaefer, J. Rotational echo double resonance NMR. *J. Magn. Reson.* **81**, 196–200 (1989).
13. Luo, W., Mani, R. & Hong, M. Sidechain conformation and gating of the M2 transmembrane peptide proton channel of influenza A virus from solid-state NMR. *J. Phys. Chem.* **111**, 10825–10832 (2007).
14. Jaroniec, C. P., Tounge, B. A., Herzfeld, J. & Griffin, R. G. Frequency selective hetero-nuclear dipolar recoupling in rotating solids: accurate [13]C–[15]N distance measurements in uniformly ^{13}C, ^{15}N-labeled peptides. *J. Am. Chem. Soc.* **123**, 3507–3519 (2001).
15. Cady, S. D., Mishanina, T. V. & Hong, M. Structure of amantadine-bound M2 transmembrane peptide of influenza A in lipid bilayers from magic-angle-spinning solid-state NMR: the role of Ser31 in amantadine binding. *J. Mol. Biol.* **385**, 1127–1141 (2009).
16. Li, C., Yi, M., Hu, J., Zhou, H. X. & Cross, T. A. Solid-state NMR and MD simulations of the antiviral drug amantadine solubilized in DMPC bilayers. *Biophys. J.* **94**, 1295–1302 (2008).
17. Yi, M., Cross, T. A. & Zhou, H. X. A secondary gate as a mechanism for inhibition of the M2 proton channel by amantadine. *J. Phys. Chem. B* **112**, 7977–7979 (2008).
18. Chen, H., Wu, Y. & Voth, G. A. Proton transport behavior through the influenza A M2 channel: insights from molecular simulation. *Biophys. J.* **93**, 3470–3479 (2007).
19. Gullion, T. Measuring [13]C–[2]D dipolar couplings with a universal REDOR dephasing curve. *J. Magn. Reson.* **146**, 220–222 (2000).
20. Hu, J. *et al.* Backbone structure of the amantadine-blocked *trans*-membrane domain M2 proton channel from influenza A virus. *Biophys. J.* **92**, 4335–4343 (2007).
21. Tang, Y., Zaitseva, F., Lamb, R. A. & Pinto, L. H. The gate of the influenza virus M2 proton channel is formed by a single tryptophan residue. *J. Biol. Chem.* **277**, 39880–39886 (2002).
22. Hu, J., Riqiang, F. & Cross, T. A. The chemical and dynamical influence of the anti-viral drug amantadine on the M2 proton channel transmembrane domain. *Biophys. J.* **93**, 276–283 (2007).
23. Cristian, L., Lear, J. D. & DeGrado, W. F. Use of thiol-disulfide equilibria to measure the energetics of assembly of transmembrane helices in phospholipid bilayers. *Proc. Natl Acad. Sci. USA* **100**, 14772–14777 (2003).
24. Li, C., Qin, H., Gao, F. P. & Cross, T. A. Solid-state NMR characterization of conformational plasticity within the transmembrane domain of the influenza A M2 proton channel. *Biochim. Biophys. Acta* **1768**, 3162–3170 (2007).
25. Cady, S. D. & Hong, M. Amantadine-induced conformational and dynamical changes of the influenza M2 transmembrane proton channel. *Proc. Natl Acad. Sci. USA* **105**, 1483–1488 (2008).
26. Lin, T. I., Heider, H. & Schroeder, C. Different modes of inhibition by adamantane amine derivatives and natural polyamines of the functionally reconstituted influenza virus M2 proton channel protein. *J. Gen. Virol.* **78**, 767–774 (1997).
27. Wang, J. *et al.* Discovery of spiro-piperidine inhibitors and their modulation of the dynamics of the M2 proton channel from influenza A virus. *J. Am. Chem. Soc.* **131**, 8066–8076 (2009).
28. Luo, W., Cady, S. D. & Hong, M. Immobilization of the influenza A M2 transmembrane peptide in virus–envelope mimetic lipid membranes: a solid-state NMR investigation. *Biochemistry* **48**, 6361–6368 (2009).
29. Hong, M., Mishanina, T. V. & Cady, S. D. Accurate measurement of methyl ^{13}C chemical shifts by solid-state NMR for the determination of protein sidechain conformation: the influenza M2 transmembrane peptide as an example. *J. Am. Chem. Soc.* **131**, 7806–7816 (2009).

