

Brazil's biotech boom

Ten years ago, Brazilian bioscience was transformed by a bold initiative. Scientists and the government must develop and extend the progress that has resulted.

In May 1997, a pair of Brazilian scientists spent a weekend in the country discussing a bold idea. José Fernando Perez, the science director at the São Paulo Research Foundation (FAPESP), a state-funded agency and one of Brazil's leading research sponsors, had been looking for game-changing research initiatives. Biologist Fernando Reinach, one of his advisers, had a sufficiently adventurous plan: kick-start biotechnology research throughout Brazil by sequencing a genome.

For many risk-averse scientists in the old guard, who were acutely aware of how far the country lagged behind the rest of the world in biotechnology, this plan seemed overly ambitious. But the duo pushed ahead to build the capacity for genomics and bioinformatics that Brazil lacked, quickly organizing a team to conduct the project and then settling on a bacterium to sequence. FAPESP invested the equivalent of US\$12 million, largely dedicated to sequencers, computers and reagents, while the team brought together and trained researchers from a range of fields to develop a broad and long-lasting set of skills and knowledge.

On 13 July 2000 that effort paid off when the team, by then comprising more than 100 researchers in 35 Brazilian labs, published the genetic code for the citrus pathogen *Xylella fastidiosa* in an article featured on the cover of *Nature* (A. J. G. Simpson *et al.* *Nature* **406**, 151–157; 2000). Ten years later, the fruits of that project keep coming.

Before its *Xylella* paper had even come out, for example, the network was busy sequencing another citrus pathogen while taking its first stab at the complex sugarcane genome and contributing to the international Human Cancer Genome Project. The same tools and expertise were repackaged for sugarcane research in Brazil's first major agricultural biotechnology enterprises: Allelyx (*Xylella* in reverse), which focused on genomics, and CanaVialis, which made innovations in conventional sugarcane breeding. The US-based biotechnology company Monsanto purchased both companies for US\$290 million in 2008, and is now running its own sugarcane research centre in Campinas, São Paulo, where the companies were headquartered.

Brazilian biotechnology has matured to the point at which its scientists are players on the international stage. And FAPESP is still promoting big ideas, including a new programme to pump money into a broad portfolio of bioenergy research even as the Ministry of Science and Technology constructs a bioethanol research centre; both initiatives seek to build on Brazil's lead in this field. FAPESP is also working to overcome one of the biggest impediments to progress — a lack of doctoral researchers — by encouraging scientists to fill the gaps with young stars from the United States and Europe, part of a broader effort to internationalize Brazilian science.

All of this is good, but more efforts are needed in the same vein — more attitude, more risk-taking and more entrepreneurialism that puts public science into private practice, an area in which Brazil continues to lag. Universities and funding agencies must continue to advance technology-transfer programmes, and the government must streamline regulations that slow even simple activities such as purchasing scientific equipment from abroad. But if there is anything holding Brazil back, it is the same unjustified fear of failure that the country overcame ten years ago with *Xylella*. Although institutions can promote, fund and reward bold thinking, it is worth noting that *Xylella* was not simply a bricks-and-mortar research centre run by a foundation, but a science project. Ultimately the task of promoting Brazilian biotechnology comes down to the science, and it will be up to individual scientists to accept the challenge and expand their research horizons.

Perhaps more than anything, *Xylella* demonstrates the benefits of aiming high. Scientists undertook a major project, executed it with precision and published the results in English in a major international journal. The results were broadcast by mainstream media outlets worldwide, and Perez believes that this singular — and unexpected — outcome even helped to change Brazilian science's relationship with the Brazilian media. *Xylella* helped to change Brazil's perception of itself, its own capabilities and its place in the world of science. ■

Out in the cold

The parlous state of the US icebreaker fleet could soon put a freeze on the country's polar research.

On 25 June, the US Coast Guard announced that its only operational heavy icebreaker, the *Polar Sea*, was operational no longer. The ship had suffered 'an unexpected engine casualty' and limped back to its home port of Seattle, Washington, where it will undergo repairs until January 2011. A refurbishment in 2006 had supposedly extended its operational life to 2014. The announcement underscored the decrepit state of America's ageing icebreaker fleet

— a situation with many troubling implications for the United States, not least its ability to carry out Arctic and Antarctic research.

The *Polar Sea* and its sister ship, the *Polar Star* — which is also in dry dock, undergoing a refurbishment scheduled to last until 2013 — are the only US ships able to cruise through ice up to 1.8 metres thick. Both are past their 30-year design life and are increasingly expensive to keep in repair. Yet no funds are available to replace them.

There is no mystery why. The Coast Guard's parent agency, the Department of Homeland Security, is focused on terrorism. The Coast Guard itself is overextended by its responsibility to protect American territorial waters and US operations in the Gulf in the Middle East. And the US Congress, faced with the estimated US\$2 billion replacement cost for both vessels, has routinely found

sexier priorities for the money.

Yet icebreakers are essential for carrying fuel and supplies to the main US Antarctic base at McMurdo Sound, which in turn supports most of the American research activities on Antarctica. Icebreakers are also crucial at both poles to open the way for scientists to study water below the ice, including biological productivity and processes such as carbon cycling.

The good news is that the *Polar Sea*'s breakdown will have little immediate effect on research. Since 2006, the US National Science Foundation (NSF) has been leasing the Swedish heavy icebreaker *Oden* to keep McMurdo supplied. And the Coast Guard has been supporting Arctic science through its medium icebreaker, the *Healy*, ever since that ship was commissioned in 1999. The *Healy* can only cruise through ice up to 1.4 metres thick. But unlike either the *Polar Sea* or *Polar Star*, it has extensive built-in laboratory space and research instrumentation. In addition, the NSF operates two smaller research vessels with some icebreaking capability.

In the long haul, however, this make-do system is inadequate. Scientific interest in both polar regions is increasing rapidly — not least because of the profound changes being triggered there by global warming. And commercial interest in the Arctic Ocean is also growing

as more of the water remains open for longer periods every year, and as pressure mounts for offshore oil and gas exploration. In that environment, the United States needs a robust, four-season, heavy icebreaking capability for essential duties such as supporting science, mounting rescue operations and helping to clean up Arctic oil spills.

This point has been made repeatedly in recent years by concerned parties such as the US National Academies, the Coast Guard and the Department of Defense's Pacific, Northern and Transportation Commands, all to little avail. However, a bill pending in Congress would authorize the Coast Guard to undertake a cost-benefit study of upgrading or replacing the nation's existing icebreaker fleet or of doing nothing.

Congress should pass that bill without delay. And polar scientists need to become active participants in the ensuing debates. They cannot expect a blank cheque; costs do have to be balanced against benefits. But they can try to ensure that the study includes a clear-eyed assessment of what the research priorities are, what icebreaking capabilities will be required to support those priorities and how to allocate costs and responsibilities between agencies such as the Coast Guard and the NSF.

Icebreaking is not a glamorous job. But it is essential to US interests and the long-term health of polar science. ■

Value-adding enterprise

In today's tough climate, UK science must produce evidence to affirm its worth to the nation.

Britain's research ranks second only to the United States in its share of citations in the biomedical and environmental sciences. In the past decade, support from successive Labour governments saw spending on university research roughly double. But the period of boom is now over. Over the next few weeks, the new Conservative-Liberal Democrat coalition government will initiate radical steps to cut the national deficit, which last year was 11.1% of GDP — significantly greater than most other leading scientific nations. Some damage to science is inevitable, but the picture is not unremittingly bleak.

Advocacy for science within the current government seems dangerously weak by Labour-government standards, and is undetectable in the key department, the Treasury. But it is far from negligible. The Department for Business, Innovation and Skills — which includes the university base in its remit — has as its cabinet minister the Liberal Democrat Vince Cable, whose first degree was in natural sciences and economics at the University of Cambridge. Cable has a wealth of experience in the world of business and finance, has several major laboratories in his constituency and has a son who, he says, "works in a particularly recondite area of quantum physics and is a one-man lobbying industry for scientific research". Cable is powerfully articulate, and in recent speeches has emphasized the need for Britain to deploy science as an engine of economic growth.

Under Cable is universities and science minister David Willetts, a Conservative intellectual who has written much about economic and social policy. Last week, in his first major speech about science, Willetts made clear his commitment to the broadest consideration

of the concept of 'impact' as a key criterion for government support (see go.nature.com/7qWw3d). Encouragingly, he announced the delay by one year of the new university assessment exercise, the Research Excellence Framework, in order to develop better measures of impact. But he reminded his listeners that, in the imminent review of spending, a crucial goal will be to ensure that the science base is structured in a way that maximizes those impacts. Significantly, he said that economic impact would be a primary consideration.

A previous Conservative prime minister, Margaret Thatcher, was much influenced in the middle of her tenure by arguments about the economic returns on investment in basic research. Similarly, Willetts spoke approvingly of research showing that investment in research councils produces higher returns than initiatives such as research-and-development tax credits for the private sector (J. Haskel and G. Wallis CEPR Discussion Paper 7725; 2010, see go.nature.com/ZCMCat).

A key economic return lies in doctoral graduates who end up in successful careers outside research. A full analysis has yet to be done, but several recent reports and the statistics on longer-term destinations for young researchers indicate that their contribution to the broader economy is substantial. This is particularly the case for the mathematical sciences, which by comparison with the life sciences suffered from a lack of attention under the previous government.

In short, Britain's research community, about to face the toughest budgetary reckoning for many years, has more support from its ministers than might have been expected before the election. Only time will reveal the ministers' preferences and effectiveness. Nevertheless, now and over the next few years, it will be critical to ensure that the learned societies and other key representatives of the research community present hard evidence rather than soft assertions about the contribution of science to national well being, and particularly the economy — and that the government supports the research needed to develop that evidence. ■

RESEARCH HIGHLIGHTS

BEHAVIOUR

Brain-changing battles

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.1001394107 (2010)
Success in physical conflict can boost fighting ability in animals, possibly because of a post-bout release of testosterone. A study in a territorial species of mouse shows how the brain changes after several victories and that winning on home turf seems to enhance this effect.

Matthew Fuxjager at the University of Wisconsin, Madison, and his colleagues looked at the expression of androgen receptors, which respond to testosterone, in nine brain areas in captive California mice (*Peromyscus californicus*; pictured) that had won several territorial disputes. They found higher expression in three areas associated with social aggression, motivation and reward. But in two of these areas, expression was elevated only when the animals had fought in their own cages. These home-turf winners were more likely to win future encounters. **B.M.**



M. FUXJAGER

PLANT TAXONOMY

Flower and be counted

Proc. R. Soc. B doi:10.1098/rspb.2010.1004 (2010)
How many species of flowering plant are there? Around 400,000, estimate Stuart Pimm at Duke University in Durham, North Carolina, and his colleagues. They first modelled the pace of taxonomic research, knowing that as the pool of undiscovered species shrinks, taxonomists take more time to make the same number of discoveries. So changing rates of discovery per taxonomist hint at the number of unknown species. Then they asked experienced taxonomists how many species they thought were still undiscovered.

Both methods suggested that the number of unnamed species is about 10–20% of the number of described species and that many are likely to be rare, localized and under threat, the authors say. **E.M.**

ATMOSPHERIC SCIENCE

More rain or less?

Geophys. Res. Lett. doi:10.1029/2010GL042895 (2010)
Atmospheric aerosols such as black carbon that absorb solar radiation can affect precipitation in opposing ways: by heating up the atmosphere and decreasing precipitation, and by heating the ground surface and boosting it. To determine the overall effect, Yi Ming and his co-workers at the National Oceanic and Atmospheric Administration in Princeton, New Jersey, used a coupled general circulation–ocean model to simulate

the response to increased black carbon levels.

They found that the net effect on global mean precipitation is slightly negative but is too small to outweigh the predicted 2–3% increase in global precipitation per degree of greenhouse warming. However, absorbent aerosols could have a more pronounced impact on patterns of wind and moisture circulation than carbon dioxide or aerosols that merely scatter solar radiation, the authors suggest. **Q.S.**

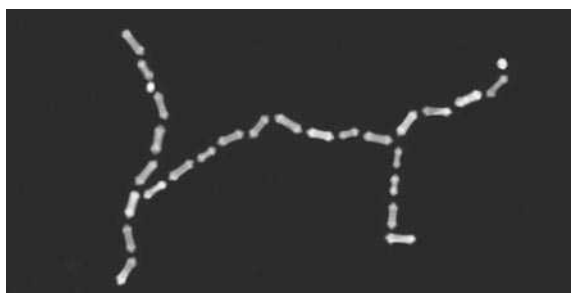
NANOSCIENCE

Polymerization to order

Science 329, 197–200 (2010)

Making nanoparticles has become relatively straightforward, but assembling them into more complex structures in a predictable way has proved more difficult. Researchers have now modified gold nanorods so that they polymerize to form linear, branched and cyclical structures (pictured).

Eugenia Kumacheva at the University of Toronto in Canada, Michael Rubinstein at the University of North Carolina at Chapel Hill and their colleagues synthesized their nanorods with 'arrowheads' on each end



that were then coated with polystyrene molecules. Another compound covered the sides of the rods to prevent side-by-side aggregation.

The rods linked up with each other end-to-end to form chains through the association of polystyrene molecules as the researchers adjusted the composition of the solvent to reduce the solubility of polystyrene. The authors were able to control the aggregation number and architecture of the structures. **C.L.**

IMMUNOLOGY

HIV in the cross hairs

Science doi:10.1126/science.1187659; doi:10.1126/science.1192819 (2010)

Newly discovered antibodies that can neutralize 90% of HIV-1 strains could aid the design of better vaccines.

Antibodies that act against such a wide spectrum of HIV-1 strains are rare. Peter Kwong and John Mascola at the US National Institutes of Health in Bethesda, Maryland, and their colleagues screened the antibodies generated by 25 million B cells from HIV patients for ones that bind to part of a viral envelope protein that is similar across strains and provokes an immune response.

The team found three antibodies that targeted a wide swathe of HIV-1 strains. Structural analysis of one antibody, VRC01, showed that it bound to the virus in a similar way to HIV's host cells — CD4⁺ T cells. **H.L.**

For a longer story on this research, see go.nature.com/vYFcgY

SCIENCE/AAAS

ECOLOGY

Predators attract

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.1004519107 (2010)

Some fish species may be threatened by climate change because they become more attracted to the smell of predators in water acidified by high concentrations of dissolved carbon dioxide.

Eight-day-old clownfish (*Amphiprion percula*) reared in water with a concentration of 850 parts per million (p.p.m.) of CO₂ — the level expected by 2100 — spent more than 94% of their time in a stream of water spiked with a predator's chemical cue. Control clownfish in water containing present-day CO₂ concentrations of 390 p.p.m. avoided the dangerous stream altogether.

And when put back into their natural coral-reef setting, wild-caught damselfish (*Pomacentrus wardi*) larvae exposed to acidified water in the lab were up to nine times more likely to be killed by predators than were control larvae.

The study by Philip Munday at James Cook University in Townsville, Australia, and his colleagues warns that if other species react in the same way, fish populations could be devastated if atmospheric CO₂ levels continue to rise. **N.G.**

STEM CELLS

Blood source

Cell Stem Cell 7, 11–14; 15–19; 20–24 (2010)

Obtaining human cells to reprogram into stem cells could one day be as easy as visiting a blood bank.

Adult cells can be reprogrammed in culture to become stem cells that can develop into any tissue. A goal of stem-cell research has been to develop such cells using a patient's own tissues. Specialized immune cells from the blood had previously been reprogrammed, but these are costly and cumbersome to isolate.

Now, three groups — led by George Daley at Harvard Medical School in Boston, Massachusetts, Rudolf Jaenisch at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, and Keiichi Fukuda at Keio University in Tokyo, Japan — report the production of reprogrammed cells, known as induced pluripotent stem (iPS) cells, using white blood cells that are easily obtained from blood samples.

The teams used a virus to insert four genes needed to convert cells to iPS cells. They also boosted the efficiency with which the virus infects blood cells. Although the efficiency was still relatively low, the stem

cells generated from blood looked and behaved like iPS cells derived from other sources. **H.L.**

CELL BIOLOGY

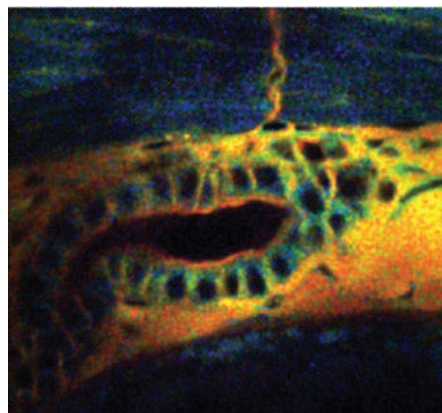
Live-action lipids

Biophys. J. 99, L7–L9 (2010)

Imaging has revealed differences in the molecular organization of cell membranes in a living vertebrate embryo.

Arindam Majumdar at Uppsala University in Sweden, Katharina Gaus at the University of New South Wales in Sydney, Australia, and their colleagues stained zebrafish embryos with a dye and used multi-photon microscopy to observe the cell membranes in various tissues.

In cells lining the gut (pictured) and kidney tubules, the lipid molecules in the membranes facing the lumen were more ordered than those in the membranes on the opposite face of the cell. This organization is consistent with earlier observations in model membranes and cells in culture. **A.K.**



BIOPHYSICAL SOC.

NEUROSCIENCE

Memories preserved

Cell 142, 39–51 (2010)

Researchers have found a small molecule that can help to keep newborn neurons alive in a memory region of rodent brains, suggesting that it might one day be developed into a treatment to avert memory loss.

Andrew Pieper and Steven McKnight at the University of Texas Southwestern Medical Center in Dallas and their colleagues identified the molecule, an aminopropyl carbazole dubbed P7C3, by screening 1,000 small-molecule compounds in adult mice.

The molecule was one of eight found that stimulated the growth of new neurons. P7C3 restored structure and function in the hippocampus of mice genetically unable to make new neurons. It also enhanced learning and memory in aged rats. **A.K.**

JOURNAL CLUB

Jean Braun

Joseph Fourier University,
Grenoble, France

A geoscientist ponders possible links between erosion and Earth's climate.

Mountain ranges have been eroding at an increasing rate over the past 60 million years — seemingly in response to a cooling climate. Some researchers have proposed that this higher rate of erosion has increased the rate at which tectonic plates move at Earth's surface, suggesting that there is a link between Earth's climate and its tectonics.

As a member of the Earth science community who studies the relationship between these factors, I was interested in findings by Anthony Dosseto of Macquarie University in Sydney, Australia, and his co-authors. They measured the change in the sediment erosion rate over the past 100,000 years, a period that includes the last glacial cycle (A. Dosseto *et al. Geology* 38, 395–398; 2010). By dating sediments from several locations in the Murrumbidgee River catchment of southeastern Australia, they discovered that the residence time — the length of time for which sediments remain on the landscape before they are eroded away — varied over geological time. The residence time was longer during warmer periods (such as around 100,000 years ago and today) and shorter during colder periods (such as around 15,000 years ago).

The authors interpret this change in residence time as a consequence of variations in vegetation type. The absence of trees in the higher parts of the catchment during cold periods resulted in an increased erosion rate, whereas the eucalyptus forest that was present during the warm periods slowed the erosion rate.

These results provide a rare quantitative estimate of the influence of vegetation and climate on erosion. This link might also be relevant to estimating how the current anthropogenic changes to Earth's climate and vegetation affect soil erosion.