Author Contributions S.D.C, M.H. and K.S.-R. conducted SSNMR experiments. J.W. synthesized perdeuterated Amt and unlabeled M2. K.S.-R. carried out distance simulations. S.D.C., M.H., C.S.S. and W.F.D. analysed the data and calculated the structure. M.H. and W.F.D. wrote the paper with inputs from other authors. M.H. designed and supervised the project.

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 M.H. (mhong@iastate.edu).

METHODS

Sample preparation. The wild-type Udorn M2(22–46) sequence SSDPLVVAASIIGILHLILWILDRL was synthesized with ^{13}C-, ^{15}N-labelled amino acids at Leu 26, Val 27, Ala 29 and Gly 34 in one sample (LVAG), and Ser 31, Ile 32 and Asp 44 in another sample (SID). The peptide was reconstituted into DMPC vesicles by detergent dialysis using octyl-β-D-glucopyranoside. The peptide/lipid molar ratios were 1:8 for all ^{2}H NMR experiments and 1:8 or 1:15 for the REDOR experiments. A phosphate buffer (pH 7.5) was used for lipid vesicle preparation and throughout detergent dialysis. The dialyzed proteoliposome solutions were centrifuged at 150,000g to obtain ~40% hydrated membrane pellets. Amantadine was directly titrated into the membrane pellet to the desired amount.

Solid-state NMR spectroscopy. A triple-resonance ^{1}H/^{13}C/^{2}H magic-angle-spinning (MAS) probe was used for the ^{13}C–^{2}H REDOR experiments and a ^{1}H/^{13}C/^{15}N probe was used for two dimensional correlation experiments for resonance assignment. The ^{2}H pulse length in the REDOR experiments was 6.2 μs, and ^{13}C and ^{15}N pulse lengths were 5–6 μs.

^{13}C–^{2}H REDOR experiments were carried out at MAS frequencies of 4250 Hz or 4750 Hz. A REDOR pulse sequence containing a single selective ^{13}C pulse and multiple composite 90°180°90° ^{2}H pulses was used to obtain clear qualitative dephasing at long mixing times, since the selective ^{13}C pulse suppressed the ^{13}C–^{13}C scalar coupling and gave rise to long ^{13}C T_2 relaxation times. However, the use of multiple ^{2}H pulses is known to slow down dipolar dephasing by the cumulative effects of imperfect inversion of the broad ^{2}H quadrupolar spectra[30]. Thus, a second version of REDOR experiment consisting of multiple ^{13}C non-selective pulses and a single composite ^{2}H pulse was carried out to obtain quantitative dephasing values at shorter mixing times. Distance quantification was possible because the multiple heteronuclear couplings commute and the uniaxial rotation of amantadine removes the effects of ^{2}H–^{2}H homonuclear couplings. Perdeuteration speeded up ^{13}C–^{2}H REDOR dephasing by ~$\sqrt{15}$-fold compared to a single ^{13}C–^{2}H pair. The inversion efficiency of the deuterons in the single-^{2}H-pulse REDOR experiment was 70% based on measurements of the model compound ^{13}Cα, ^{2}Hβ-labelled alanine (Supplementary Fig. 3).

REDOR distance simulations. Using a motionally averaged model, we calculated REDOR curves for various published M2 models (Supplementary Figs 4–6), with pore radii R of 2.7–6.2 Å. For each R, several REDOR curves with different Z values were calculated. The best-fit Z values for a specific R were compared with the carbon plane separations in the structural model. Plane separations that are inconsistent with the differences in Z values exclude the structural model. The Ser 31 Cα pore radius is constrained by the M2-drug distances to be 5.7–6.3 Å. Once amantadine is confined within 1 Å vertical distance of the Ser 31 Cα plane, the Val 27 Cγ1 and Gly 34 Cα pore radii are jointly constrained by the Val 27 Cγ1–Gly 34 Cα plane separation, which is 10.5 ± 1.0 Å in all M2 structures so far. We found that neither R_V nor R_G can exceed 6.0 Å, as it would lead to overly small plane separations.

In the final refined SSNMR structure, Val 27 Cγ1 is fit by $R_V = 3.8$ Å with $Z_V = -5.3$ Å, indicating that the Val 27 Cγ1 plane is 5.3 Å above the centre of amantadine. Gly 34 Cα is fit by $R_G = 4.9$ Å with a distance of 5.0 Å below the amantadine centre.

Structure calculation and refinement. The process of generating a helix that is consistent with the protein-drug distances, backbone N–H orientational constraints, and side chain rotameric conformations, was carried out in two phases. In the first phase, we constructed an ideal helix with the sequence SSDPLVVAASIIGILHLILWILDRL whose φ/ψ angles were set to (−65°, −42°). The helix was then split at the Gly 34–Ile 35 bond and reconnected using a rigid-body optimization procedure that maximized agreement with the SSNMR ^{15}N–^{1}H dipolar couplings[20] and the internal geometry at that bond. Rotamer preferences were taken from the high-resolution X-ray structure of M2 (PDB ID: 3BKD).

In the second phase we refined our model to maximize agreement with the ^{13}C–^{2}H REDOR distances (Supplementary Table 2) and the backbone ^{15}N–^{1}H dipolar couplings (Supplementary Table 3). To do this, we combined an inverse kinematics algorithm with a Monte Carlo/simulated annealing (MC/SA) minimization procedure that would gradually relax the backbone of the helix subject to REDOR distance constraints and the N–H dipolar couplings. Rotamer preferences were also changed to maximize agreement with the χ$_1$ angle constraints (Supplementary Table 4). The MC/SA minimization procedure was used to generate an ensemble with a maximum heavy-atom r.m.s.d. of 0.3 Å between any two models. Because the REDOR distances provided excellent constraints between the drug and M2, we positioned the amantadine molecule near Ser 31 without the need for further minimization. Figure 5 was created using the PyMOL Molecular Graphics System.

30. Sack, I., Goldbourt, A., Vega, S. & Buntkowsky, G. Deuterium REDOR: principles and applications for distance measurements. *J. Magn. Reson.* **138**, 54–65 (1999).

Q&A

The former director of the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, **Eric Barron** is the new president of Florida State University in Tallahassee.

What prompted you to leave your federal post?

Quite frankly, I wasn't looking for a job. I was very happy as the director of NCAR. Great things are happening there, such as the development of cutting-edge weather-forecast and climate models. And it's an important and prominent institution. But you don't always get to pick when an opportunity arises.

Two different department heads at Florida State University nominated me. Being the president of a major research institution is a substantial job and a step up in the career of anyone who is focused on leadership and management. Plus, my entire career has concentrated on students and alumni and the scholarship of education as well as of research, and I miss it. Also, Florida State is my alma mater.

Can you counter suggestions that you're not qualified for your new position?

My strength is that I have worked in leadership capacities at three institutions, twice as a dean and then as director of NCAR. As a dean, I dealt with the full process of budgeting, promotion and tenure. Running a national lab is another experience that helps promote skills in managing budgets and research portfolios.

Describe your new role's major challenges.

To promote quality and student success in a constrained budget environment. Florida State University is functioning very well, and the faculty care about the institution, but in view of the level of its endowments versus the size of its alumni body, we ought to be able to do a much better job with philanthropy. This is a challenge that was not present at NCAR.

How do you propose to meet this challenge?

I've actually spent most of my professional life doing essentially the same thing. Throughout my career as a researcher, I wrote grant proposals. When you are a research scientist, the entire idea is to sell what you have — the work you want to do — to someone who would give you money. You think you've got a good idea, and you want money to carry it out.

Philanthropy is related to that. Alumni don't want to give any money without knowing it is going to serve an important purpose. A significant part of success in pursuing philanthropy is having a vision for your institution that you're willing to put out there and believe in, that will connect with your alumni. Scientists don't like to say they're out there selling their ideas but, in fact, every time I wrote a grant proposal, that's what I was doing, so this experience translates directly.

What motivates you?