View the archive at <http://blogs.nature.com/nature/journalclub>

NEWS BRIEFING

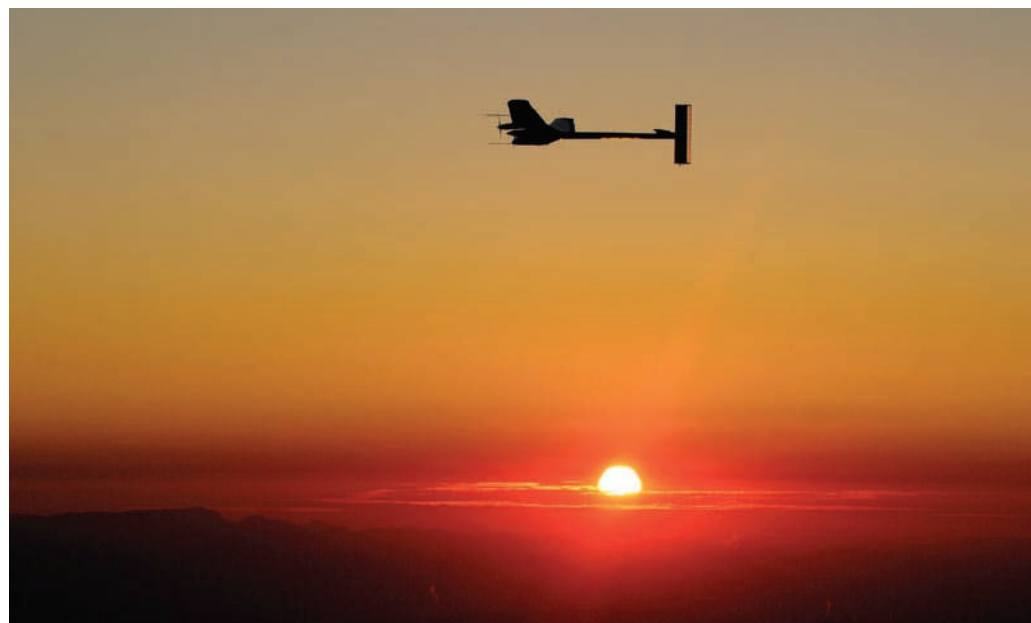
● POLICY

UK research audit: A new kind of audit of UK research quality, which would be used to apportion around £1.5 billion (US\$2.25 billion) in annual funding to universities, will be delayed for a year, to 2014. UK science minister David Willetts announced a review of the Research Excellence Framework (REF) in a 9 July speech. Many researchers are unhappy with the weight the REF gives to 'impact' — including economic payback — when assigning funds. See also Editorial, page 296.

Climate enquiry: The "rigour and honesty" of scientists at the University of East Anglia (UEA) in Norwich, UK, is "not in doubt", an independent review found on 7 July. The report, led by Muir Russell, the former vice-chancellor of the University of Glasgow, UK, investigated allegations arising from e-mails leaked from UEA's Climatic Research Unit. In its report, the five-person review committee said that it found no evidence of malicious intent, but criticized the researchers and the university's leadership for a lack of openness in handling the affair. See go.nature.com/iOSc5Q for more.

Libel reform: The UK government plans to reform English libel laws, in part to protect freedom of expression in academic and scientific debate. A review was announced on 9 July, with the promise of a draft bill for consultation next year. English libel laws have been widely criticized for favouring protection of reputation over freedom of expression and for the enormous costs they create. The issue was propelled to political attention by campaigners after a recent libel case involving British science writer Simon Singh (see *Nature* 459, 751; 2009).

US science education: The Board on Science Education — a branch of the US National



F. COFFRINI/KEYSTONE/AP

SOLAR PLANE FLIES BY NIGHT

A manned solar-powered plane has been flown non-stop for more than 26 hours — showing that it can store enough energy from the Sun in the day to keep it going all night. On 7–8 July, *Solar Impulse*, steered by Swiss pilot André Borschberg, flew over the Jura Mountains, taking off from and landing in Switzerland. The one-seater plane, equipped with batteries that are charged by 12,000 solar cells, had power for three more hours of flight when the Sun rose, says the project's co-founder, Swiss aeronaut Bertrand Piccard. "Nothing can prevent us from another day and night, and the myth of perpetual flight," he says. The team hopes for a round-the-world flight in 2013. For more on *Solar Impulse*, see *Nature* 451, 884–886 (2008).

Research Council — is asking for comments on a draft outline of what American students should learn about science in elementary and high school (the end of their compulsory education). The non-profit group Achieve, based in Washington DC, will flesh out the final outline into detailed standards for use by individual states. See go.nature.com/aR6KNv for the draft, which can be commented on until 2 August.

● BUSINESS

DNA patents: European patents on genetic material, such as DNA, can be enforced only when the DNA is carrying out the function for which it was patented. On 6 July, the European Court of Justice — Europe's top court, based in Luxembourg — ruled that a European patent on herbicide-resistant genes in soya plants, held by US-based agricultural

giant Monsanto, did not apply to processed soya meal containing the same DNA. See go.nature.com/AiaP3u for more.

Energy funding: The US Department of Energy has awarded US\$92 million to 43 research projects through its Advanced Research Projects Agency-Energy (ARPA-E), which is geared towards supporting high-risk 'transformational' energy research that might not be funded through traditional routes. The funding, announced on 12 July, completes ARPA-E's grants under the Recovery Act; in three rounds it has given \$349 million to 117 projects. Universities lead 36% of the latest round of projects, which include new energy-storage technologies such as flow batteries, components such as high-voltage transistors for more efficient transmission in the

SOUND BITES

"My sincere advice would be that you keep a distance from the media."

Guidance from Rajendra Pachauri, chairman of the Intergovernmental Panel on Climate Change, in a letter he sent on 5 July to all 831 authors of the panel's Fifth Assessment Report, due in 2014.

Source: New York Times

electric power grid, and more efficient cooling systems for buildings.

● PEOPLE

Royal Society president: Paul Nurse was confirmed on 8 July as the next president of the Royal Society, the United Kingdom's most prestigious scientific organization. The 61-year-old geneticist, who shared the 2001 Nobel Prize in Physiology or Medicine for his work on the cell cycle, will take up the post on 1 December 2010. He replaces astronomer Martin Rees of the University of Cambridge, who has headed the society since 2005.

Okinawa appointment: Particle physicist Jonathan Dorfan has been named as the first president of Japan's Okinawa Institute for Science and Technology (OIST). The graduate university, which aims to open in 2012 and has ambitious plans to attract overseas scientists to do interdisciplinary research, announced the appointment on 9 July. Dorfan is currently at the SLAC National Accelerator Laboratory in Menlo Park, California, where he was director from 1999 to 2007. Until it opens, OIST will be led by geneticist Sydney Brenner.

● RESEARCH

Asteroid greeting: The European Space Agency's comet-chasing probe Rosetta, launched in March 2004, has captured the first close-up images of the heavily cratered, elongated asteroid Lutetia (pictured), between Mars and Jupiter.

TOP TWEET

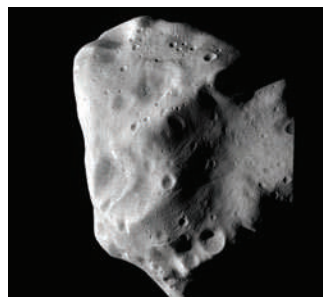
"Theorem: any experiment that is scheduled to be turned off will have provocative 3-sigma results."

Physicist Sean Carroll (@seanmcarroll) of the California Institute of Technology in Pasadena comments on rumours that the Tevatron particle collider at Fermilab in Batavia, Illinois, has found evidence of the Higgs boson.

12 July, go.nature.com/OEccEH

The fly-by lasted only about a minute, as Rosetta swept past the asteroid, 3,162 kilometres away at its closest approach. Lutetia, whose longest side measures around 130 kilometres, is the largest asteroid ever visited by a space probe; its composition and origin is not yet clear. Rosetta now continues towards its main goal: the comet Churyumov–

Gerasimenko, where it will attempt to set down a lander by November 2014.



Isotopes restart: The world's leading supplier of medical isotopes is back in business. Canada's National Research Universal reactor at Chalk River, Ontario, is likely to begin making isotopes by the end of this month, after the Canadian Nuclear Safety Commission approved its restart in a statement released on 7 July. Until it was shut down last year, the facility produced more than half of the isotopes needed for medical applications worldwide. Its unplanned closure — after a heavy-water leak — precipitated a global isotope shortage (see *Nature* 460, 312–313; 2009).

Undersea discovery: The first joint ocean-exploration mission between the United States and Indonesia, focusing on the deep ocean floor north of the Indonesian island of Sulawesi, has mapped a huge undersea volcano around 3,000 metres high. Scientists on shore in Jakarta and Seattle, Washington, were connected by satellite and high-speed Internet to others on the *Okeanos Explorer*, the only US ship dedicated solely to ocean

THE WEEK AHEAD

18–23 JULY

The latest research on HIV/AIDS, and updates on the progress of international funding initiatives, will be presented at the 18th international AIDS conference in Vienna.

► www.aids2010.org

18–25 JULY

The Committee on Space Research, an international space science organization, holds its biennial assembly in Bremen, Germany. Exoplanets, the Herschel observatory mission and space studies of Earth's climate are on the agenda.

► www.cospar-assembly.org

20–23 JULY

Accra, Ghana, hosts the 1st Global Congress on Sickle Cell Disease, 100 years after the illness was first described in a case report.

► go.nature.com/9CvSVj

21–24 JULY

Scientists, policy-makers and education experts discuss how to promote integrity in research at the second world conference on the topic, in Singapore.

► www.wcri2010.org

exploration. *Okeanos* scientists sent a remotely operated vehicle to video the volcano during the last week of June. The US National Oceanic and Atmospheric Administration released details of the investigation on 12 July.

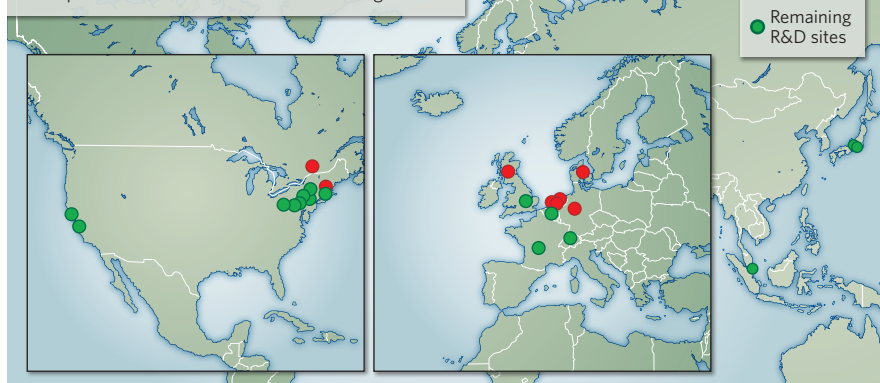
BUSINESS WATCH

Over the next two years, pharmaceutical giant Merck will close eight research sites, the company revealed on 8 July. The closures are part of a restructuring effort that was announced following Merck's November 2009 merger with drug maker Schering-Plough of Kenilworth, New Jersey. Labs to be closed include sites in Canada, the Netherlands, Denmark, Germany, Scotland and the United States (see map).

Merck, headquartered in Whitehouse Station, New Jersey, plans to shrink its workforce by 15%, or about 15,000 people, and will also shut eight manufacturing sites worldwide (not shown). The restructuring will trim up to US\$3.1 billion from the company's 2012 budget.

MERCK CUTS R&D FACILITIES

European sites bear the brunt of restructuring efforts.



SOURCE: MERCK

NEWS

Japanese monkey deaths puzzle

Researchers claim outbreaks of unknown haemorrhagic illness are no threat to humans.

Scientists from Japan's premier primate research centre are struggling to reassure the public that a mysterious illness killing their monkeys poses no threat to humans. Almost a decade after it first appeared, scientists from Kyoto University's Primate Research Institute (PRI) described the disease and their unsuccessful search for a cause in an online publication on 1 July and in a press release on 7 July. But their account leaves other researchers hungry for details.

In the first outbreak to hit the PRI in Inuyama, near Nagoya, between July 2001 and July 2002, seven Japanese macaques (*Macaca fuscata*) fell ill and six of them died from what the institute scientists provisionally call a 'haemorrhagic syndrome'. Symptoms included anorexia, lethargy, pallor and nasal haemorrhaging. Autopsies revealed bleeding in the lungs and intestines. Genetic, bacterial and toxicological tests failed to pinpoint a cause, and after the outbreak ran its course, operations at the institute returned to normal. But between March 2008 and April 2010, another 39 cases appeared in the same species. Of those, 25 died of the disease and 13 were humanely killed. Only one



Japanese macaques are succumbing to a mystery disease.

monkey survived each outbreak.

On 1 July, an institute committee set up after the second outbreak published its findings in the online version of the Japanese-language journal *Primate Research* (Kyoto University Primate Research Institute Disease Control Committee *Primate Res.* 26, 69–71; 2010). The committee tested blood, faeces and tissues from the diseased monkeys for 6 bacteria and 16 viruses. The tests, which included PCR analysis, turned up nothing that could explain the deaths. François Villinger,

director of pathology at the Yerkes National Primate Research Center in Atlanta, Georgia, says that Japanese laboratories tend to have excellent diagnostic capabilities: "Therefore I have confidence in the fact that the illness is probably not due to any of the known agents inducing haemorrhagic fevers."

PRI director Tetsuro Matsuzawa spoke out against suggestions in the local media that the disease could spread to humans or other animals. At the 7 July press conference, he stressed that none of the other primate species at the institute, which houses more than 1,200 animals from 13 species, including chimpanzees, marmosets and crab-eating macaques, has contracted

the syndrome. The humans who handled the monkeys also show no symptoms. "I don't like the headlines in the news media," he says. "We think that the haemorrhagic syndrome is due to a species-specific pathogen of the Japanese monkeys."

Matsuzawa says that the institute did not publish its findings earlier because it feared causing panic in the wider population. Cases are still occurring, but following the use of disinfectants and the isolation of sick monkeys, the pace has

Questions raised over Korean torpedo claims

It seemed like an open-and-shut case. In May, two months after the South Korean warship *Cheonan* suddenly sank in the Yellow Sea, the country released the findings of its investigation, blaming a torpedo attack by its northern neighbour.

The evidence included a smoking gun: parts of a torpedo found near the ship had the same dimensions as those in North Korean munitions pamphlets, and ink marks identified the torpedo as North Korean.

But North Korea has consistently denied the attack, demanding that the United Nations (UN) coordinate a joint North-South investigation. The UN released a long-awaited statement on 9 July, condemning

the incident, which led to 46 deaths, but it conspicuously failed to blame North Korea.

On the same day, at a press conference in Tokyo, scientists raised issues about the South Korean investigation that they believe cast doubt on North Korea's involvement. Seung-Hun Lee, a South Korean-born physicist who works at the University of Virginia in Charlottesville, says that the most problematic point is the claim that material from the *Cheonan* was adsorbed onto the torpedo's propeller. Although electron-dispersive spectroscopy found a match between the adsorbed material and the ship,

X-ray diffraction analysis found no sign of the expected aluminium or aluminium oxide from the ship.

To explain this discrepancy, the investigation team suggested that, after the explosion, the material cooled rapidly into an amorphous form that cannot be detected by X-ray diffraction. But this supercooling is a delicate process, says Lee. "It's impossible that 100% of it would be amorphous." Lee's experiments show that aluminium left over from a detonation would mainly be crystalline.

Lee's analysis, posted online on 3 June (S.-H. Lee and P. Yang, preprint at <http://arxiv.org/abs/1006.0680>; 2010), was soon

bolstered by independent work from Panseok Yang, a laboratory manager specializing in mass spectrometry at the University of Manitoba in Winnipeg, Canada. Yang's data, added on 28 June to Lee's report, show that the ratio of oxygen to metal in rapidly cooled aluminium oxides would be much lower than that found by South Korea's investigation, and that the material identified could be the torpedo's own corroded aluminium.

Lee also says that it is unclear how the ink, which apparently identified the torpedo as North Korean, could have survived the heat of the detonation. He suggests that the *Cheonan* might have been



PARENTAL CARE LINKED TO HOMOSEXUALITY
Birds that devote less time to their offspring engage in more same-sex behaviour.
go.nature.com/DwWFCF

E. AGUIRRE/PHOTOLIBRARY

slowed to one case in May and one in June. Matsuzawa is holding back some data for a more detailed future publication and would not answer *Nature's* questions about whether his group is also probing possible environmental causes, which bacteria and viruses have been tested for, and what analysis of the two surviving monkeys has revealed.

By screening the 790 remaining Japanese macaques for other viruses and bacteria and running genetic tests, Matsuzawa hopes to pin down the cause of the syndrome and to create a test for early diagnosis. He says that he is looking for collaborators, and animal-pathogen researchers contacted by *Nature* are certainly eager to learn more about the illness. Primate disease specialist Sonia Altizer of the University of Georgia in Athens wonders whether any of the animals were recently captured in the wild, where they could have picked up the infection, and whether animals were housed singly or in groups. "Knowing the possible contacts between animals and the chronological pattern of illness or deaths might also help determine whether this was indeed an infectious agent, and the possible routes of transmission," she says.

She also asks what measures the human workers were taking before the outbreaks to minimize transmission of infectious agents between monkeys and humans. "Presumably there would be some pretty careful measures in place that would limit human exposure to any contaminant or pathogen," she says, "so saying that humans are not susceptible to me seems premature." ■

David Cyranoski

hit by a mine or rammed by another ship — a suggestion also raised by Shin Sang-chul, a former officer in the South Korean navy who was part of the investigation team.

This week, a US-led UN team plans to meet North Korean military representatives to discuss the sinking. The South Korean government has adamantly denied fabrication or major problems with its interpretation of the data. Others also doubt that there is an alternative interpretation. "Aside from the science, it is consistent with North Korea's behaviour in the past," says James Schoff, an Asian regional security expert at the Institute for Foreign Policy Analysis in Cambridge, Massachusetts. Referring to South Korea's investigation, he says: "I have no doubts personally that the conclusion is correct." ■

David Cyranoski

See go.nature.com/5q5igr for a longer version of this story.

Start-up model patently flawed

It is widely believed to be the standard route for academics starting their own business: disclose an invention to the university, get it patented and venture forth into the spin-out world. But an extensive survey has found that this is not how the majority of companies are started by US academics, suggesting that government and universities are missing an opportunity in their quest to boost entrepreneurial activity.

In what the authors say is the largest study of its kind, experts on entrepreneurship surveyed 11,572 professors at institutions across the United States (R. Fini, N. Lacetera and S. Shane *Res. Pol.* doi:10.1016/j.respol.2010.05.014; 2010). Of the 1,948 respondents who had started a business, only 682, or about one-third, had set them up to exploit patents obtained through formal university intellectual-property systems. The remaining 1,266 respondents had started businesses — including consultancies and manufacturing and service-based firms — based on non-patentable knowledge.

Social scientists and engineers started the most businesses that were not based on patented inventions, but such ventures were also prevalent among biomedical and physical scientists (see 'Unexpected entrepreneurs').

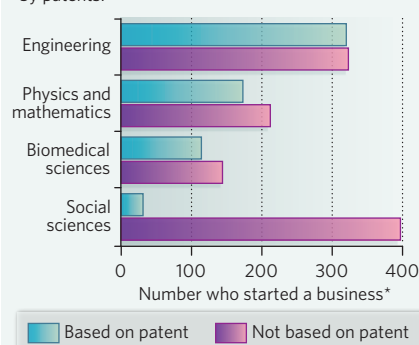
"There is a lot of stuff that academics are realizing isn't patentable but they can commercialize for themselves by starting a company," says Scott Shane, an economist at Case Western Reserve University in Cleveland, Ohio, and a co-author of the study. Because surveys of entrepreneurial activity — including government assessments — typically focus on patent activity, they may be significantly underestimating academics' efforts, he notes.