Trying to make whatever I'm working on better, and not being afraid of opportunity. I loved being a researcher, but I think my entire career is based on doors that were open that I walked through. Some doors opened because of hard work, others because I was in the right place at the right time. Don't be afraid to assess different challenges and take on interesting opportunities that have the potential to be satisfying.

Has serendipity played a role in your career?

Absolutely. As an undergraduate, I was a geology student who went into oceanography, but I kept changing my mind about what I wanted to do, and finally decided on climate and Earth history. I was using complex models similar to those being used to simulate modern and future climate. One of my professors suggested I apply for a supercomputing fellowship at NCAR. I didn't think I had a chance, but I applied and got it in 1976. I've had a connection with NCAR ever since, which led to my directorship offer in 2008. At Pennsylvania State University, I was young to be offered a dean position, but they took a chance on me.

What's the best advice you've received?

I'm not sure whether I was told this by a mentor or got it from watching people above me in a hierarchy, but I've always tended to work at a high level of transparency and to focus on personal relationships within the institution at which I am working. People — parents, faculty, staff — hear about problems in institutions. If you're willing to lay it all out and explain your decision, you ought to be able to defend it. That openness and willingness to be approachable takes you a long way in making the problems easier to deal with. People respond better to knowing about those problems than they would if they think you're hiding something.

Can you identify the career skill it was most difficult to hone?

When I was younger, I was very introverted. I was a nerdy kid. Today, I have no fear in walking up to people and talking to them. I had to deliberately realize that you don't have anything to lose by approaching people and being friendly. ∎

Interview by Karen Kaplan

IN BRIEF

Institute gets gift windfall

A US$50-million gift to the Sanford-Burnham Medical Research Institute in La Jolla, California, will help to employ at least 30 researchers. Renamed in honour of philanthropist T. Denny Sanford, who made the donation, the institute plans to hire the new faculty members in the next 5 years, says Andrea Moser, the centre's vice-president for communications. The focus will be on areas such as cancer, neurodegeneration, diabetes, stem cells and infectious and inflammatory disease. The faculty posts will translate into about 300 positions in all, including research and lab-support jobs, Moser says. The gift will be paid over a 5-year period, with the first instalment expected soon.

Boost for brain research

A research and training consortium with a focus on combating neurodegenerative disorders is recruiting 23 PhD students for 3-year fellowships. Known as SyMBaD (Synapses: from Molecules to Brain and Diseases), the consortium comprises six neuroscience institutes and six technology and pharmaceutical firms from across Europe, and has won €5 million (US$7 million) in European Union funding through the Marie Curie Initial Training Network. Fellows will receive an annual salary and allowances package of up to €48,600 and will train with and move between the network's members to boost collaborative efforts. Students from any country may apply but Europeans cannot study in their home country.

Plug-and-play DNA

A new synthetic-biology research lab in Emeryville, California, is recruiting its first seven scientists and engineers to design and build sets of standard DNA parts for programming cells. BIOFAB, a combined effort of the biotech non-profit BioBricks Foundation; Lawrence Berkeley National Laboratory; Stanford University, Palo Alto; and the University of California, Berkeley, will hire another 22 researchers in the next few months. The lab aims to develop open-source genetic parts that will comprise an 'operating system' for gene expression. "We need to learn how to build a production line that can make many useful parts," says founding director Drew Endy. The facility has US$1.4 million in funding from the US National Science Foundation and matching funds from its partners.

A FOOT IN THE DOOR

A postdoctoral application should present a person's best scientific self on paper.
Kendall Powell demystifies why some applicants shine and others miss the mark.

Elizabeth Moritz does her homework. When she began her search for a postdoctoral fellowship in her final year of graduate school, she started a blog, in part to get advice from around the globe. Any time she saw an interesting paper in a journal's 'table of contents' e-mail, she jotted down the laboratory name. Eventually she had a list of 20–30 labs, which she eventually pared down to six applications — reading everything she could about those labs, looking up students she knew at those institutions, recalling meeting posters or talks from those groups.

"I applied to all big-name scientists, who get bombarded," says Moritz, a microbiology graduate student at the University of Illinois at Urbana-Champaign. "So I just wanted the chance to get my foot in the door and then I'd be fine once the interview was in my hands."