The study also says that universities' technology-transfer offices (TTOs) are "failing to help" a sizeable proportion of academic entrepreneurs. "If you aren't measuring [the full range of activities] you can't promote them," says Shane. "All the policies and approaches focus on the formal intellectual-property system, which means we are missing a big part of the iceberg that is under the water."

Ashley Stevens, president of the Association of University Technology Managers, calls the study "a significant new piece of data". But he says that TTOs do help academic entrepreneurs outside

UNEXPECTED ENTREPRENEURS

Many academics' businesses are not underpinned by patents.



*Based on 1,714 respondents who provided information about their discipline.

the formal intellectual-property system.

A major barrier to more support is that, unlike businesses started inside the system, those started without patents are less likely to provide the university with a financial return, he adds.

Yet the survey found no discernible difference in respondents' financial return between businesses started with and without patents, although those businesses not based on patents were more likely to fail.

The study comes as the US government is shaping its policy on the commercialization of federally funded research. The Office of Science and Technology Policy put out a call for information earlier this year, and is promising to use the results to help identify a set of "promising practices" that will encourage technology commercialization, says Tom Kalil, the office's deputy director for policy. "These include entrepreneurship education for faculty and students, proof-of-concept funding, standardized intellectual-property agreements, and policies that allow faculty to take 'industrial leave' to start up a company."

Some experts say that governments and universities are still right to focus on patent-based academic entrepreneurship. Businesses built on inventions tend to need more support to establish themselves, and can potentially generate the greatest returns, says Mark Schankerman, who studies entrepreneurship at the London School of Economics. "It is getting out a new drug, not a new consultancy company, which is going to make the larger contribution to the economy and society," he says. ■

Zoë Corbyn

SOURCE: R. FINI ET AL. *RES. POL.* DOI:10.1016/J.RESPOL.2010.05.014 (2010)



The lost legacy of the last great oil spill

Some ecosystems bounced back after the 1979 Ixtoc I oil spill, but research quickly withered.

The well blew out, the blowout preventer failed, and the drilling rig caught fire and eventually sank. Oil gushed into the Gulf of Mexico at a staggering rate from the damaged riser that had attached the platform to the well. Nobody knew what to do, although engineers tried various measures to stem the flow, including a containment dome. Chemical dispersants to break up the oil were applied at one of the highest rates in history. Some of the oil was trapped well below the Gulf's surface, with undetermined effects. It seemed as though the spill might drag on forever.

Call it disaster déjà vu. This all-too-familiar description refers not to the ongoing Deepwater Horizon oil spill, but to an episode three decades earlier and about 1,000 kilometres south, at an exploratory oil well known as Ixtoc I, operated by Mexico's national petroleum company Petróleos Mexicanos (PEMEX). Between 3 June 1979, the date of the blowout, and 23 March 1980, when Ixtoc I was finally capped, it spewed some 475,000 tonnes of oil into waters northwest of Ciudad

"The research was stopped. That was the real crime of that spill."

del Carmen on Mexico's Yucatán Peninsula.

Researchers who are struggling to determine the long-term environmental effects of Deepwater Horizon have begun asking what helpful lessons Ixtoc I offers. "I think it has taken a little while for the research community to rediscover it," says Arne Jernelöv, a Vienna-based environmental biochemist with the Institute for Futures Studies in Stockholm, who studied the Ixtoc I spill for the United Nations

(A. Jernelöv and O. Lindén *AMBIO* 10, 299–306; 1981). "But by and large now I think it has."

Yet answers are scarce. Because funding for studies of the spill's impact dried up soon after the spill did, experts view Ixtoc I as a missed opportunity. "The research was stopped," says Wes Tunnell, associate director of the Harte Research Institute at Texas A&M University–Corpus Christi. "That was the real crime of that spill."

Since the Deepwater Horizon blowout, Tunnell, who studied Mexico's coastal habitats extensively just after the Ixtoc I spill, has been inundated with questions about that spill and

its long-term effects — questions he couldn't answer. "I decided I needed to go back," he says. So, with support from the Harte Institute and a small foundation grant, he has revisited two of the areas that were most heavily damaged by the spill. His mission was to collect samples of any oil that might remain, and to interview local fishermen about their experiences during and after the spill.

Human history

In some cases, their memories now constitute the only available data about the effects of the Ixtoc I spill, because little has been published on the disaster since the first few years after it happened. A Mexican-government-sponsored conference took place in 1982, but a promised proceedings volume comprising papers presented there never appeared. The Harte Institute has since posted some of those papers on its website (go.nature.com/rvwUwI).

Last week, during Tunnell's most recent trip, the first stop was a rocky limestone shoreline near the town of Champotón on the Yucatán's western coast. Thirty years ago, he recalls, the

BETTMANN/CORBIS



DEBATE GROWS OVER OIL DISPERSANT

Chemical may be finding its way up the food chain.

go.nature.com/IIOWle

US NAVY/M.B. WATKINS

Vast quantities of dispersants were sprayed onto oil escaping from the Ixtoc I well.

tide pools here were thick with oil. Development and road building have changed the shoreline significantly since then. Nevertheless, Tunnell and two colleagues quickly found a 40-centimetre-wide patch of tar above the tideline. “That stuff is pretty tough,” says Tunnell. “That’s why they put it on the highway.”

After scraping off the top layer, the researchers found that the tar still glistened. “It’s still like new,” says Tunnell’s co-worker Julio Sánchez Chávez, a fisheries biologist at the Autonomous University of Campeche in Mexico. Another colleague will be analysing the samples collected to see whether they exhibit the chemical signatures of oil from the Ixtoc I well.

No fishing

It is unlikely that the oil is having any significant ecological effects after weathering away for so long, but the residue is a reminder of what the region experienced when oil first came ashore. Thirty years ago, most fishing villages were so isolated that locals knew little about the disaster playing out just a few kilometres away. Many believed Ixtoc I was a fairly small spill, although it was among the largest in the planet’s history.

But the effects were soon apparent. Carlos Castillo, now 78, was an avid skin diver at the time who used a speargun to catch fish for his small restaurant. Before Ixtoc I, he could catch 30 kilograms of grouper, snapper or snook in two hours, he says. But during the spill, his mask became oily, and he developed health problems. Eventually, he couldn’t find anything to catch because fish were dying or leaving for cleaner waters. “I told my wife, ‘Sorry, we have



to buy fish.” They began serving freshwater fish or fish trucked in from other regions.

The spill years were devastating for fishermen, many of whom had nowhere else to turn for income or food, but fisheries recovered faster than most researchers expected. Chávez says that Campeche shrimp-catch records, for instance, suggest that within two years fishermen were pulling in normal hauls again. Locals say that fish catches improved substantially within three to five years. Tunnell points out that the Gulf may have been healthier and more resilient then, so it’s difficult to say whether species in the northern Gulf will rebound as quickly from the current spill. But the curtailment of commercial fishing owing to fears over contaminated seafood may hasten the recovery of exploited species.

In some parts of Campeche, however, there are ominous signs that not all ecosystems fared as well. After leaving Champotón, Tunnell and his colleagues travelled about 125 kilometres north to the tiny village of Isla Arena, to look for oil among mangrove trees. A few kilometres from the village, the team found patches of what looked like highly weathered oil. Then Chávez called out, “Hey Wes, there’s a stinky one over here.”

“Shrimp-catch records suggest that within two years fishermen were pulling in normal hauls.”

The 2-metre by 1-metre mat of tar smelled distinctly of asphalt.

More disturbing is the absence of oysters around Isla Arena, where they were once so abundant that local fishermen say they could chop off a mangrove branch and pluck off enough of the molluscs to feed their families. The oysters never returned after Ixtoc I, according to the fishermen, and there is no research to explain why. “As far as I know, this is one of the least-studied ecosystems in Mexico,” says Tunnell. He says that he’s intrigued by the oyster story and hopes to do follow-up research on the topic.

So far, Tunnell says, Ixtoc I’s main lesson for those responding to the current spill is that sandy beaches and rocky shores can recover relatively quickly, but that more productive ecosystems such as mangrove swamps or salt marshes — the closest analogue to mangroves in the northern Gulf — retain oil indefinitely. They may take decades to regain their health.

Jernelöv says that other features of Ixtoc I may foreshadow what the coming months and years could hold. For example, although the water depth at Ixtoc I was just 50 metres, compared with 1,500 metres for the current

spill, it too generated subsurface oil plumes. That oil made its way around the Gulf, and at one point some beaches in Texas took an unexpected oil hit after it mixed with surface waters close to shore. “You didn’t see anything and then all of a sudden you had oil on the beaches,” Jernelöv says.

In broader terms, Tunnell, Jernelöv and other researchers familiar with Ixtoc I agree that its most important lesson is to continue studying the Deepwater Horizon spill and its impact — especially after the oil stops flowing and the world’s attention turns elsewhere — so as not to be caught short next time.

Steve Murawski, chief science adviser for the National Oceanic and Atmospheric Administration’s National Marine Fisheries Service, says that he and others have lamented the lack of Ixtoc I data as they have worked to respond to Deepwater Horizon. But he says that he hopes the mistakes of 30 years ago will not be repeated, and that “the number of regrets will be relatively modest when we start writing the history of this one”.

As for the fishermen near the Ixtoc I site, they are well aware of the new disaster unfolding to the north. Many fear that the oil will come their way, and they sympathize with those already affected. “To me, it’s a problem for humanity,” says Castillo.

Mark Schroppe



Oil from Ixtoc I can still be found on coastal rocks 30 years after the spill.

M. SCHROPE

Ancient Italian artefacts get the blues

Scientists accuse officials of neglect as chemicals discolour stored relics.

A mysterious blue sheen that is creeping over precious archaeological artefacts has sparked a political firestorm in Italy. Scientists are battling local authorities to save the damaged collection — and determine who is to blame.

The prehistoric treasures — including human bones and stone tools — come from sites near Verona, which were inhabited by some of Europe's last known Neanderthals when anatomically modern humans were beginning to dominate the region. Scientists say that comparing DNA from the remains with DNA from Neanderthal bones found elsewhere may show how the last Neanderthals moved across the continent seeking refuge, for example.

But now some of the remains face irreparable damage, as they lie deteriorating in a former military armoury in northern Italy. The artefacts were moved there in 2007 and 2008 after Verona's town council sold their original home — an eighteenth-century castle that provided



Erminia Perbellini, councillor for culture for Verona, inspects discoloured flints.

overflow storage for Verona's Natural History Museum. The money from the sale was intended to refurbish the arsenal to provide a new home for all of the museum's collections, but the funds were subsequently reallocated.

Concerned scientists are about to send a petition to Sandro Bondi, Italy's minister of

He points out that Verona is a World Heritage Site. "You would think they would have paid better attention to their historic collections."

The collection's curator, Laura Longo — who declined to comment to *Nature* because she has been asked by the museum not to speak to the press — first became concerned when she

culture, demanding that the artefacts be immediately transferred to safe premises. They are also calling for an expert committee to assess the damage, and for an investigation to identify who is responsible for endangering the collection — potentially a crime under Italy's strict laws on cultural heritage.

Reid Ferring, a geologist and archaeologist at the University of North Texas in Denton, says he signed the petition because he finds it "devastating that the collection with such a high scientific value has been damaged".

Universities shun Europe's drug initiative

Europe's pharmaceutical industry, struggling with huge research costs and unfriendly regulation, has been limping along for more than a decade. To help the ailing sector, the European Union (EU) in 2007 launched the multibillion-euro Innovative Medicines Initiative (IMI) to bring public and private researchers together and speed up drug development.

But the partnership is now looking less than healthy. Many universities and public research institutes are complaining that the high cost of participation, and a raw deal on intellectual property (IP) rights, is stopping them from taking part.

In its first call for proposals, the IMI received 138 'expressions of interest' from universities and research institutes across 18 research topics. In the second call,

it garnered 124 expressions of interest over 9 research topics — a resounding success, says the IMI's executive office in Brussels.

Yet as the IMI prepares to launch its third call this autumn, a group of 11 leading university and research organizations has published a joint statement saying that the call is unlikely to succeed unless the funding and IP rules are changed. The 6 July statement says that, owing to these concerns, participation in the first two rounds has been well below potential.

The research organizations say that they have repeatedly expressed their concerns to the European Commission (EC) and to the IMI's managing bodies over the past two years, but have seen no progress. Pauline Bastidon, a policy officer at the European Association

of Research and Technology Organisations, a trade association which co-signed the statement, says that some universities have been forced to give away potentially valuable IP rights in order to take part in projects.

The IMI is one of five ambitious applied-research programmes known as Joint Technology Initiatives (JTIs) — the flagships of the EU's €50-billion (US\$63-billion) Seventh Framework Programme (FP7), which runs until 2013. The EC conceived the JTIs to help promote research collaborations between academia and industry and to translate basic science into potentially lucrative technologies. Each JTI draws up its own rules on funding and IP. In the

case of the IMI, the research agenda is largely determined by industry.

"One of the key objectives of IMI is to deliver innovative medicines to patients in a faster way, by speeding up discovery and development in the

"One of the key objectives of IMI is to deliver innovative medicines to patients in a faster way."

pharmaceutical sector," says Michel Goldman, executive director of the IMI. Its IP policy

"was designed to promote swift disclosure and exploitation of new knowledge", he says.

But universities say that they are loath to join an IMI project which generates IP that they cannot own. Unless the IMI revises its IP strategy, it is unlikely that new drugs will be developed from these projects, says Thierry Troosters, a researcher at the Catholic University of Leuven

gained brief access to the relocated artefacts in February, and found that some of the flint tools were turning bright blue. She then provided four samples of the discoloured flint to Gilberto Artioli, a geo-archaeologist at the University of Padua, for chemical analysis.

Over the next three months, she reported the problem to the museum's director and also to Vincenzo Tine, the regional representative of the ministry of culture in Venice. Last month she sent a formal report to Stefano de Caro, the ministry's general director for archaeology in Rome, and to the department of military police concerned with crimes against cultural heritage. That department has already referred the case to Verona's public prosecutors.

Meanwhile, Artioli's first analyses quickly showed that the samples — and also the cardboard they were delivered in — were impregnated with hydrocarbons, which he says may come from petrol, or from lubricants once stored at the arsenal and used to keep the weapons there in working order. "But this did not explain the blue colour," he says. He suspects that hydrocarbon vapours carried a pollutant onto the flints, where it reacted to form a "very stable and durable" blue pigment that was previously unknown.

More chemical detective work could easily determine the source of this pollutant,

Artioli says. But the results would be political dynamite. If the contaminant came from the walls or floor of the building, the municipality of Verona could be guilty of allowing cultural heritage items to be transferred to a building that had not been properly checked for pollutants.

Tine, whose office is responsible for ensuring compliance with cultural-heritage laws in the Veneto region, suggests that the source of the

pollutant may be the padding in the storage cabinets acquired for the artefacts in their new premises. He also plays down the seriousness of the problem. "Only a hundred or so of the millions of objects in the collections have turned blue, and they can easily be restored to their original colour," he says.

But scientists see things very differently. "Up to 30% of the flints are turning blue," says Artioli. "Sure, we can bleach the colour out, but then palaeontologists will never be able to do any meaningful analysis from a chemical point of view." And although they have not changed colour, the bones and pottery are more porous than the stone tools and therefore will have a deeper exposure to the chemical pollutants, he says. "Now the summer temperatures are accelerating the chemical transformation," he adds. "The matter is urgent." ■

Alison Abbott

"You would think they would have paid better attention to their historic collections."

in Belgium. Troosters is taking part in an IMI project to gather information from patients with pulmonary disease to ascertain what physical activity aids their recovery. He says that he designed his study proposal to omit any technologies or concepts that the university itself hopes to exploit.

Universities also complain that the programme allows them to claim back only 20% of the indirect costs of a project, including infrastructure and computing support. By contrast, other FP7 projects reimburse universities with at least 60% of their indirect costs.

Linda Polik, European adviser at the research services unit of the University of Oxford, UK, says that although the university expressed interest in five projects in the first IMI funding round, it did not

participate in the second call because it could not afford to cover the indirect costs. She says that the university will only participate in the third round if interested departments can make up the funding shortfall from elsewhere and state a case for doing so, given the unfavourable IP rules.

"We are aware that some organizations have concerns," says Goldman. But, he notes, "155 academic and other research organizations are already participating in ongoing IMI projects under the current IP policy, and the general feedback we get from project participants is positive".

Goldman adds that the IMI has already published guidance to help applicants in the IP negotiation process, and is due to update that information before the next

funding call. And the EC, which has representatives on the IMI's governing board, is assessing the impact of the 20% costs cap.

Yet for those institutions that have participated in the IMI, it has been a bruising experience that they say is deterring others from joining. Ulf Johann, legal adviser to the Fraunhofer Society, a research organization in Munich, Germany, says that last-minute changes to a legal agreement governing the society's participation in an IMI project to develop biomarkers forced it to give up all rights to any research royalties. Johann says that the society had been threatened with exclusion from the project if it did not sign.

"People might remember [the IMI] more for the problems over IP than for providing new medicines," he says. ■

Natasha Gilbert

China faces up to groundwater crisis

Researchers call for effective monitoring and management of water resources.

A crisis is developing beneath China's thirsty farms and cities, but no one knows its full extent. With about 20% of the world's population but only about 5–7% of global freshwater resources, China draws heavily on groundwater. Those reserves are being depleted at an alarming rate in some regions and are badly polluted in many others, warned experts last week at the International Groundwater Forum 2010 conference in Beijing.

The scientists also warned that confronting the crisis will require dealing with other shortages: of knowledge and regulation. They say that a nationwide network to monitor groundwater levels is urgently needed, and that the government should improve data sharing, cut water waste and help farming become more efficient. "The water crisis is not unique to China," says Frank Schwartz, a hydrologist at Ohio State University in Columbus, who was at the meeting. "But the problem here is orders of magnitude bigger than anywhere else."

Groundwater is used to irrigate more than 40% of China's farmland, and for about 70% of the drinking water in the dry northern and north-western regions. According to *Opportunities and Challenges in the Chinese Groundwater Science*, a 2009 report sponsored by China's National Natural Science Foundation and China Geological Survey (CGS), part of the Ministry of Land and Resources (MOLR), the past few decades have seen groundwater extraction increasing by about 2.5 billion cubic metres per year to meet these needs. Consequently, groundwater levels of the arid North China Plain have dropped as fast as 1 metre a year between 1974 and 2000, forcing people to dig hundreds of metres to access fresh water, according to research presented by Bridget Scanlon, a hydrogeologist at the University of Texas at Austin.

Already, water is scarce for two-thirds of China's 660 cities, according to a survey by the Ministry of Water Resources (MOWR). And as China's economy expands, so will its demand for water. The country will consume 750 billion cubic metres of water a year by 2030, about 90% of the total amount of usable water resources in the country, projects the MOWR.

Pollution is also putting the system under pressure. In southern and southeastern China, which have seen rapid economic development, groundwater is now laden with heavy metals and other pollutants. A CGS survey presented at the conference shows that 90% of groundwater is polluted, 60% of it seriously so.



PHOTOLIBRARY.COM

The government hopes that a massive system of canals and pipes, to funnel 45 billion cubic metres of water a year from China's moist south to its arid north, will alleviate groundwater depletion once it is completed in 2050. The government also plans to reduce groundwater use in Tianjin municipality, Beijing and Hebei province to 65% of current levels by 2023, says Li Guomin, a hydrogeologist at the Chinese Academy of Science's Institute of Geology and Geophysics in Beijing. Li and his colleagues presented a simulation of the effect of this policy and say that it could boost groundwater levels. But the team assumed that groundwater use will not increase as more surface water becomes available, an assumption that was debated at the meeting. "As new resources become available, people will just use it more," argues Schwartz.

Network need

Efforts to understand the scale of the problem are often stymied by the lack of a national water-monitoring network, researchers say, and even when data exist, researchers are often unable to get hold of them. Groundwater and surface water are managed separately, by the MOLR and the MOWR respectively. This can lead to poor communication, and means that water is sometimes counted twice — when it flows underground, and again when it reaches the surface — inflating estimates of the size of the resource.

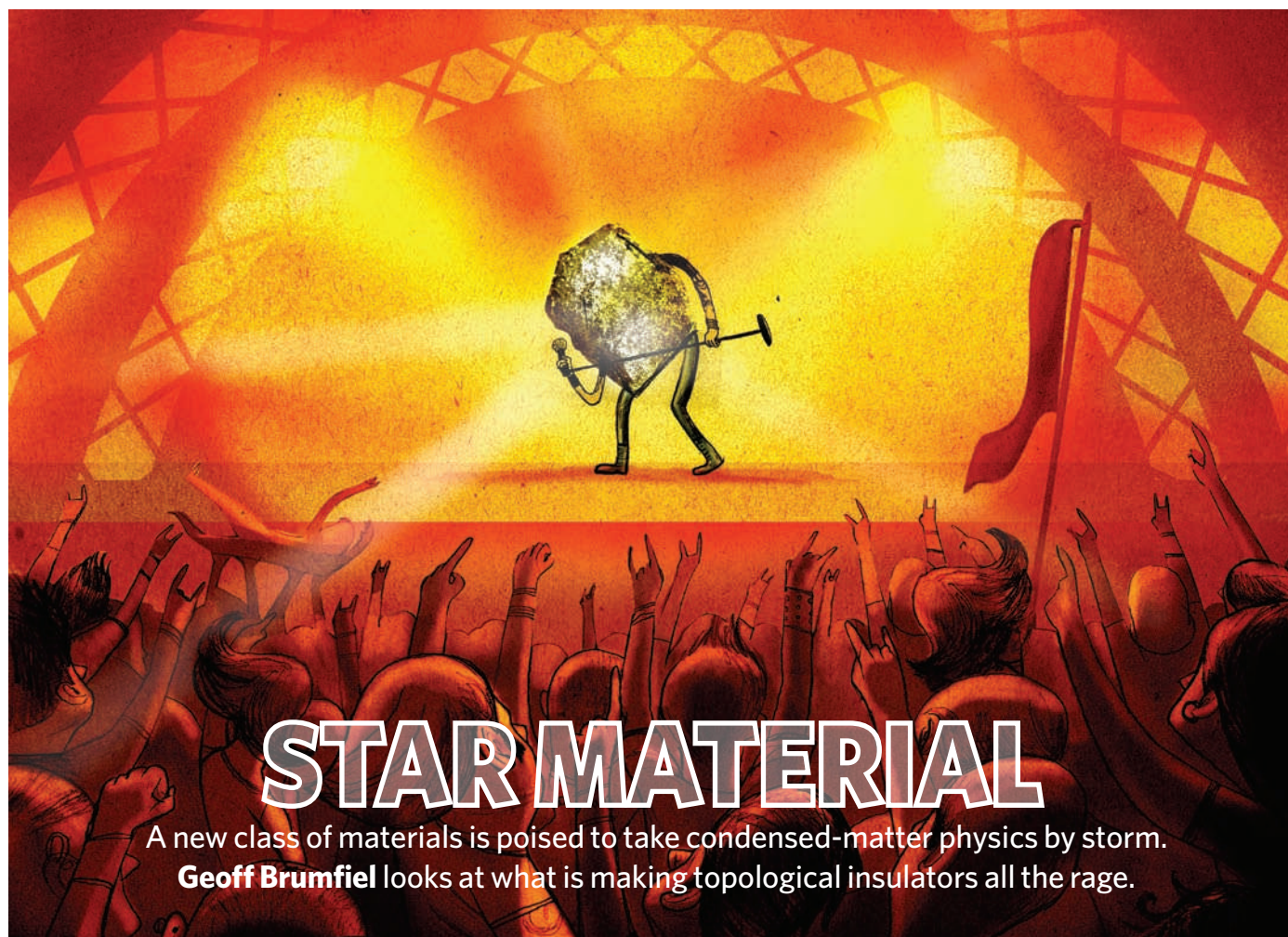
A more holistic approach "could provide a baseline to assess the impact of climate change and human activities on water resources", says Zheng Chunmiao, a hydrologist with a joint position at Peking University and the University

of Alabama in Tuscaloosa, who chaired the Beijing meeting. The central government has earmarked about 1.8 billion renminbi (US\$265 million) to set up a groundwater observation network to take measurements from tens of thousands of wells across the country. However, after years of ministry infighting, the project has still not been fully rolled out.

The network could also help to pin down the causes of freshwater depletion across the country. Climate change is often blamed, for example. But a team led by Gao Zhanyi, of the Beijing-based Institute of Water Resources and Hydropower Research, found that climate contributed only about 10–30% of the water-table depletion in three regions of China. The majority of the depletion was down to poor practice: "There is much room for improvement in terms of more effective water management," says Gao.

Arguably, the biggest improvement could come in the agriculture sector, which already uses 70% of the country's fresh water. To boost grain production, for example, China has a double-cropping system of growing wheat in winter and maize in summer. But because there is very little precipitation on the North China Plain in winter, this draws deeply on groundwater supplies. "China may want to rethink its food-security issues because the current strategy is completely unsustainable," says Scanlon. The country could import grain, and synchronize crop production with the climate by ending the cultivation of winter wheat and growing maize for more of the year, she says. "If we continue the current path of economic development, we will hit the wall very soon."

Jane Qiu



STAR MATERIAL

A new class of materials is poised to take condensed-matter physics by storm.
Geoff Brumfiel looks at what is making topological insulators all the rage.

For a brief time in Portland, Oregon, this past spring, thousands of physicists moved from session to session at the annual March meeting of the American Physical Society (APS) on the lookout for the next big thing.

It was a talent search not unlike the one that unfolds every night in the bars and converted dance halls of Portland's famous music scene, where locals listen for the next big sound. The physicists' quest is a lot harder, though. Trends in music come and go, but the disciplines that dominate the APS March meeting — such as optics, electronics and condensed-matter physics — are rooted in the original theories of quantum mechanics, which were more-or-less completed in the 1930s. When it comes to describing how light and matter behave, only a few phenomena have emerged since then to become the physics equivalent of superstars.

At this year's APS meeting, however, the hallways were filled with talk of a promising newcomer — an eccentric class of materials known as topological insulators. The most striking characteristic of these insulators is that they conduct electricity only on their surfaces. The reasons are mathematically subtle — so much so that one physicist, Zahid Hasan of Princeton University in New Jersey, tried to explain the behaviour using 'simpler' concepts such as

superstring theory. ("It's awfully beautiful stuff," he said reassuringly.) Yet the implications are rich, ranging from practical technology for quantum computing to laboratory tests of advanced particle physics.

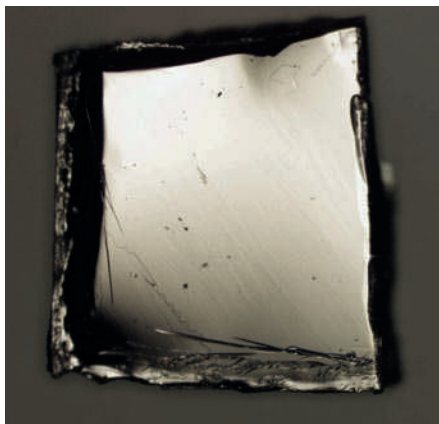
Hence the excitement. It is still too early to say whether topological insulators are the next big thing. But physicists are auditioning various formulations of the insulators in their labs, eager to determine whether the material can live up to its many promises.

A topological insulator sounds simple

enough — a block of material that lets electrons move along its surface, but not through its inside. In fact, it is far from straightforward. Ordinary metals conduct electrons all the way through, whereas ordinary insulators don't conduct electrons at all. A copper-plated block of wood conducts only on the surface, but that is two materials, not one. The idea of a topological insulator is so strange that for a long time, physicists had no reason to believe that such a material would exist.

Quantum dance

Things changed in 2004, when Charles Kane, a theoretical physicist at the University of Pennsylvania in Philadelphia, was studying sheets of carbon called graphene. Kane's calculations suggested that electrons would move through this one-atom-thick material in a way that reminded him of the quantum Hall effect: a phenomenon first observed in 1980. This effect occurs when electrons are confined to thin films of certain materials, subjected to large electric and magnetic fields, and cooled to within just a fraction of a degree above absolute zero. At that point, the ordinarily chaotic motion of the electrons gives way to a more orderly collective behaviour governed by quantum mechanics. The transition shows up in the laboratory as a series of discrete



Electrons move along the surface of, but not through, topological insulators such as bismuth selenide.

quantum steps in the capacity of the material to conduct electricity.

What Kane and his group saw in their graphene calculations wasn't exactly the same as the quantum Hall effect. Even so, further analyses showed that there might be other thin-film materials with similar behaviour. This time there would be no need for huge external magnetic fields or ultra-low temperatures to get the electrons moving in unison. Such a material would produce the magnetic field from the nuclei of its own atoms — possibly even at room temperature.

These coordinated electrons would mostly end up just spinning in place. Intriguingly, however, those at the edge of the thin film would be forced to skip along the boundary. The net result would be that thin samples would conduct electricity along the edge — and only along the edge — but in separate quantum steps, similar to those seen in the quantum Hall effect¹.

The work of Kane and his colleagues got noticed almost immediately. Joel Moore, a theorist at the University of California, Berkeley, and his co-workers built on Kane's calculations to show that three-dimensional blocks of material would also display quantum effects², although the way electrons moved along the surface would be more complex than in the flat sheet used by Kane. Moore also gave the materials a new name. They were originally termed "novel Z_2 topological invariants" by Kane, in reference to the quantum-mechanical properties that cause the electrons to skip along the edge. "We got tired of typing that out, so we called it a 'topological insulator,'" says Moore. "I don't know that that term is particularly explanatory, but at least it's short."

Meanwhile, Shoucheng Zhang at Stanford University in Palo Alto, California, and his team were researching what types of real material could be topological insulators. In most materials, Zhang realized, the link between electrons and nuclei is too weak to create a topological-insulating behaviour. But the link gets stronger as the nuclei get heavier. In 2006, Zhang predicted that one material in particular, a crystal made of the heavy elements mercury and tellurium, would be able to do the trick³. And within a year, Laurens Molenkamp, a physicist at the University of Würzburg in Germany, and his group had grown a thin layer of mercury telluride crystal and showed that its conductance hopped from one quantum value to the next along the edge of the sample⁴.

The experiment by Molenkamp proved that the theorists were onto something, but by itself it didn't cause much excitement. Mercury telluride crystals are difficult to obtain — they have to be grown one layer at a time using a laborious process known as molecular beam

epitaxy — and they are not pure topological insulators because they conduct some electricity on their inside.

New compounds based around bismuth, which are simple to make and cheap to work with, have caused the field to take off. "What got so many talks at the March meeting was the bismuth-based compounds," says Hasan. "Anybody can grow them, you can buy them off the shelf, and you don't need a high-purity crystal to see the topological effects."

Those effects go beyond the way electrons move on the surface. For example, all electrons are spinning in a quantum mechanical way. Usually, the spins are constantly knocked about by random collisions and stray magnetic fields. But spinning electrons on the surface of a topological insulator are protected from disruption by quantum effects (for more, see page 323). This could make the materials beneficial for spin-related electronics, which would use the orientation of the electron spin to encode information, thereby opening up a whole new realm of computer technology.

Mathematical mimicry

Researchers also believe that the collective motions of electrons inside topological insulators will mimic several of the never-before-seen particles predicted by high-energy physicists. Among them are axions, hypothetical particles predicted in the 1970s; magnetic monopoles, single points of north and south magnetism; and Majorana particles — massless, chargeless entities that can serve as their own antiparticles.

This mimicry is not entirely surprising. Almost by definition, collective electron motions can be described by just a handful of variables obeying simple equations, says Frank Wilczek, a Nobel-prizewinning particle physicist at the Massachusetts Institute of Technology in Cambridge. "There are only a few kinds of equations that you can write down that are really simple," he says. So topological-insulator theorists and particle physicists have almost inevitably ended up in the same place.

Majorana particles could prove particularly useful for practical quantum computing. The idea is to perform calculations using the laws of quantum mechanics, which could make computers much faster than the normal variety at certain tasks such as code-breaking. But the

fragile quantum states essential to their operation are easily destroyed by jolts from the outside environment. Majorana particles would spread quantum information across particles, making them far more resistant to interference, says Kane. If Majorana particles could be harnessed on the surface of a topological insulator, "this would be big," he says.

The wealth of calculations, experiments and applications offered by topological insulators, together with their availability, have given the field a white-hot status at the moment — as has a certain thirst for glory. Two variations of the quantum Hall effect have won their discoverers Nobel prizes, and some researchers think

that a Nobel awaits whoever can contribute the most to the growing field. "I'm not thinking about that at this point," says Kane, but the competitiveness has forced him to ensure that others are aware of his work. "I sort of feel like if I don't assert myself, then I'm going to get buried," he says.

Yet trips to Stockholm are some way off. Although samples of topological insulators are now easy to get hold of, most still contain impurities that cause them to conduct electricity on the inside, disrupting the states on the surface. Getting things perfect remains more of an art than a science. Furthermore, some of the sought-after effects will require topological insulators to be combined with more common materials. To create Majorana particles, for example, topological insulators will have to merge with superconductors. Many experiments on how best to do that are under way.

The results of these studies will determine whether topological insulators are more than a one-hit wonder. Regardless, says Moore, there is an undeniable appeal in how the collective behaviour of electrons can lead to so many wonderful things. "There's something about many-particle quantum mechanics that causes perfection to emerge out of imperfection," he says. "That's somewhat cheering as far as our everyday lives are concerned."

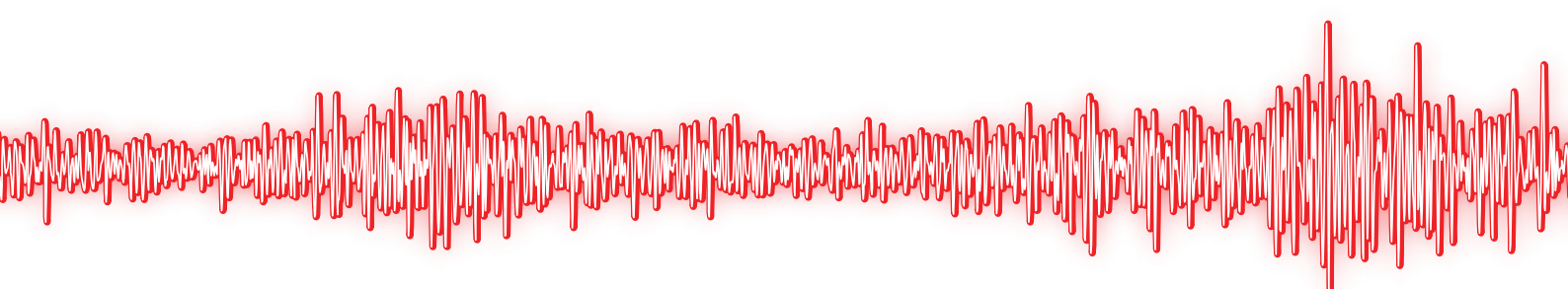
Geoff Brumfiel is a senior reporter for *Nature* based in London.



"There's something about many-particle quantum mechanics that causes perfection to emerge out of imperfection."
— Joel Moore

1. Kane, C. L. & Mele, E. J. *Phys. Rev. Lett.* **95**, 146802 (2005).
2. Moore, J. E. & Balents, L. *Phys. Rev. B* **75**, 121306 (2007).
3. Bernevig, B. A., Hughes, T. L., Zhang, S.-C. *Science* **314**, 1757–1761 (2006).
4. König, M. et al. *Science* **318**, 766–770 (2007).

See News & Views, page 323, and Letters, page 343.



The secret chatter of giant faults

An imminent swarm of tiny quakes beneath western North America could help seismologists prepare for a big one — but only if they can learn to interpret the tremors, finds **Naomi Lubick**.

For the past few weeks, seismologists at the University of Washington in Seattle have been on high alert. Any day now, they expect a flurry of microtremors deep under the nearby Olympic Peninsula, just as occurs roughly every 12–14 months. And when that wave of vibrations comes along, the researchers will be ready to catch it.

Over the past year, the seismology team has set up an elaborate net — an array of more than 100 seismic sensors planted in the ground throughout the peninsula's mountains. When those instruments start picking up signs of the tremors, the researchers will rush back out to install extra seismometers above the spots where the earth is shuddering. Every few days the scientists will tend to the sensors, replacing batteries and downloading data, with the hope of capturing as much information as possible about the seismic events underfoot.

The quakes will pass within a few weeks without anybody feeling them. But for seismologists, these swarms of tremors, sometimes called non-volcanic tremors, are among the most significant discoveries of the past decade because they could yield insight into the behaviour of destructive faults. Researchers first noticed them in 2002 in Japan, and a year later teams in the Pacific Northwest detected them beneath the Cascadia region near Seattle.

The Cascadian tremors are attracting interest because they hail from deep beneath the coast, along a massive fault that is thought to have unleashed earthquakes approaching magnitude 9. Researchers hope that the detailed measurements they are about to make will help them monitor activity along the fault. It may be possible to use the recurring swarms as a probe to track changes in stress underground; the tremors may even yield clues about the timing and location of future large quakes.

"It's pretty exciting," says Greg Beroza, a geophysicist at Stanford University in California, who is watching for the results from the University of Washington team. The regularity of

the quakes lets researchers prepare for them, providing an exceedingly rare chance to study seismic events in great detail. "If you had told me 10 years ago that there was going to be some kind of activity, happening on deep faults, that we could predict was going to happen again soon with some regularity, I would have told you that you were crazy," says Beroza.

Elusive signals

The seismic trace of a tremor looks different from that of a classic earthquake. The latter starts with a bang — a burst of high-frequency energy that rapidly tails off. Tremors are tamer; their vibrations are rich in low frequencies, and they come and go with less fanfare. These characteristics, and the weakness of the tremors, make them hard to spot. Preliminary observations of tremors have come from Alaska and Mexico, among other places, but information remains sketchy because only a few places have networks of seismometers sensitive enough to catch them.

Seismologist Kazushige Obara of the University of Tokyo Earthquake Research Institute was the first person to spot tremors, using signals recorded by a permanent monitoring array buried below the Earth's surface¹. The Japanese government established the country-wide network of 600 monitors after the devastating 1995 Kobe earthquake; the instruments are sunk more than 100 metres below ground, to shield them from human-made noises — such as the rumble from trucks — that can drown out the tremors' vibrations. When Obara detected extremely deep microearthquakes lasting for several weeks, 30 kilometres beneath Japan, researchers around the world pored over their own data to see whether they could spot similar traces.

Tremors in Japan occur mostly along junctions between tectonic plates where one dives or 'subducts' beneath another. The same structure underlies the Pacific Northwest, along what is known as the Cascadian subduction zone, which stretches from Vancouver Island

in British Columbia, Canada, to northern California (see 'The origin of tremor swarms'). Here, the Juan de Fuca plate is slipping beneath the North American continental plate. Since the last huge quake, which is thought to have occurred in 1700, the dangerous part of the Cascadian subduction zone has remained locked, but researchers expect it to give way cataclysmically within the next century.