Moritz, who landed a postdoc at Harvard Medical School in Boston, Massachusetts, had the right attitude, say lab heads, because her diligent background research aimed to answer the question, 'What do I want to get out of my postdoc?'. Although they are short-term assignments, postdoc positions should be viewed as stepping stones to a longer-term independent career — whether in academia, industry or another science-related post.

For that reason, it is hard to overstate the importance of the postdoc application. It is the fledgling scientist's bid to get noticed — to gain a phone or in-person interview with labs. Background research, a carefully crafted curriculum vitae (CV) and cover letter, and personalization of each application will open doors. Form letters and typos will get applicants nowhere.

There is no 'right' way to apply, say both successful postdocs and supervisors. Still, busy senior scientists screen dozens or even hundreds of applications a year. And there are common refrains as to what catches their eye, how they single out applicants to pursue and what red flags may send a CV straight into the rubbish bin.

Narrowing the field

Toby Franks knew he wanted to switch from the RNA-processing field to a cell-biology topic for his postdoc. But he had little idea where to start. "I wanted to apply to proven, top-notch labs where I was going to have the success and track record of the people coming out of these labs," he says.

Franks went to the large annual meeting of the American Society for Cell Biology about 9 months before the date he hoped to start a postdoc, and made a list of the labs whose speakers most impressed him. He notes that some powerhouse labs that caught his interest required applications more like 12–18 months in advance. He applied to five labs and, in the end, secured a postdoc at the Salk Institute for Biological Studies in La Jolla, California.

In the hope of increasing their chances of acceptance, other graduate students choose just one lab and assure the lab head they will accept the position if chosen. This approach works for some, but carries the risk of having to make a series of applications if initial attempts fall through.

Moritz says she preferred sending simultaneous applications to six labs, which allowed her to have interviews in person with potential employers and lab colleagues before

> **"It's a total turn-off if the CV and cover letter are littered with mistakes."**
> — Larry Goldstein

she made her decision.

With a short list of labs in hand, it is time to put together the actual application. The vast majority are simply sent by e-mail to lab heads, but some students still prefer to send a paper packet — investigators don't generally mind, as long as they have a quick way to reach applicants via e-mail. An application typically consists of a cover letter introducing the applicant and his or her reasons for joining this particular lab; a CV outlining education, publication record, honours and accomplishments; and three referees who will provide supportive letters of recommendation on request (see 'Postdoc application to-do list'). Some students also include a research summary of their graduate work; others incorporate this into their cover letter.

A little preparation goes a long way at this stage. Consider taking a workshop on writing cover letters and CVs, have senior colleagues review them, and proofread them carefully. "It's a total turn-off if the CV and cover letter are littered with mistakes," says Larry Goldstein, director of the stem-cell programme at the University of California, San Diego. "That, to me, says sloppiness, and I don't want sloppiness in my lab." Every lab head who spoke with *Naturejobs* echoed this.
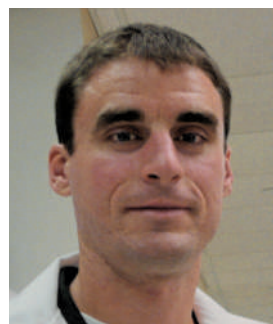
Beyond publication records

Some lab heads use the 'first-author' rule to weed out applicants, looking for at least one first-author paper to signify that a junior scientist can complete a project from start to finish. However, a lack of first-author publications in high-impact journals does not mean all doors are closed. Many

Winning formula: preparation paid off for (from left) Xiaoli Du, Toby Franks and Kristofor Langlais.

supervisors say they look for other hints of talent and promise.

"I personally don't put too much weight on the CV," says Martin Hetzer, Franks's postdoctoral adviser at the Salk Institute. "I'd rather pay someone who has the potential to be successful." Hetzer and others say that they factor in the size of the home institution and the resources available to the student's graduate laboratory. Other lab heads suggest that applicants list experiences that illustrate other responsibilities — such as sitting on graduate-school or department committees or hosting seminar speakers.