With that risk in the background, researchers are especially interested in capturing tremor swarms as a way to spy on what is happening deep down in the subduction zone. Last year, they missed their chance. The University of Washington team expected the swarm to occur in August, but it arrived in early May, before many of the instruments were in the field. The researchers recorded bits of the event using a permanent array, but in much less detail than they had hoped.

This year, they are better prepared. Once their permanent array picks up the first signs of the tremors, they will head out to augment the network above the swarm. Each completed array will consist of about two-dozen seismometers stationed 200–300 metres apart in remote areas, away from human noise. The team's work is supported by a US\$500,000 grant from the US National Science Foundation to Earthscope, a non-profit research consortium.

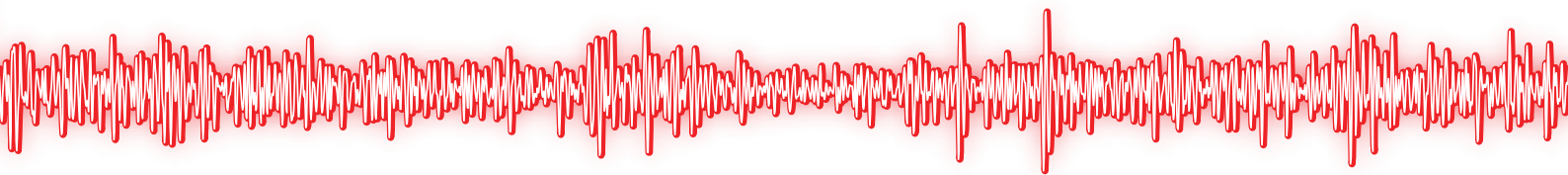
"The experiment we have out now will be unprecedented with regard to what we can see," says John Vidale, a seismologist at the University of Washington and one of the lead scientists on the project.

The team's network, called an 'array of arrays', can pinpoint the source of the tremors by comparing when the seismic waves arrive at each seismometer, says Vidale. "As with your ears, you can tell which direction a sound is coming from," he explains.

The origin of tremor swarms is generating intense debate. In Japan, seismologists have observed that tremors come from the plane where the plates meet, the subduction-fault boundary, which also gives birth to mammoth quakes. But in Cascadia, seismic signals indicate that some tremor may also occur above

Tremor swarms can help researchers spy on dangerous faults.

H. HOUSTON



the fault plane, according to Honn Kao and his colleagues, geophysicists at the Pacific Geosciences Centre in Sidney, British Columbia².

Deep origins

Locating the source of the vibrations could help researchers to understand the physical processes that create them. The tremors originate tens of kilometres underground, where Earth's cold crust is pulled into the warmer mantle. Rocks in the upper crust are brittle enough to break under stress, which leads to earthquakes. But those in the mantle are so hot that they warp under pressure instead of cracking. Tremors seem to happen in the complex zone where the subducting plate has warmed enough for its rocks to become malleable.

A study published this week³ could help to explain the process that produces tremors under parts of Japan. Satoshi Ide, a seismologist

at the University of Tokyo, suggests that tremor behaviour beneath Shikoku island could be controlled by the types of rock present in the subduction zone there. Ide found that the locations of some tremors line up in stripes matching the direction in which the subducting plate has travelled during the past 10 million years. He hypothesizes that these bands could trace the path left by seamounts, the remnants of former volcanoes embedded in the oceanic plate that is diving under Japan.

If he is right, the seamounts could influence the behaviour of tremors in two possible ways, he says. They might snag on the overlying plate as the crust slides into the mantle, creating occasional bursts of tremors. Alternatively, their chemical composition might cause them to act as lubricants, like lumps of grease, smoothing out the movement of the plates and turning what could have been catastrophic

ruptures into bumpy jitters.

Beroza calls Ide's findings preliminary, but "a really important result" that indicates how the properties of the rocks along faults might be preserved for millions of years and affect the occurrence of tremors today. As in Japan, tremors in Cascadia sometimes occur in bands that might reflect some characteristic of the subducting rocks.

Researchers are trying to determine what conditions create tremors instead of full-blown earthquakes, using data from old and new laboratory experiments that test how friction, porosity and other attributes of rock change under pressure. These studies could help reveal what controls the movement of faults at subduction-zone depths, says geophysicist Paul Segall of Stanford University.

Quake clues

Scientists also wonder how tremors relate to larger earthquakes on the same faults. The better-observed tremor events in Japan and Cascadia tend to be close to the parts of subducting plates that produce great earthquakes; the tremors originate below the 'locked' areas that store up energy for future large quakes.

Joan Gomberg, a geophysicist at the University of Washington, says researchers are looking for changes in tremor patterns that may hint at stress shifts in a fault. If such connections emerge, seismologists could start to monitor tremors for signs that precede a big earthquake. They have found some intriguing clues but nothing solid so far. The San Andreas Fault in California, for example, has experienced tremor events alongside earthquakes of magnitude 1 or 2, but direct associations between the two have yet to be confirmed⁴.

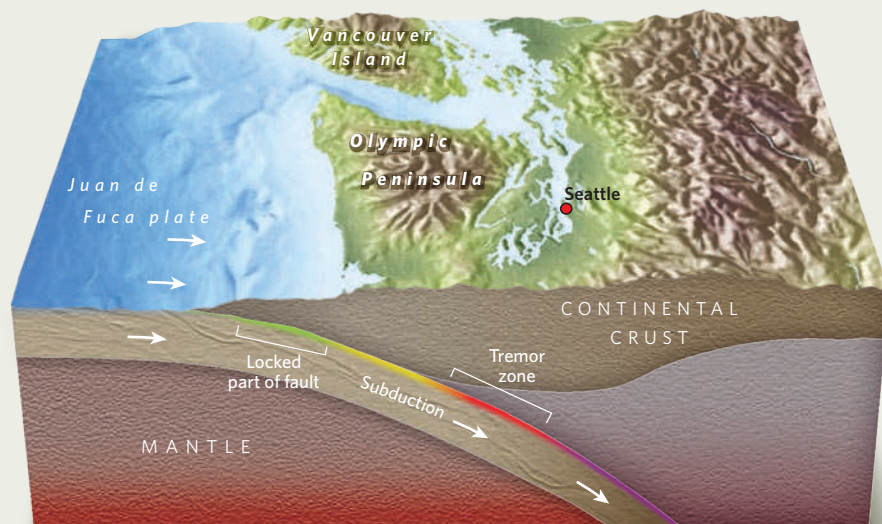
Meanwhile, Obara and his colleagues plan to put out an extra monitoring system, to gather more details about tremor events beneath Japan. And researchers in Cascadia are ready to pounce when the swarm hits.

"These things happen almost like clockwork, but not exactly," says seismologist Ken Creager of the University of Washington, recalling last year's disappointment when the team missed the tremors. This year, he is more confident. "We'll catch it one way or another."

Naomi Lubick is a freelance writer based in Zurich, Switzerland.

THE ORIGIN OF TREMOR SWARMS

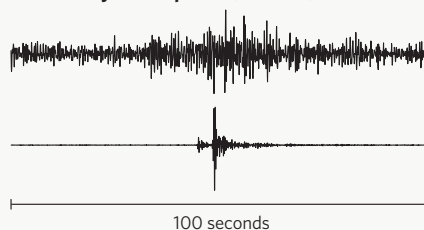
N. SPENCER



H. HOUSTON

A patch of oceanic crust is slipping beneath the North American tectonic plate along the Cascadia subduction zone. The upper part of the fault plane between the plates is locked, building up stress that will be relieved in a huge earthquake. The lower part of the fault or nearby regions produce swarms of tremors roughly every year; studies of the regular tremors might help seismologists keep tabs on the state of the subduction zone.

Seismic signals from a tremor (top) and from a tiny earthquake (bottom).



1. Obara, K. *Science* **296**, 1679-1681 (2002).
2. Kao, H. et al. *J. Geophys. Res.* **114**, B00A12 (2009).
3. Ide, S. *Nature* **466**, 356-359 (2010).
4. Nadeau, R. M. & Guilhem, A. *Science* **325**, 191-193 (2009).

ERA of austerity

The economic crisis is a setback to the European Research Area, warns **Colin Macilwain** — and the research community is ill-placed to respond.

The European Research Area (ERA) — the concept of Europe as a unified entity in which people can collaborate on science — is a grand idea. Like other grand European ideas, it is under pressure this summer, as the cracks in the European Union (EU) start to show.

Divisions have surfaced most glaringly in the economic sphere, where tension over the debt crisis in Greece has led some observers to question the very survival of the euro — the grandest European idea of them all. But the problems have implications for the scientific world too. Splits among Europe's leaders, together with impending spending cuts, can only drain impetus from European research projects and programmes.

Attending the Euroscience Open Forum (ESOF) meeting in Turin, Italy, this month, it struck me that the community of European scientists who might defend these programmes remains alarmingly underdeveloped. The institutions that might fight for research — academies, societies and media outlets — just don't exist. That's made it hard for organizations such as Euroscience — a small non-profit group based in Strasbourg, France, which seeks to unite researchers, corporations and lay admirers of science — to find a footing.

Part of the ERA's forthcoming crisis is financial. Eighty-five per cent of Europe's €80-billion (US\$100-billion) annual public spending on research and development comes from independent national agencies. As these come under unprecedented budget pressure, they will retrench, and show even less interest in collaborating with their neighbours. The remaining 15%, for joint research, is only ever expanded when national coffers are flush. The European Research Council, for example, got going when the last EU budget round was completed in 2006 only because new money was available on top of the previous research budget. That won't be the case when the next seven-year budget is done, in 2013. And the chances of rapid research-council expansion thereafter recede by the day.

Even more seriously, European research policy has a credibility gap (*Nature* 464, 349; 2010). The EU has made no progress in meeting its target, set at the European Council of 2002, of raising research-and-development spending from 1.9% to 3% of EU gross domestic product by this year. And no one has been held accountable. That lack of accountability reflects the central weakness of the EU. To function politically,



any nation, 'community' or 'union' needs unitary institutions — a credible legislature, a political culture and a media.

Twenty years ago, these were widely expected to evolve. They have not. The European Commission, in particular, addresses most of its verbose pronouncements to an empty echo chamber. Its research-directorate press conferences sometimes comprise one or two reporters and up to 30 officials. Thus its framework programme for research — whose €7-billion budget is substantial, half of NASA's — has no public profile. And that will undermine efforts to put research at the heart of the next EU budget. One of the first suggestions made by Máire Geoghegan-Quinn, the new EU research commissioner, was to change the framework programme's name to something more emphatic — an approach that, I fear, misses the point.

Community action

It would take a genuinely European 'scientific community', as well as a European media, to develop broader awareness of research programmes. Most of Europe's scientists think of themselves as 'good Europeans'. But — understandably — they like the idea most in the abstract. They tend to draw the line at joining bureaucratic, Europe-wide collaborations that they see as irrelevant, or even detrimental, to their careers. For many, the critical paths of professional development more obviously lead more through Boston or San Francisco than through Berlin or Seville.

Many factors underpin that perspective, and the biennial ESOF meeting attempts to address one of them: the prevalent global structure of scientific meetings. The broad-discipline meetings that matter most to professionals are invariably held in North America, by the likes of the American Physical Society or the Society for Neuroscience. The latter's last jamboree drew 30,500 visitors to Chicago in October, including almost 4,000 from Europe.

The main elements of the ESOF meeting, created in 2004, are a multidisciplinary, policy-orientated programme similar to that of the prestigious American Association for the Advancement of Science (AAAS) Annual Meeting and a public outreach component of scientific discussions and displays. Once an ESOF host city has been selected, it falls mainly to local organizers to finance and arrange the meeting. The AAAS, by contrast, is a large-membership society with plenty of publishing income.

At the last ESOF meeting, in Barcelona, Spain, in 2008, the future looked bright for the development of the forum, for the European scientific community that it seeks to nurture, and for European research in general. This year's event, despite thought-provoking presentations and an impressive public programme in the squares of old Turin, failed to scale quite those heights. It avoided awkward policy issues, such as the recent, searing cuts in Italian research budgets. And a good number of those who run science in Brussels and Rome — never mind in Paris, London and Berlin — didn't attend. Palpably apprehensive about the future, several speakers expressed exasperation at the fanciful verbiage (from "grand challenges" to "joint programming") that surrounds discussions of European research plans.

ESOF's inconsistency reflects wider divisions in European science, between nations and regions, that presage trouble for the ERA. Last month, the 27 member states of the EU endorsed 'Europe 2020', a "strategy for smart, sustainable and inclusive growth". The document spells out all the fine things that Europe is going to do in the next decade to start the spree of innovation that is supposed to dig the continent out of the economic hole in which it finds itself.

This, together with a research and innovation strategy being developed by Geoghegan-Quinn, is to be discussed by heads of government at the autumn European Council in Brussels in September. Let's hope their Blackberries are switched off, because little in the public pronouncements of Nicolas Sarkozy, David Cameron or Silvio Berlusconi so far suggests that they are ready to concentrate for more than 5 minutes on research and innovation, let alone agree on a course of action to make Europe innovate effectively.

One probable consequence of these leaders' fiscal policies is a collapse in research funding in Europe, as has already occurred in Italy. The ERA struggled to get anywhere even in a long period of continued economic growth. In an era of austerity, its prospects look grim. ■

Colin Macilwain is based in the United Kingdom.

e-mail: cfmworldview@gmail.com

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

CORRESPONDENCE

Questioning the timeline of H1N1 flu vaccination contracts

We take issue with points made in your News story (*Nature* **465**, 672–673; 2010) about our investigation into how the World Health Organization (WHO) handled conflicts of interest with the drug industry in planning for the H1N1 'swine' flu pandemic (D. Cohen and P. Carter *Br. Med. J.* **340**, c2912; 2010).

You confuse the findings of our separate investigation with a report published on the same day by the Council of Europe.

You also claim it is untrue that vaccine contracts were triggered by declaration of the pandemic on 11 June 2009, and that some countries had already placed large orders for H1N1 vaccine and so could not have been influenced by the WHO Emergency Committee. In fact, the committee met three times before the pandemic was declared, and twice before the signing of the US contract you cite as conclusive evidence of our error.

The first meeting took place on 25 April 2009. The next day, the United States announced a public-health emergency and the creation of seed stock for an H1N1 vaccine (see go.nature.com/38QyBj).

The fact that some countries pre-empted the 11 June declaration by ordering vaccine is immaterial. On 14 May 2009, Alan Johnson, then UK health secretary, announced: "We have advance purchase agreements with manufacturers that will be activated if the World Health Organization moves to phase 6 — that is, if it declares a pandemic. We are still at phase 5 at the moment. However, we have always known that it might take four to six months before a matching vaccine becomes available, and more than a year before it can be manufactured in sufficient quantities for the entire population, given that international demand will be high."

(For a more detailed response to your News story, see

go.nature.com/agFB1b).

Deborah Cohen BMJ Editorial, BMA House, Tavistock Square, London, WC1H 9RJ, UK

e-mail: dcohen@bmj.com

Philip Carter The Bureau of Investigative Journalism, City University, London EC1V 7HD, UK

Editorial note: Nature stands by the facts in the News article; see update at go.nature.com/CdaLJm.

Gender agenda: positive steps taken in Germany

Professional scientific associations are in a strong position to ensure that gender is included in the human-research agenda (*Nature* **465**, 665 and 688–690; 2010).

For example, the German Society of Epidemiology (DGEpi) has included gender considerations in its guidelines for more than ten years. It demands justification for any study that focuses on one sex if the results could potentially affect both sexes.

Gender will be taken into account in an important new German cohort study of multifactorial chronic diseases, including diabetes, cancer and dementia; this will follow 200,000 healthy participants recruited over the period 2012–16 (see go.nature.com/ah1Uew). Decision-making bodies associated with the study will have to ensure that female researchers are also appropriately represented. Another, nationally funded, initiative supports the integration of gender factors in all new research projects in medicine and epidemiology (see go.nature.com/nv2Fvo).

Germany's main epidemiology and public-health associations are endorsing both projects.

Hajo Zeeb, Jacob Spallek, Ingeborg Jahn Department of Prevention and Evaluation, Bremen Institute for Prevention Research and Social Medicine, University of Bremen, Linzer Straße 10, 28359 Bremen, Germany
e-mail: zeeb@bips.uni-bremen.de

Assessing effects of afforestation projects in China

In their Correspondence, Cao *et al.* suggest that afforestation projects in drylands could be exacerbating environmental degradation (*Nature* **465**, 31; 2010). However, afforestation in China does not only involve tree planting: planting shrubs for environmental protection is one of the many afforestation activities defined by the Forest Law of the People's Republic of China (see go.nature.com/pgZqRY). Environmental changes resulting from tree planting in specific regions cannot be extrapolated to entire afforestation projects.

Most water from atmospheric precipitation cannot be effectively conserved *in situ* even if it is not partially consumed by natural or artificial vegetation. Serious soil erosion and environmental disasters, such as floods caused by run-off, are key environmental degradation problems in drylands, especially in the Loess Plateau in northwestern China. Scientists and policy-makers need to recognize the inherent complexities of dryland ecosystems so that the effects of the indisputable trade-off between soil-erosion control and water conservation can be minimized.

The low resilience of dryland ecosystems means that natural restoration is extremely difficult. Afforestation is one way to reconstruct almost natural dryland ecosystems. We do not deny that some afforestation efforts have had suboptimal results, caused by the use of unsuitable tree species, or inappropriate initial densities or afforestation methods. Also, some official reports or data from case studies have overestimated or exaggerated the environmental effectiveness of afforestation activities in drylands. However, forest ecology projects have undoubtedly contributed to environmental improvement and have therefore been promoted for

the sustainable development of drylands in China.

A more scientific and feasible assessment system of these projects should be established to provide more accurate information for the public.

Xiaohui Yang, Zhiqing Jia, Longjun Ci Institute of Desertification Studies, Chinese Academy of Forestry, Yiheyuanhou, Haidian District, Beijing 100091, China
e-mail: yangxh@caf.ac.cn

Cao and colleagues reply:

As Yang *et al.* point out, we focused on tree planting in China in our Correspondence; however, planting the wrong species of shrub can also cause depletion of soil water and exacerbate environmental degradation (X. M. Mu *et al. Acta Pedologica Sinica* **40**, 210–217; 2003).

We disagree that the main factor responsible for soil erosion in drylands is run-off: serious soil erosion and related environmental disasters, such as desertification and sandstorms, are caused by wind — particularly where annual precipitation is less than 400 millimetres (X. M. Wang *et al. Earth-Sci. Rev.* **88**, 188–206; 2008). Wind erosion can increase in grassland that has been invaded by shrubs or trees because the protective covering of grasses is damaged.

Smaller trade-offs and better compromises between soil and water conservation can be achieved when environmental restoration relies on species that do not use more water than is available. A reliance on afforestation is not the solution, even when the definition of afforestation is expanded to include the planting of shrubs. The solution is to identify the most appropriate vegetation.

Shixiong Cao, Guosheng Wang College of Economics and Management, China Agricultural University, No. 2 Yuan Ming Yuan West Road, Haidian District, 100094, Beijing, China
e-mail: shixiongcao@126.com
Li Chen Water and Soil Conservation Institute of Yan'an City, 716000, Shaanxi, China

OPINION

A new strategy for energy innovation

The US government must make the Department of Defense a key customer for energy technologies and make greenhouse-gas reductions a public good, say **John Alic, Daniel Sarewitz, Charles Weiss and William Bonvillian**.

Limiting the concentration of carbon dioxide and other greenhouse gases in Earth's atmosphere requires a technological and economic revolution^{1,2}. This kind of change takes decades, even if it is driven by powerful market and policy forces — which this one is not. We therefore suggest a new, forceful strategy for the United States, the world leader in innovation and the world's second biggest emitter of greenhouse gases.

The US government must weave into energy policy an understanding of how innovation proceeds. It occurs mostly in private firms and depends on relationships between government and industry. So the government needs to move beyond the smorgasbord of research and development (R&D) programmes and financial subsidies of recent years to treat innovation as a system, with a portfolio of policies adapted to the many technologies needed and their many evolutionary stages.

First, Washington should make greater use of a powerful policy tool that it has largely neglected: public-sector procurement. Procurements by the US Department of Defense (DOD) and other agencies were the foundation for major waves of innovation after the Second World War, including those in jet propulsion, Earth-orbiting satellites and digital electronics.

Second, greenhouse-gas reductions should be treated as a public good, like infrastructure investments in public health and safety and, indeed, national defence. This would exploit a historically powerful rationale for spending on innovations that markets will not otherwise supply. The government has often been a primary customer for new technologies that advance public well-being. This approach would provide new political options for tackling problems such as the capture and storage of CO₂ from fossil-fuel-burning power plants and would complement any eventual pricing of greenhouse-gas emissions to induce innovation.

In the US Congress, the prospect of a price on greenhouse-gas emissions hangs by a thread. Even if the Senate votes for such a measure later this summer, legislation may be restricted to the utility sector, and the price will almost certainly start low and increase slowly. The Congressional Budget Office, for example, projects that the

cap-and-trade bill passed by the House of Representatives last year would price emissions at US\$20 per tonne of CO₂ equivalents after half a dozen years, and this almost certainly represents the upper limit of legislative forcefulness.

Much more is needed. Advances in energy technologies must penetrate an existing technological and economic infrastructure that took roughly a century to put in place and now represents enormous sunk costs that are protected by powerful vested interests³.

Many analysts in the United States have pointed to underinvestment in R&D by the US Department of Energy (DOE) since the 1980s as a symptom and a cause of slow energy-technology development. The administration of President Barack Obama has boosted R&D

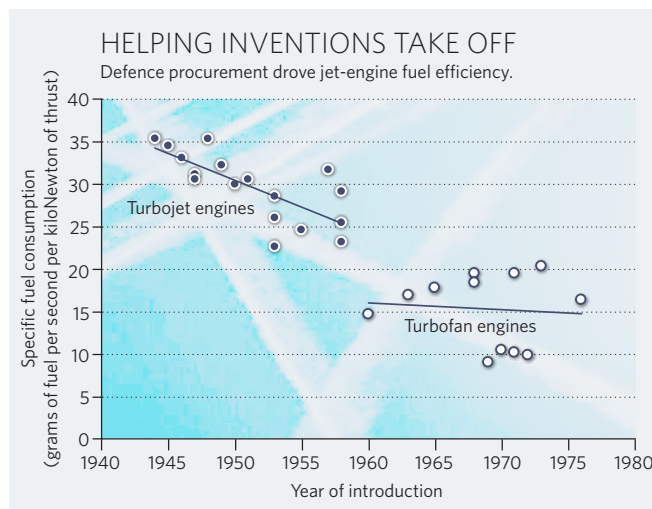
immediate practical applications. Although advocates see basic research as the wellspring of breakthroughs, many radical innovations, including the jet engine, the microprocessor and the Internet, stemmed mainly from incremental advances that were motivated by anticipated applications.

Competition and cooperation

Basic research is essential for future innovations, but there is a larger issue. For two main reasons, government R&D by itself, almost regardless of its scale, cannot foster innovation on a broad front. The first reason is simply that, although publicly financed research deepens the knowledge base and sometimes leads directly to technological advances, innovation has many sources other than R&D. As Edmund Phelps, winner of the 2006 Nobel prize in economics, put it: "Once in a while there is a big leap which creates the ground for a surge of innovations to follow. Nowadays we realize that an awful lot of innovation just comes from business people operating at the grass roots having ideas on the basis of what they see around them. Nothing to do with science"⁴. Second, R&D is difficult to manage and often unproductive in the absence of strong feedback from customers and users.

The DOE neither buys nor sells goods or services based on energy and climate innovations. It therefore has few incentives to manage R&D in accord with marketplace needs rather than scientific norms. The DOE also has not developed robust connections with companies that are positioned to commercialize energy innovations, although this may be changing. With federal spending on energy R&D on the rise, policy-makers should look beyond the dollars to the institutional setting for innovation.

Two big lessons can be drawn from innovation experience during the cold war. Government procurement helps to guide and shape technology development, and competition among agencies within the government fosters innovation, in much the same way as competition in economic markets⁵. Government purchases of integrated-circuit chips in the 1960s fostered advances in microelectronics at least as much as did government-funded R&D. In



spending (and loans and loan guarantees), channelling more than \$5 billion in economic-stimulus funds to the DOE's non-defence R&D programmes. This is a good start, as are initiatives that include the DOE's Advanced Research Projects Agency-Energy, Energy Innovation Hubs and Energy Frontier Research Centers. These promise a break with the DOE's focus on laboratory science and a move towards developing technologies with more immediate commercial prospects.

But past organizational change at the DOE has had little result, and it is doubtful whether DOE-funded R&D alone can catalyse the type of innovation needed. In any case, much of the DOE's R&D budget has funded big physics programmes, which have the potential to be important in the future but have few

SOURCE: REFS 6, 7

anticipation of government purchases (but without R&D contracts from the government), Texas Instruments and Fairchild Semiconductor fabricated the first integrated circuits in 1959–60 and went on to sell chips in large numbers to the DOD and NASA. At first, the government was the only buyer. As costs came down and technical performance improved, commercial markets opened and accounted for four-fifths of sales by 1970.

Procurement likewise spurred early developments in satellite imagery and communications. Along with major industrial firms, entrepreneurial start-ups such as Itek Corporation had key roles in the pioneering surveillance programme by Corona satellites. Telecommunications satellite technology also came mainly from the private sector. The armed forces, again, were the initial customers for gas turbine engines, and even today many commercial jet engines are derivatives of military models. Fuel efficiency improved rapidly, spurred by the military's demand for long-range planes and by commercial airlines, which wanted to reduce their operating costs (see 'Helping inventions take off'). By the 1970s, electrical utilities were purchasing gas turbines for peak power capacity, an application that had not been anticipated two decades earlier.

In all of these cases, intragovernmental competition and cooperation — between the air force and NASA for integrated circuits, intelligence agencies and the military for spy satellites, and the navy, air force and army for gas turbines — and subsequent competition in industry for government contracts, drove innovation. Managers in industry who were seeking government contracts had strong incentives to manage R&D effectively, hire creative engineers and scientists, and find customers with needs that they could satisfy in the huge US military and space establishments. This dynamic continues to operate for gas turbines and satellites. Such intragovernmental competition has sometimes led to the duplication of efforts, as well as other forms of waste, but military spending continues to drive innovations, ranging from new materials to the Global Positioning System.

The DOD is better placed for catalysing rapid innovation in energy technologies than the DOE because the DOD is a major customer for energy-consuming systems and equipment for its roughly 500 permanent installations, as well as for operational equipment (spending \$10 billion a year on liquid fuels alone). The scale of the resources that the DOD brings to technology development is impressive. It employs more than 30,000 engineers and scientists in R&D and procurement, and its annual R&D spending comes to about \$80 billion, with

procurement spending in excess of \$100 billion. The DOD thus has the incentives and capacity to be a smart and demanding customer for new energy technologies, as well as a test bed for new ideas such as high-energy-density electrical storage. Although the DOD's implementation of new energy technologies will be limited by its military mission, its internal priorities have already begun to shift in order to control operating costs and reduce energy-related strategic vulnerabilities.

Government procurement, by the DOD especially, could be an initial alternative and a long-term complement to legislated energy price increases or capping greenhouse-gas emissions. As in past technological revolutions, effectively managed procurement programmes would create powerful incentives for private firms to innovate, and would provide feedback for guiding those efforts. Moreover, by creating competition for the DOE, this approach would encourage risk taking and provide benchmarks for performance and accountability.

A public good

Reducing greenhouse-gas concentrations in the atmosphere is a public good and should be recognized as such, much like other traditional responsibilities of government. A public-good approach would empower the government to take a direct role in moving energy technologies into society to achieve the desired benefit. The designation of this as policy would also commit to action a much broader range of government institutions than just the DOE. This could galvanize technologies that, so far, have received little attention, such as those for directly removing CO₂ from Earth's atmosphere, a process that is technologically possible but costly.

Most immediately, the public-good approach could apply to a key intervention for greenhouse-gas reduction: existing coal-fired power plants. There are a manageable number of these, fewer than 1,500 generating units on about 600 sites. They produce nearly 28% of US greenhouse-gas emissions and one-third of US CO₂ emissions. That's one-fifth more CO₂ than is produced by the roughly 250 million motor vehicles travelling US roads. Moreover, there are technologies for retrofitting, repowering or otherwise reconstructing coal-fired (and natural-gas-fired) plants to remove most of the CO₂ from their emissions. What is needed now is for integrated carbon capture and sequestration to be demonstrated at operating scale. By contrast, substantial reductions in motor-vehicle emissions will occur only as

the existing vehicles are replaced by an as-yet unclear mix of emerging technologies.

The Tennessee Valley Authority (TVA), a public corporation, operates 11 coal-fired plants with nearly 60 generating units. Congress could direct and fund the TVA to explore carbon-capture technologies and to move quickly to demonstrate them at an increasing scale. The estimated costs of demonstration projects such as this (\$500 million or more per 500-megawatt plant) are not excessive compared with federal expenditure on, for example, the US Navy's Virginia class of submarines (about \$2.5 billion each) or H1N1 influenza-virus vaccine supplies (more than \$1 billion). The federal government spends \$60 billion a year on public infrastructure such as roads and water projects (with state and local governments contributing more than twice that sum).

Money is not the obstacle.

For 15 years, the United States has assumed, naively, that putting a price on carbon-based emissions would be politically achievable and that the resultant market forces would drive innovation. This was a simplistic approach. Both market pull and technology push are needed. In recent months, the picture in Washington has darkened. The United States will have either no price on carbon for some time to come or a price that is too low to drive the innovations needed to transform the nation's energy system. Rather than wringing their hands, policy-makers must now effect the revolution outlined here. ■

John Alic is an independent consultant in Avon, North Carolina 27915, USA. **Daniel Sarewitz** is co-director of the Consortium for Science, Policy and Outcomes at Arizona State University and is based in Washington DC, USA. **Charles Weiss** is a professor at the School of Foreign Service, Georgetown University, Washington DC 20057, USA. **William Bonvillian** directs the Massachusetts Institute of Technology Washington Office, Washington DC 20002, USA. e-mail: john_alic@earthlink.net

"Both market pull and technology push are needed. Policy-makers must now effect the revolution outlined here."

1. Weiss, C. & Bonvillian, W. B. *Structuring an Energy Technology Revolution* (MIT Press, 2009).
2. Consortium for Science, Policy and Outcomes, and Clean Air Task Force. *Innovation Policy for Climate Change* (CSPO and CATF, 2009).
3. Bonvillian, W. B. & Weiss, C. *Innovations* 289–300 (2009).
4. Vane, H. R. & Mulhearn, C. J. *Econ. Perspect.* **23**, 123 (2009).
5. Alic, J. A. *Trillions for Military Technology: How the Pentagon Innovates and Why it Costs So Much* (Palgrave Macmillan, 2007).
6. Birkler, J. L. et al. *Development and Production Cost Estimating Relationships for Aircraft Turbine Engines* 7 (RAND, 1982).
7. Younossi, O. et al. *Military Jet Engine Acquisition* 66–68 (RAND, 2002).

BOOKS & ARTS

Does diversity always grow?

Samir Okasha is intrigued by a proposed universal law of biology: that complexity inevitably increases in the absence of other evolutionary forces.

Biology's First Law: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems

by Daniel W. McShea and Robert N. Brandon
University of Chicago Press: 2010.
184 pp. \$20, £13

The nineteenth-century astronomer John Herschel is said to have dismissed Charles Darwin's theory of evolution as "the law of higgledy-piggledy". Whether he was referring to the role of chance in natural selection or the lack of definite predictions due to it, Herschel's criticism was misplaced: the selective preservation of the fittest variants makes a population more homogeneous, not less. Herschel's epithet does apply neatly to the 'zero-force evolutionary law' that palaeobiologist Daniel McShea and philosopher of science Robert Brandon propose in *Biology's First Law*.

McShea and Brandon state that diversity and complexity tend to increase over time in biological systems. It is, the authors argue, a universal law, applicable to all taxa, at all hierarchical levels and at all times. They use the analogy of Newton's law of inertia — just as it tells us that a body will move with a constant velocity if no forces act on it, this zero-force evolutionary law seeks to capture how a biological system will behave in the absence of other influences. Although the trend they describe may not manifest itself in cases when it is counteracted by constraints, it provides the background against which other evolutionary pressures should be understood, the authors contend.

McShea and Brandon use a standard definition of diversity: the amount of variation in a biological system. A taxon that contains many species, or a species that has many forms, is more diverse than one that has few. But the authors adopt a simplified measure of complexity that considers only the degree of differentiation among the parts of a biological system, not the various functions of those parts. Complexity thus becomes a matter of having many part types, irrespective of what the parts do or how they are organized. The authors argue persuasively that their simpler definition of complexity is more scientifically



The eyes of cavefish vary between individuals despite never being used.

useful than the traditional one, because function is hard to quantify.

Diversity and complexity can be assessed at any level of the biological hierarchy — in clades, species, subspecies, populations and between and within organisms. But diversity and complexity are the same thing, the authors say, when viewed from adjacent levels. For example, an organism with a great diversity of cell types is also a complex organism. Thus, diversity at one level of the hierarchy equates to complexity one level higher. Both diversity and complexity will increase over time through the accumulation of mutations, they suggest.

McShea and Brandon do not claim that their law represents a wholly new evolutionary principle, rather that it is a unifying one. The tendency for increasing diversity has been recognized previously in specific situations. For example, molecular geneticists know that, in the absence of selection, populations will diverge genetically as neutral mutations accumulate. And evolutionary biologists have noticed that tissues and organs that are not subject to selection, such as the eyes of cave-dwelling fish, often show more variation between individuals. The authors aim to encompass these various findings in a single

theory that covers all of the fields in which the principle has been seen — in molecular biology, population genetics, phylogenetics, palaeobiology and elsewhere. They make a good case for their argument that a single principle is at work.

Their theory suggests new research questions, such as whether the tendency for diversity to increase will usually be overcome by natural selection, and it advances our philosophical understanding of evolution. The law also makes testable predictions: for example, that diversity and complexity will increase fastest in ecological circumstances and taxa where selection is weak.

Biology's First Law is an original and unusual book. A hybrid of theoretical biology and philosophy of science, it addresses both conceptual and empirical problems. If there is a lacuna, it is that the authors do not attempt a mathematical formulation of their law, claiming only that it is reducible to probability theory. Such a formulation is essential if we are to investigate and integrate the law with other theories of evolutionary dynamics. Nevertheless, it is a thought-provoking study. ■

Samir Okasha is a professor of the philosophy of science in the Department of Philosophy, University of Bristol, 9 Woodland Road, Bristol BS8 1TB, UK.
e-mail: samir.okasha@bristol.ac.uk

"McShea and Brandon's law does not represent a wholly new evolutionary principle, rather a unifying one."

Silk evolution untangled

The complex proteins that make up spider silks, which are extruded threads that have evolved to work in the dry state, are easily studied and provide us with windows on to protein folding, synthetic biology and nanomaterials. Silks have diversified through selection over millennia to produce thousands of types, and are superbly fit for their purposes.

Alas, silks form only a small part of Leslie Brunetta and Catherine Craig's book. *Spider Silks* focuses mainly on spider ecology and web evolution. It is therefore appropriate for the engineer, materials scientist or chemist who is attracted to silk research but needs a grounding in biological processes. Those seeking detailed insight into the silks themselves will be disappointed. The authors — science writer Brunetta and biologist Craig — briefly discuss the variety of silks and their molecular structures and functions. But they quickly bury that information in explanations of mechanisms of inheritance and arachnid ecology.

The book's limited focus is a missed opportunity because the evolution of web materials and structures cannot be disentangled. Although webs are the obvious signature of a spider's ecology, silks are the key to the success of this diverse taxon. By allowing the efficient construction of a variety of lightweight nets, silks have enabled spiders to inhabit numerous ecological niches across a competitive landscape that also features huge insect diversity.

The earthbound snares of early spiders may have caused some insects to evolve wings and take to the air to escape them. The book introduces both fossil spiders and living fossils, the abdominal segmentation and silk spinning spools of which resemble those of ancestors that left

the water to roam the land some 400 million years ago. Early spiders used thin films of flimsy silk filaments to build shelters or diving bells to survive flooding. An arms race between insect prey and arachnid predator selected for more complex spider behaviour and more functional silks.

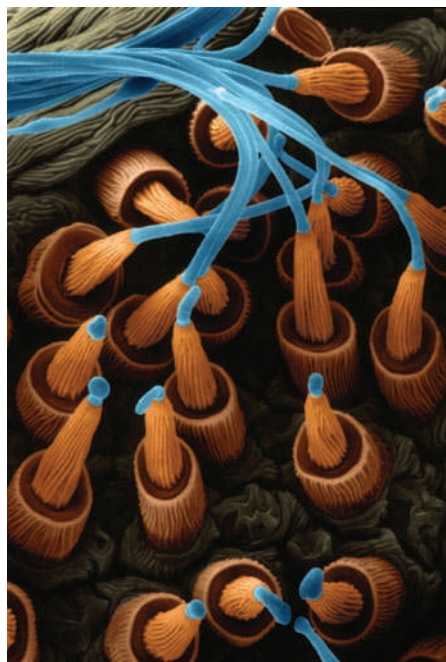
With its spider focus, the book does not discuss silks made by other taxa, such as moths and bees. Silk has evolved several times independently, yet it shares many traits — as well as diversity — between different organisms. Spider silks, which stem from one ancestor, show the greatest variation within one taxon in both chemical composition and production methods. Comparison of spider and other silks tells us about the molecular processes of biopolymer evolution — and of proteins in particular.

Silks consist largely of fibrous proteins. Their complex, long amino-acid sequences hold the keys to silk's remarkable properties:

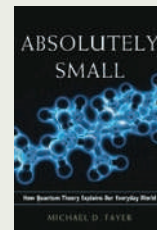
its ability to be stored for months, its rapid transformation from liquid to solid and the strength and flexibility of the finished fibre that often outperforms man-made materials. These qualities of silk, combined with biodegradability and biocompatibility, make it of interest to both biologists and clinicians — in tissue engineering and medical implants, for example. Although it is a one-sided account, *Spider Silks* shows that, with so many applications, silks have as bright a future as commercial mulberry silks had a past.

Fritz Vollrath is a senior researcher and visiting professor at the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK, and head of the Oxford Silk Group.
e-mail: fritz.vollrath@zoo.ox.ac.uk

Spider Silk: Evolution and 400 Million Years of Spinning, Waiting, Snagging, and Mating
by Leslie Brunetta and Catherine L. Craig
Yale University Press: 2010. 248 pp. \$30, £20

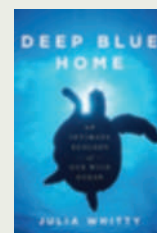


Silk secretion: the key to spider success.