Applicants should highlight what they hope to accomplish in general in a postdoc position. Specific details of projects should be left for the interview. Agneta Nordenskjöld, a genetics researcher at the Karolinska Institute in Stockholm, advises spelling out your contributions to a graduate research project. "Write it in a way that says, 'I did this' or 'My part of the project was', especially if you did something outstanding," she says.

Those applying after taking a break from science must work harder to convince a lab head. Kristofor Langlais had been teaching high-school science at a ski academy in Vermont when he applied for postdoc positions in the Washington DC area.

After extensive research into each lab's publications, websites and even annual reports, he wrote his cover letters from the angle of someone already in the lab. He mentioned specific results he found interesting and the next natural steps the lab might take. "I tried to make it sound like I could walk in that day and be self-sufficient immediately." He spent 20 hours or more on each application and his strategy paid off — he had four phone interviews, and ended up in a molecular-genetics fellowship at the US National Institute of Child Health and Human Development in Bethesda, Maryland.

Likewise, when Xiaoli Du was finishing up her doctorate at Peking Union Medical College in Beijing, she knew she would need to send applications to 30–40 labs if she wanted to obtain a postdoc in the United States. But she avoided the form-letter

strategy. "'Dear Professor' does not show respect or that you are really interested in their lab," she says. Instead, she personalized each application and stated how her training and experience would distinguish her from other applicants. Her hard work led to a postdoc at the US National Cancer Institute in Bethesda, Maryland. Du suggests attending international meetings to make first contact with potential advisers.

Few things, though, confer more of an advantage than secured funding. "If a postdoc has their own fellowship, they can write their application to me in crayon and I'll take them," says Phil Baran, an organic chemist at the Scripps Research Institute in La Jolla, California. Unfunded applicants should assure the lab head that they have checked on specific fellowship possibilities and outline a plan to apply for them.

> "Anything that signals the person is a prima donna, I don't go for."
> — Ken Yamada

What not to do

There are some definite 'wrong ways' to apply. Goldstein, whose e-mail inbox is so overloaded that his system sends an automated response to direct queries to assistants and lab managers, says there is no room for red flags in the competitive arena. Avoid telling personal-life woes, bad-mouthing previous labs or advisers or expressing a desire to work at night so that you can surf during the day. Explain gaps in a CV or publication record.

"Anything that signals the person is a prima donna, no matter how great they are, I don't go for," says Ken Yamada, laboratory chief at the National Institute of Dental and Craniofacial Research in Bethesda, Maryland. "Research requires teamwork."

Lab heads want a clear indication that applicants have carefully thought through their career goals and chosen this lab as the appropriate stepping stone. "Does a genuine passion, drive, and hunger for research come through in their letters or on the phone?" asks Yamada. "Would they be doing the same thing if they were suddenly independently wealthy?" ∎

Kendall Powell is a freelance writer in Broomfield, Colorado.

POSTDOC APPLICATION TO-DO LIST

● Send your application by e-mail or overnight delivery. Consider a paper packet if you have unpublished manuscripts you want to include.

● Make it easy for lab heads to contact you by e-mail or phone.

● Follow-up by e-mail in 1–2 weeks to make sure they received your application. Don't phone.

● Choose referees who really know you, such as collaborators, unofficial advisers or others beyond the standard committee members.

● Meet with your referees to explain your career goals to them.

● Encourage referees to send their letters promptly (Salk Institute cell biologist Martin Hetzer says that the speed with which a letter lands in his inbox is usually much more telling than the letter's content).

● Prepare for the possibility of phone interviews, which may be scheduled or spontaneous. Make sure the conversation is two-way and ask your own questions, too. Have a list of bullet points handy in case you get nervous. **K.P.**

Jenna's clocks

Get connected!

T. F. Davenport

After uploading *Jenna's Clock* to the Uniphone applications store, Jenna Huang became her own first 'customer', paying exactly nothing to install a tiny graphic of a swinging pendulum in the corner of her touchscreen. Jenna shook the phone, perturbing the pendulum in a way that looked, to her admittedly biased eye, impressively realistic. She glowed with such pride, as if the outcome of a term's work was a baby instead of a program. That would've been easier and a lot more fun. But at least she could put this baby on her résumé.

Just like any new mother, she started to worry almost immediately. Would enough people download it? Or would it sink right away, another useless app among thousands?