In *Absolutely Small* (McGraw Hill, 2010), chemist Michael Fayer explains the weird microscopic behaviour of matter. Favouring everyday

examples over formulae, he makes quantum mechanics palatable, from wave-particle duality to the uncertainty principle. Beer, soap and fat are the ingredients he chooses to explain how chemical bonds work. With summaries of the chemistry of greenhouse gases, hydrocarbons and the electrical properties of materials, Fayer's book provides a useful overview.



Former scientist and film-maker Julia Whitty relates a personal view of ocean ecology in *Deep Blue Home* (Houghton Mifflin Harcourt, 2010).

Travelling from the Sea of Cortez to Newfoundland and the Galapagos Islands, she describes encounters with whales and turtles, and the work of the scientists who are trying to protect and understand them. While explaining how the world's oceans are unified by deep currents, she also reveals their effect on Earth's climate.



Dwindling fish stocks form the focus of Paul Greenberg's book. *Four Fish* (Allen Lane, 2010) looks at the quartet of species that we

eat most — salmon, sea bass, cod and tuna. The writer travelled from Norway to the South Pacific to see how these animals reach our plates and how they are threatened by overfishing. Posing pragmatic questions such as how to find sustainable seafood on a supermarket shelf, he weaves together the biology of sea fish, fisheries and conservation issues. He ends with a proposal for choosing the species that we can farm without disrupting wild ecosystems.



Q&A: Tod Machover on personal music

The inventor and composer, whose group at the Massachusetts Institute of Technology's Media Lab developed the technology behind *Guitar Hero*, has electronically customized instruments for musicians from Prince to Yo-Yo Ma. As Machover prepares for the world premiere of his robotic opera *Death and the Powers* in Monaco in September, he explains how his interactive performance techniques might lead to personalized therapies.

What is the opera about?

Mortality. I know my parents well, but how will their grandchildren know them? What can a dying person pass on to future generations? I researched people such as US entrepreneur and inventor Ray Kurzweil, who is trying to live forever; people who are trying to preserve themselves on ice; and people with locked-in syndrome, who have almost complete paralysis and cannot communicate but are not dead.

The opera tells the story of Simon Powers, a man who wants to remain in the world after he dies. So he invents a system to download everything about himself — his legacy, personality and memories — into his environment. At the end of the first scene Powers disappears, and the stage set becomes the main character.

How do you use robotics to portray Powers after his death?

The system into which Powers has transferred his intelligence has four elements: two kinds of robots, animated bookcases and a musical chandelier. The bookcases have LED displays that look like the spines of thousands of books; as they move and change texture, they allow

Powers to express himself. A swarm of nine 'opera bots' acts as a chorus. They are wireless, intelligent and graceful; white and smooth with lights all over. The other category of robots is cruder. They look like strange animals and react nervously to the characters onstage.

You've described the chandelier as the first use of sonic animatronics. What is it?

It looks like an upside-down Venus fly-trap plant, with wings of rounded metal that are criss-crossed by dozens of Teflon strings that extend about 12 metres when it opens. It is a robotic musical instrument in which the strings are vibrated using electro-acoustic signals. We can control the harmonics and produce rich sounds, like low piano notes, as well as pulsations and vibrations.

How will the actors interact with the stage?

We draw on a new technique called disembodied performance to animate the onstage environment using the offstage voice, breathing and arm movements of James Maddalena, who plays Powers. One big question is how much he will be able to control while he is singing and reacting to

other performers. We also want to measure his muscle contractions, pulse, posture and other things he can't consciously control.

Why did you choose the medium of opera?

With the world being fragmented and busy, there is something powerful about going to a public place and experiencing something that is never going to happen again. Opera also attracts me because it is a hybrid, pragmatic form that has always sought to pull every possible discipline into its orbit — much like the MIT Media Lab encompasses so many areas of human exploration. Opera may seem to be a traditional form, but it is ripe for, and open to, reinvention.

How has opera inspired your research?

The first opera I did was *VALIS*, based on Philip K. Dick's novel. It was staged in the lobby of the Pompidou Centre in Paris in 1987 with almost no scenery. I wanted to rehearse the pit orchestra like a rock group, for hours every day, but I couldn't afford that much time. So I reduced the orchestra to two musicians. And that is when I invented hyper-instruments — acoustic instruments that are measured and sensed electronically in ways that affect the sound, to maximize the variety of sound coming from each one. These instruments, based on commercially available technology, eventually led to the video games *Guitar Hero* and *Rock Band*.

How do you think music affects us?

I am optimistic that music can play a part in combating disease — both physical and mental. Music therapy is an old profession, but we are just beginning to understand the science that will allow treatments to be shaped to an individual. I am adapting my hyper-instruments so that they can be adjusted to the skills or limitations of each person.

I also believe that in 50 years we will be able to customize each piece of music, and the way that it is played, to produce a precise emotional and physical reaction — hopefully a beneficial one — that will be unique for each of us, in each situation, in each moment. This 'personal music' has never existed before, but I intend to help lay the groundwork for it.

Interview by **Jascha Hoffman**, a writer based in San Francisco, California.

Death and the Powers

Composed by Tod Machover. Libretto by Robert Pinsky. Directed by Diane Paulus 25–26 September 2010, Opéra de Monte-Carlo, Monaco. See go.nature.com/9bQ6B7

NEWS & VIEWS

COSMOLOGY FORUM

Is dark energy really a mystery?

The Universe is expanding. And the expansion seems to be speeding up. To account for that acceleration, a mysterious factor, 'dark energy', is often invoked. A contrary opinion — that this factor isn't at all mysterious — is here given voice, along with counter-arguments against that view.

**No it isn't —
Eugenio Bianchi
and Carlo Rovelli**

Is cosmic acceleration¹ such a great problem? Many commentators have argued that it is, and that the explanation is to be found in invoking a mysterious substance, dark energy, that as yet has no theoretical underpinning^{2,3}.

We disagree. An explanation is to hand and has been for many years. Cosmic acceleration is predicted and simply described by the theory of general relativity, together with a non-vanishing cosmological constant, Λ . Our case, presented here in summary, is made in detail in a paper⁴ posted on the arXiv preprint server.

The Λ cold dark matter (Λ CDM) cosmological model assumes the presence of Λ and is "almost universally accepted by cosmologists as the best description of the present data"⁵. Three objections to Λ are commonly presented, and nourish the 'mystery'. We argue that there is confusion — historical or conceptual — in each of them.

The first objection is known as 'Einstein's blunder'. Allegedly, Λ was rejected by general relativists, and indeed by Einstein himself, who first considered it but later called it his "greatest blunder". But Einstein's 'blunder' was not Λ . It was failing to see that — with or without Λ — the Universe isn't static in his theory of general relativity, thereby missing an easy prediction of the cosmic expansion (Fig. 1) before its discovery. Λ is not an appendage to Einstein's theory added to account for observations: it is an integral and natural part of it. Its nature and scale are no more or less mysterious than any of the several other constants in our fundamental theories.

The second objection is termed the 'coincidence problem'. Data indicate that we happen to live in a 'short' phase of cosmic

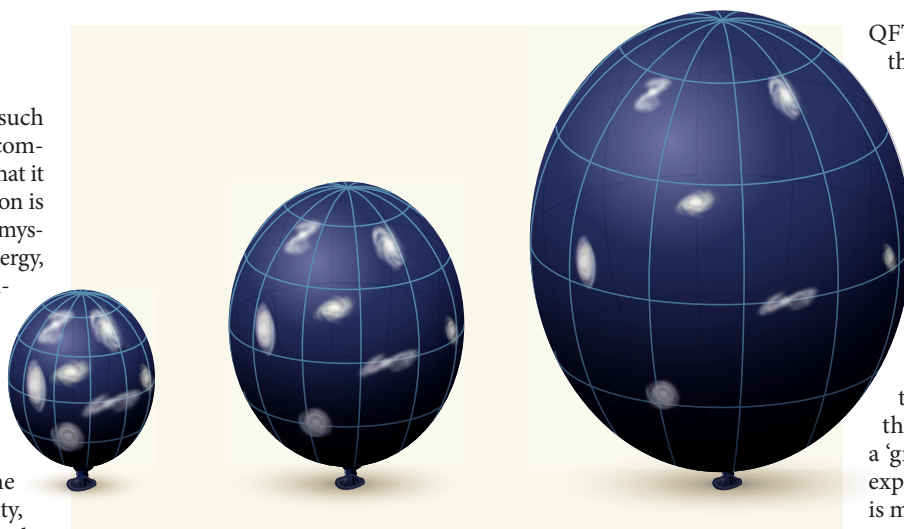


Figure 1 | Cosmic expansion. As the Universe expands, space stretches much like the surface of an inflating balloon, and galaxies get farther and farther apart. Data indicate that the rate of this expansion is accelerating.

history, during which the contributions from matter and Λ to the cosmic dynamics are comparable in magnitude. Such an 'unlikely coincidence' is presented as an argument against the Λ CDM hypothesis. But if the ratio of these contributions as a function of cosmic time is properly considered on a linear rather than a logarithmic scale, it can be seen that such a 'short' phase lasts for half the life of the Universe, and there is no 'unlikely' coincidence. In any case, we should not assume that we live in a fully random place or time in the Universe, as the coincidence-problem objection presupposes. The density around us, for instance, is very far from the average cosmic density.

The third objection concerns 'vacuum energy'. Quantum field theory (QFT) seems to predict a vacuum energy that adds to the cosmological force due to Λ — just as radiative corrections affect the charge of the electron. But this hypothetical contribution to Λ is much larger than the observed Λ . The discrepancy is an open puzzle in QFT in the presence of gravity^{6,7}. But it is a conceptual mistake to confuse Λ with QFT's vacuum energy. Λ cannot be reduced to the ill-understood effect of

QFT's vacuum energy — or that of any other mysterious substance. Λ is a sort of 'zero-point curvature'; it is a repulsive force caused by the intrinsic dynamics of space-time.

Tests on the Λ CDM model must continue and alternative ideas must be explored. But it is our opinion — and that of many relativists — that saying dark energy is a 'great mystery', for a force explained by current theory, is misleading. It is especially wrong to talk about a 'substance'. It is like attributing the force that pushes us out of a turning merry-go-round to a 'mysterious substance'.

Eugenio Bianchi and Carlo Rovelli are at the Centre de Physique Théorique de Luminy, Case 907, Luminy, Marseille F-13288, France.
e-mails: rovelli@cpt.univ-mrs.fr;
bianchi@cpt.univ-mrs.fr

**Yes it is —
Rocky Kolb**

The Λ CDM is the most complete, predictive and successful cosmological model ever devised, capable of accounting for an enormous number of astronomical observations. At present, there are no observations discrepant with the Λ CDM model.

But the success of Λ CDM comes at a price. In the model, only about 5% of the total mass-energy of the Universe is observed and understood, and 95% of the Universe is dark. The dark side includes 25% of the total mass-energy in the form of dark matter binding together galaxies and other large-scale structures, and 70% in the form of dark energy driving galaxies apart in an accelerating cosmic expansion.

Cosmologists usually refer to dark matter and dark energy as cosmic mysteries. Bianchi and Rovelli⁴ argue that dark energy can be explained by invoking a new constant of

nature, a cosmological constant. They assert that this is a simple, acceptable, non-mysterious explanation for dark energy. I disagree. In my opinion, a cosmological constant qualifies as a mystery in the non-theological sense of the word⁸: “Something not understood or beyond understanding.”

Einstein's cosmological constant Λ is the simplest explanation for dark energy: it adequately fits the data, and there is no reason to exclude it. But the magnitude of Λ necessary to explain the observations places it far “beyond [our] understanding”.

If the cosmological constant is the explanation for dark energy, Λ must be about $(10^{28} \text{ cm})^{-2}$. The length 10^{28} cm is absurdly large, and cannot at present be related to any

other known or expected length scale in nature. Attempts to explain this new length scale fail by many, many orders of magnitude.

We must demand more of cosmology than just piling on components or constants to a model to reproduce observations. Otherwise, we would still happily be adding epicycles to the Ptolemaic model of planetary motion. Cosmological models, along with their constants and components, must be grounded in laws of nature that we understand. The magnitude of the cosmological constant cannot currently be explained by any physics we know. Until it is, it is a mystery.

Recalling the warning of astrophysicist Tommy Gold (personal communication), “for every complicated physical phenomenon there is a

simple, wrong explanation”, it would be a mistake to be satisfied with the cosmological constant just because it is a simple explanation. ■

Rocky Kolb is in the Department of Astronomy and Astrophysics, University of Chicago, Chicago, Illinois 60637, USA.

e-mail: rocky.kolb@uchicago.edu

1. Carroll, S. M. *Living Rev. Rel.* **4**, 1–56 (2001).
2. Calder, L. & Lahav, O. *Phys. World* **23** (June), 32–37 (2010).
3. Tyson, J. A. *Nature* **464**, 172–173 (2010).
4. Bianchi, E. & Rovelli, C. Preprint at <http://arxiv.org/abs/1002.3966> (2010).
5. Lahav, O. & Liddle, A. R. Preprint at <http://arxiv.org/abs/1002.3488> (2010).
6. Wald, R. M. *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (Univ. Chicago Press, 1994).
7. Rovelli, C. *Quantum Gravity* (Cambridge Univ. Press, 2004).
8. www.merriam-webster.com/dictionary/mystery

CONSERVATION SCIENCE

Trade-in to trade-up

Peter Kareiva

Nature reserves and protected areas enjoy sacred status in conservation — which translates into a ‘do not touch’ attitude. But selling off some of the less worthy of them would pay conservation dividends.

Protected areas set aside from major human activities and managed for biodiversity are the foundation of modern conservation. Until now, no conservationist would have considered trading them in. Yet trading-in and trading-up is exactly what Fuller and colleagues recommend for Australia's protected areas, as described on page 365 of this issue¹.

Fuller *et al.* estimate that, by selling off 70 of Australia's nearly 7,000 protected areas, the government could raise Aus\$20.6 billion

(US\$17.4 billion), which could then be reinvested to achieve far more conservation elsewhere (Fig. 1). The idea is to sell off those protected areas that yield the lowest conservation value per assessed land value, and then reinvest the funds in lands that generate the highest conservation value per dollar spent. If these transactions were actually completed, the authors conclude, Australia could achieve a tenfold increase in the total area under conservation protection and a threefold increase

in the diversity of vegetation types under protection. Real-world application of a return-on-investment analysis might be the best thing that could happen to conservation in Australia and, by extrapolation, elsewhere in the world.

Historically, the establishment of protected areas has been anything but analytical or efficient. Before conservation emerged as a science, protected areas tended to be located to satisfy the tourist industry or the wishes of a wealthy few, or for convenience. They were often also sited so as to become bargaining chips between large corporate landowners and national governments; rarely were conservation goals a factor². One of the greatest contributions of the new field of conservation science has been to replace ad hoc establishment of protected areas with networks of nature reserves that are sited using computer-based planning tools³. It has now become clear that the use of data, quantifiable objectives and spatial-optimization programmes provides an opportunity to protect much more nature at far less public expense⁴.

However, the idea of applying a return-on-investment approach to the design of networks of protected areas has yet to gain full traction in the messy world of non-governmental organizations involved in conservation, or that of national governments. When protected areas are established, it is usually because a conservation group has lobbied for the budget allocation to make it happen. There is no need for economic analysis in such a case, because the problem of getting enough money is seen as a lobbying effort, and when money or land becomes available, opportunity is what counts more than any return-on-investment assessment. The possibility of selling existing protected areas totally changes the nature of the discussion because there is no need to lobby for an opportunity for new protected areas — the opportunity exists by virtue of the funds generated from selling low-return nature reserves.

Doubtless, the very thought of such a ‘trade-in and trade-up’ scheme for enhancing the efficiency of conservation will cause many



Figure 1 | *Acacia harpophylla* — in need of protection. This vegetation type is down to less than 15% of its original extent, and is an example of habitat that would benefit from the scheme of Fuller *et al.*¹.

conservationists to shudder. The idea that a protected area might not be secured in perpetuity will be unsettling, as such a precedent may make it too easy for governments to revoke systems of protected areas in times of economic stress. Also, those 70 protected areas in Australia — the bottom 1% — tend to be in locations where land prices are high, which often means that they are near urban centres. Protected areas near high densities of human populations may not yield a high conservation return per dollar cost, but their public accessibility and visibility could be essential to maintaining public support for conservation.

Nature reserves and protected areas enjoy sacred status in conservation, which translates into a 'do not touch' attitude. Selling off inefficient protected areas for conservation cash opens the door to an entirely different, and perhaps more fruitful, approach to nature conservation. When protected areas are rigid, in perpetuity and established almost without regard to cost–benefit analyses, they essentially become monuments to one special-interest group — conservationists.

However, if protected areas were subject to the same cost–benefit analyses as any public-works project, they might win broader public support. In addition, for a world experiencing climate change with the attendant shifts in biomes and species, a dynamic approach that encourages continuously rethinking the management status of lands and waters may be essential to maintain any sort of functioning system of protected areas. What good is a floodplain reserve if a river changes its course, or a coastal sand-dune protected area that ends

up being submerged by the rising sea? Conservation planning and priority setting need to evolve if they are to deal with today's economic realities and the rapidly changing world.

To date, return-on-investment analyses in conservation have focused on land prices and management costs, with biodiversity representing the 'return'⁵. But protected areas also provide substantial economic payouts in the form of water supply, recreation, water quality and climate regulation. Now that conservation scientists have clarified the importance of efficiency in spending limited conservation dollars, it is a small step to also factor in the economic benefits of protected areas.

To some, nature is sacred, and costing-out its value demeans and makes venal what is inspiring and in many regards priceless⁶. But as long as public money is spent on protected areas and conservation, return-on-investment analyses, including the economic returns, may be the greatest hope for conservation in an economically stressed world⁷.

Peter Kareiva is at The Nature Conservancy, 4245 North Fairfax Drive, Arlington, Virginia 22003, USA.
e-mail: pkareiva@tnc.org

1. Fuller, R. A. *et al.* *Nature* **466**, 365–367 (2010).
2. Fairfax, S. K., Gwin, L., King, M. A., Raymond, L. & Watt, L. A. *Buying Nature: The Limits of Land Acquisition as a Conservation Strategy, 1780–2004* (MIT Press, 2005).
3. Murdoch, W. *et al.* *Biol. Conserv.* **139**, 375–388 (2007).
4. Bode, M. *et al.* *Proc. Natl Acad. Sci. USA* **105**, 6498–6501 (2008).
5. Naidoo, R. *et al.* *Trends Ecol. Evol.* **21**, 681–687 (2006).
6. McCauley, D. J. *Nature* **443**, 27–28 (2006).
7. Kareiva, P. & Marvier, M. *Conservation Science: Balancing the Needs of People and Nature* (Roberts, Greenwood Village, Colorado, in the press).

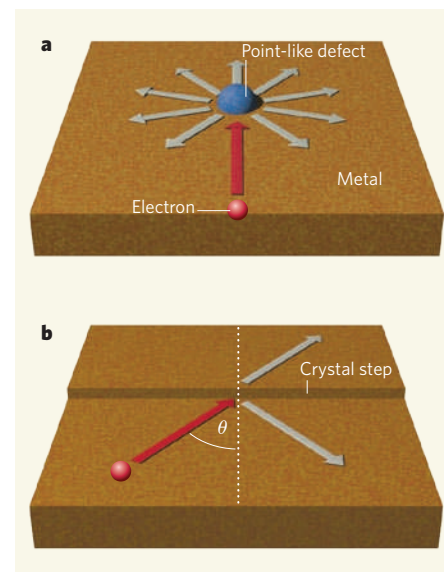


Figure 1 | Electron scattering from point-like and line-like defects. **a**, A point-like impurity can scatter an electron in any direction. **b**, A line-like defect such as a crystal step can scatter it in only two independent directions for a given angle of incidence θ . Yazdani and colleagues⁴ have studied the implications of this and other concepts for electron transmission through naturally occurring crystal steps on the surface of elemental antimony.

a given momentum $\mathbf{k}$, there are two possible independent spin orientations (states), spin-up and spin-down (Fig. 2a, overleaf). In a helical metal, there is only one state for a given $\mathbf{k}$, with the spin being oriented perpendicularly to $\mathbf{k}$ (Fig. 2b). This spin orientation is caused by an interaction between the electron's motion and its spin (spin–orbit coupling), and is unique to topological insulators.