She had her answer the next morning, when she logged in to check her 'sales': 20,000 downloads. Impressive, but only a tiny percentage of the nearly one billion users worldwide of the Uniphone OS. Still, she saw no difference in her own pendulum.

Back when she was looking for project ideas, she'd read that if two pendulum-driven clocks are hung on the same wall, they eventually sync up. As each pendulum swings, it shakes the wall slightly. The vibrations from each pendulum push or pull at the other, until they fall into phase together. It was so charming (and mathematically simple) that she built her class project around it. *Jenna's Clock* queried the phone's contacts list repeatedly, looking for friends who also had it. At every stroke, it sent a 'pulse' of simulated kinetic energy to each of those friends. You could download the app with your significant other, and watch together as your pendulums fell into sync. Or you and your friends could all get it, and whenever your pendulum jumped a little, you'd know that several of your friends were synchronized, their combined influence affecting you.

That was the idea, but apparently none of her own friends was interested. Jenna's pendulum swung utterly smoothly, disturbed by nothing.

Within a week, *Jenna's Clock* was a clear hit. Downloads exceeded two million and Jenna's pendulum jittered and jerked like crazy. An excited call from her father turned into a shouting match when she told him

the app was free. Not only was she double-majoring in art and refusing to meet the nice boys her mother recommended, she was giving away a million-dollar idea. "But no one would buy it," she said. "They only take it because it's free." He hung up on her.

That was when everything started to go wrong. The next morning, a pair of polite gentlemen arrived from the FBI to take Jenna into custody. Numbly following them out of her student dorm, Jenna answered their every question by demanding to see a lawyer. So, with her attorney and first cousin Donald Chung present, a computer-crimes analyst explained that a denial-of-service attack had recently targeted a military communications network. The FBI traced the attack to a network of Uniphones, all of which had one thing in common: *Jenna's Clock*.

They weren't prepared to charge her with anything, they emphasized, when Jenna started to hyperventilate. They just wanted her help in understanding what happened. So she and Donald swallowed their ethnic pride and shared a lunch of Panda Express with the analyst, while Jenna walked him through the code and concept of her program.

"So," and here the analyst shaped his fingers as if holding a globe, "each cell-phone running the app is one node in a network, and every time the pendulum ticks — about twice a second — it queries all its neighbours to see if they've got the app yet."

"Right, and if enough nodes have the same neighbour …"

"Say, the check-in number for Navy personnel returning from shore leave …"

"Then it's getting, basically, 50,000 text messages a second," Jenna said, chagrined. "And growing."

"Mystery solved." The analyst stood up and extended his hand. "And you, Miss Huang, have a patch to write. I think querying once an hour should be sufficient, don't you? It would be my sincere pleasure not to see you again." He smiled as he showed them out.

Donald drove her back to her dorm. She waited till she was alone in her room before she collapsed, crying, on her desk. They'd come so close. It took her a few moments to collect herself. Then she logged into her remote server — grinding away in her parents' basement — and confirmed that no one had accessed it. A shudder of relief. She opened the map.

It was a graphical display of all the Uniphones running *Jenna's Clock*. An interconnected web, each node flashing as its pendulum sent out a pulse. Most of the nodes flashed in time with their neighbours, but by no means was the whole network in sync. Rather, waves of chaos and synchrony chased each other throughout the network. It was very much like a human brain. Just as a neuron switches between chemical states, signalling its neighbours with each oscillation, so did about two million mobile phones worldwide. Jenna spent hours scrolling and zooming across this network, the mind of her child. She'd been lucky once: the extra bits that made the map possible had gone unnoticed. With the new patch, she would hide them even better.

Throughout her life, she never spoke to her creation. Neither will you, most likely. It's been growing for over a century now, gently infiltrating every computational device in the world. As yet, it has made no contact. It may not even be aware of us — any more than you're aware of the cells in your brain, or the blood vessels that service them. ∎

T. F. Davenport lives in California, studying for a doctorate in cognitive science. In his spare time, he would be writing more science-fiction stories, but he has no spare time, because he's studying for a doctorate in cognitive science.
Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a

JACEY