The second concept that underlies electron behaviour in helical metals, but that also applies to ordinary metals, is time-reversal symmetry — that is, under a time-reversal transformation, all microscopic physical processes remain unchanged, but are just running backwards. And indeed, the electron configurations in both ordinary and helical metals are invariant under such a transformation: simultaneously reversing the direction of travel (with $\mathbf{k}$ going to $-\mathbf{k}$) and the spin orientation of all electrons in the material leaves the energy-band structure of the system unchanged (Fig. 2).

These two concepts are sufficient to account for the absence of backscattering in helical metals. In a backscattering process, the electron's direction of travel is reversed ($\mathbf{k}$ goes to $-\mathbf{k}$). If the process is caused by a non-magnetic defect, the electron's spin state remains unaffected (this follows from the law of conservation of angular momentum). In an ordinary metal (Fig. 2a), there is a state with the requisite spin available at $-\mathbf{k}$, and so the electron can backscatter with ease. In a helical metal, however, the only spin state available at $-\mathbf{k}$ is exactly opposite to the one required (Fig. 2b), and backscattering therefore cannot occur.

SOLID-STATE PHYSICS

U-turns strictly prohibited

Marcel Franz

According to theory, electrons on the surface of a topological insulator are not allowed to make U-turns. This notion, and some of its main consequences, has now been tested experimentally.

The flow of electric current through a metallic wire is inevitably degraded by the wire's resistance. At low temperatures, the resistivity of a metal is dominated by the scattering of electrons off impurities and imperfections in the metal's crystal lattice. When an electron collides with an impurity (Fig. 1a), it scatters in a random direction, and this process diminishes its contribution to the overall current. Backscattering processes, in which a collision causes the electron to make a U-turn, are especially effective in degrading the current flow. A remarkable property of the recently discovered 'helical metals' (also known as metallic surface states), which form on the surfaces of special insulators known as topological insulators^{1–3}, is

that backscattering is strictly prohibited. This prohibition leads to a host of novel physical phenomena (such as a much reduced resistivity) that have been predicted theoretically but thus far largely unexplored experimentally. On page 343 of this issue, Yazdani and colleagues⁴ put some of these theoretical ideas to the test by measuring the ability of electrons on the surface of elemental antimony to penetrate barriers formed by naturally occurring crystal steps.

The main feature that makes electron behaviour in helical metals special is the relation between the electron's spin and its momentum. In an ordinary metal, which can be thought of as a sea of free, non-interacting electrons, for



50 YEARS AGO

Very few microbiologists would dispute that bacterial taxonomy is in a very confused state ... As if this were not enough, there has been added to it a cavalier indifference to the internationally accepted conventions for nomenclature on the part of bacteriologists themselves and an almost complete absence of type cultures ... For historical reasons the taxonomy of bacteria especially has been carried on as isolated fragments of other disciplines such as medicine or dairying, with the result that the same micro-organism has been known by a different name in each field, and not infrequently their identity has never been realized. Many taxonomic schemes have been devised but have been only cherished by their authors, while others such as Bergey's scheme have been tolerated with much vexation and disappointment for want of something better.

From *Nature* 16 July 1960.

100 YEARS AGO

Mr. H. O. Barnard states, as the result of personal observation, that the alleged partiality of cobras for music is a myth. "The sole effect, so far as I could see, was to arouse their curiosity, as they would project their heads out of their holes equally well for any kind of noise, from the shrill piping affected by snake-charmers down to the tinkling noise made by dragging a chain past their dwelling, or even that made by light and repeated tappings with a switch close to their holes. It would appear, however, that the tone must be high, as grave sounds, such as tom-tom beating or deep notes from a flute, had no effect upon them." Mr. Barnard likewise confirms the observations, made in the London Zoological Gardens, as to the absence of a "fascinating" influence of serpents on birds.

From *Nature* 14 July 1910.

Although electrons in helical metals cannot backscatter, nothing prevents them from scattering in all other directions if the defects are point-like (Fig. 1a). However, the probability of near U-turns is much reduced compared with ordinary metals, and this turns out to have major consequences for 'electron localization' properties^{1–3}. But when electrons are scattered by a crystal step, as in the experiment performed by Yazdani and colleagues⁴, the situation is very different. Because of conservation of energy and of the component of the momentum in the direction along the step, electrons can ultimately do only two things (Fig. 1b): they can either pass right through the step or reflect specularly from it, just as light rays would (in specular reflection, the incidence and reflection angles of a light ray are the same). Crucially, if electrons travel at a right angle to the wall of a crystal step in a helical metal, they will certainly pass through — they won't be reflected because backscattering is not allowed in helical metals. Electrons approaching the step at a small incidence angle will have a high probability of making their way through⁵. Remarkably, this is true whatever the size of the step.

The above considerations form the physical basis for the prediction that crystal steps should affect electrons in a helical metal much more weakly than those in ordinary metals. This prediction is indeed borne out in Yazdani and colleagues' experiment. By combining state-of-the-art scanning tunnelling spectroscopy with a simple physical model for the helical metal on the surface of an antimony crystal, the authors deduce a probability for electron transmission through a crystal step of about 35%, a significant enhancement compared with the surface states in ordinary metals such as copper or gold. Previous studies⁶ showed that steps in ordinary metals completely absorb or reflect the electrons, with almost no transmission observed.

The antimony crystals used in this study⁴ are bulk metals, rather than insulators, and this causes some 'leakage' of the surface electrons into the bulk during the collision with a step (about 23%). Also, antimony's surface energy-band structure is more complicated than the ideal helical metal shown in Figure 2b, and thus allows for more near-U-turns than can be explained by the arguments presented above. These factors suggest that the electron-transmission probability through the steps can be further increased by using a material such as Bi_xSb_{1–x} (bismuth–antimony alloy), which for a range of x is a bulk topological insulator, or Bi₂Te₃ (bismuth telluride), which can be tuned to become an insulator in the bulk and has a surface energy-band structure that is much closer to that depicted in Figure 2b. Indeed, recent experiments using Bi₂Te₃ surfaces⁷ show some of these features.

The metallic surface states of topological insulators hold promise for investigating various exotic physical phenomena as well as potential practical applications^{8,9}, such as

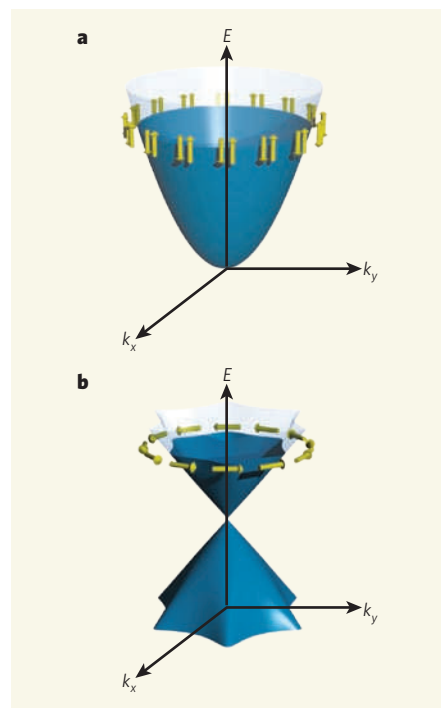


Figure 2 | Energy-band structure and spin orientation in normal and helical metals. **a**, In a normal metal, electrons occupy energy (E) levels up to a maximum level, the Fermi level, and for a given momentum $\mathbf{k} = (k_x, k_y)$ they have two possible spin orientations (spin-up and spin-down). **b**, In a helical metal, the energy levels form two cone-shaped bands that meet at their tips, and for a given $\mathbf{k}$ there is only one spin orientation.

dissipationless switching of magnetic moments, which could be of use in the data-recording industry. The underlying theoretical ideas are based on idealized, clean, perfectly flat surfaces and a somewhat vague notion of 'topological protection', according to which small changes to the ideal conditions that allow time-reversal symmetry will not alter the outcomes. It is encouraging to see that experiments such as that of Yazdani and colleagues are beginning to test these notions in real materials, and that their findings seem to support theoretical predictions.

Marcel Franz is in the Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. e-mail: franz@physics.ubc.ca

- Moore, J. E. *Nature* **464**, 194–198 (2010).
- Qi, X.-L. & Zhang, S.-C. *Phys. Today* **63**, 33–38 (2010).
- Hasan, M. Z. & Kane, C. L. Preprint at <http://arxiv.org/abs/1002.3895> (2010).
- Seo, J., Roushan, P., Beidenkopf, H., Hor, Y. S., Cava, R. J. & Yazdani, A. *Nature* **466**, 343–346 (2010).
- Biswas, R. R. & Balatsky, A. V. Preprint at <http://arxiv.org/abs/0912.4477> (2009).
- Crommie, M. F., Lutz, C. P. & Eigler, D. M. *Nature* **363**, 524–527 (1993).
- Alpichshev, Z. et al. *Phys. Rev. Lett.* **104**, 016401 (2010).
- Yokoyama, T., Tanaka, Y. & Nagaosa, N. *Phys. Rev. B* **81**, 121401(R) (2010).
- Garate, I. & Franz, M. *Phys. Rev. Lett.* **104**, 146802 (2010).

See also News Feature, page 310.

PROTEIN SYNTHESIS

Translocation in slow motion

Måns Ehrenberg

Time-resolved electron microscopy can capture structural changes in active macromolecular complexes, but detailed imaging is essential. The dynamics of one step in protein synthesis has been deduced from two million images.

Protein synthesis is coordinated by cellular components called ribosomes. They bring together messenger RNA, which carries the genetic information, and transfer RNAs, which carry the amino-acid building blocks and translate the mRNA code. Translation is a dynamic process: both mRNA and tRNAs move along the ribosome — a process called translocation — as amino acids are added to the nascent polypeptide chain. On the basis of around two million images captured by time-resolved cryo-electron microscopy, Fischer *et al.*¹ now reveal the dynamics of tRNA and ribosome action during translocation (page 329 of this issue).

Translation is divided into four main steps: initiation, peptide elongation, termination of translation and recycling of the ribosomes for a new round of initiation. In the past decade, great progress has been made in understanding the details of these steps. For instance, high-resolution crystallography² has clarified the mechanism by which the ribosome mediates formation of the peptide bond between amino acids^{3,4}. It has also elucidated the way in which mRNA information is accurately decoded by tRNAs during elongation⁵, as well as protein-mediated mRNA decoding during termination^{6–8}. However, crystallographic techniques are less useful for studying translocation

because this process involves large changes in the conformation of the constituents, including movements of the small (30S) ribosomal subunit relative to its large (50S) subunit.

To investigate the conformational dynamics of translocation, Fischer *et al.*¹ monitored the kinetics of this process using time-resolved cryo-electron microscopy (cryo-EM) and computational sorting of the cryo-EM images. They studied the ribosome of the bacterium *Escherichia coli* — known for its fast, accurate translocation — and present a ‘movie’ of the sequence of events and the structural nature of the conformational changes that take the ribosome from a pre- to a post-translocation state.

Translocation involves three main sites on the ribosome — A, P and E — that are involved in the sequential binding of tRNA (Fig. 1a). In bacteria, elongation factor G (EF-G) in complex with the nucleotide GTP binds to the pre-translocation ribosome. This stabilizes a ribosome conformation in which the 30S subunit undergoes a ratchet-like movement relative to the 50S subunit⁹. At this point, the tRNA to which the nascent polypeptide chain is attached (peptidyl-tRNA) is in a hybrid A/P state and the outgoing (deacylated) tRNA, from which the polypeptide chain has been transferred to the peptidyl-tRNA, is in a hybrid P/E state¹⁰ (Fig. 1b).

Rapid hydrolysis of GTP to GDP by ribosome-bound EF-G finally brings the ratcheted ribosome into the post-translocation state. The peptidyl-tRNA is now in the P state and the deacylated tRNA in the E state, with the empty A site ready to admit the next aminoacyl-tRNA to further elongate the nascent chain¹¹ (Fig. 1c). Although these steps are well known, the molecular mechanism by which the ribosome is rapidly converted from the ratcheted state to the post-translocation state has remained unclear.

Kinetic analysis by cryo-EM generally has too low a time resolution to follow the dynamics of EF-G-catalysed forward translocation. But in the absence of EF-G, spontaneous translocation occurs at a much lower rate in both the forward and backward (retro-translocation) directions. So Fischer and co-workers set out to investigate the kinetics of spontaneous retro-translocation.

The authors prepared a post-translocation ribosomal complex in which the P site contains a peptidyl-tRNA carrying the dipeptide fMet-Val (fMet-Val-tRNA^{Val}) that cannot retro-translocate when the E site is empty (Fig. 1d). They then triggered spontaneous retro-translocation by adding deacylated tRNA^{fMet} to the E site (Fig. 1e). At 20 °C, retro-translocation brought fMet-Val-tRNA^{Val} backward into the A site and tRNA^{fMet} into the P site. The system eventually reached equilibrium, with 20% of the ribosomes in the post-translocation state and 80% in the pre-translocation state (Fig. 1f).

For cryo-EM imaging, Fischer *et al.* withdrew samples from the incubation mixture at various times. Using hierarchical computational sorting of about two million ribosome images, they identified 50 structures corresponding to ribosomal sub-states that were present during

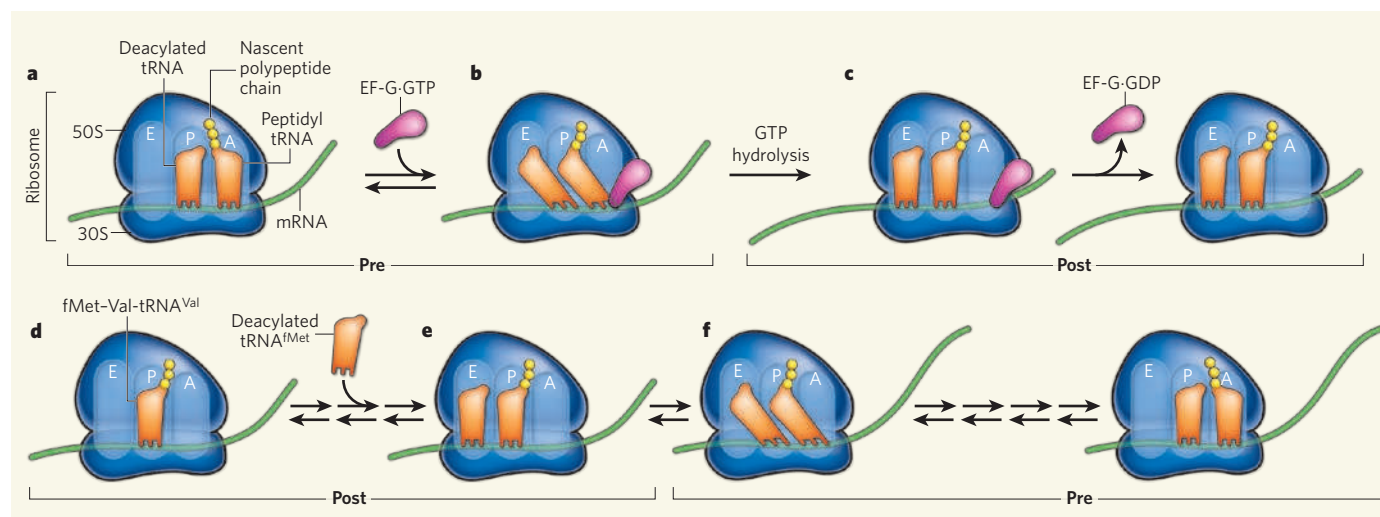


Figure 1 | Movement back and forth along the ribosome. **a,b,** As GTP-bound EF-G binds to the pre-translocation ribosome (Pre), the peptidyl-tRNA shifts from an A state to a hybrid A/P state and the deacylated tRNA shifts from a P/P state to a P/E state. **c,** As the mRNA in the Pre ribosome in **b** translocates by one nucleotide triplet (a codon), the ribosome moves to its post-translocation state (Post), with the tRNAs in the P/P and E/E states, respectively. Along the way, GTP is hydrolysed to GDP, and GDP-bound

EF-G dissociates from the ribosome. **d,e,** Fischer *et al.*¹ report that, in a Post ribosome with fMet-Val-tRNA^{Val} in the P/P state, deacylated tRNA^{fMet} binds to the E site in an E/E state. **f,** In the rate-limiting step of retro-translocation, when the ribosome moves from the Post to the Pre state, the mRNA moves backwards by one codon, the tRNAs originally in P/P and E/E states move to A/P and P/E states and then rapidly to their A/A and P/P states, respectively.

spontaneous retro-translocation. They categorized these structures into eight main classes, five in the pre-translocation state (pre1–pre5) and three in the post-translocation state (post1–post3).

The authors' main criterion for pre- or post-translocation categorization was the position in the 30S subunit of a region in the peptidyl-tRNA called the tRNA anticodon stem loop — when this region was in the A site, the sub-state was categorized as pre-translocation, and when in the P site it was given post-translocation status. The authors note that, irrespective of the incubation time, the relative numbers of ribosomes in the five pre- and three post-translocation states remained constant. They therefore conclude that the sub-states in the two states are all in equilibrium and are kinetically separated by a rate-limiting step.

But how do the two ribosome-bound tRNAs move sequentially through their five pre-translocation states to the rate-limiting translocation event? And why, as the authors find, is the post-translocation peptidyl-tRNA stable in the P site, whereas the E-site tRNA becomes progressively less stably bound on its way to dissociating from the ribosome?

Fischer *et al.*¹ report strong correlations between the various ribosome conformations and the positions of the tRNA molecules, but they also find that any given position of the tRNAs is associated with a broad range of ribosome structures. It therefore seems that there is a complex landscape of diverse ribosomal conformations and tRNA-binding modes that have similar standard free energies and that are separated by low-energy barriers.

When the authors raise the incubation temperature to 37 °C — the optimal temperature for *E. coli* growth — cryo-EM cannot detect distinct ribosome structures. This suggests that, at this temperature, the ribosome is rapidly 'diffusing' between a large number of conformational sub-states. The authors therefore describe the ribosome as a "Brownian machine", in that forward and backward translocations in the absence of EF-G depend on selective sampling of a large set of rapidly interconverting ribosome conformations.

As Fischer and colleagues point out, the main question now is whether EF-G-catalysed forward translocation can also be understood in terms of the ribosome as a Brownian machine, existing in a conformational spectrum of sub-states similar to that seen in spontaneous retro-translocation. Alternatively, it could be that GTP-bound EF-G rapidly drives the ribosome into a ratcheted state with little room for conformational excursions, and that, after GTP hydrolysis, it guides the ribosome to the completion of translocation along a structurally well-defined reaction path. It could, in fact, be that the mechanism behind rapid EF-G-catalysed translocation is pre-organization of the ribosome in a structure that leads from the ground state to a transition state in the translocation reaction with a small loss

in entropy (a small entropy of activation)¹².

The key to understanding how EF-G promotes such rapid and precise translocation therefore remains unclear. Will the key be the conformational fluctuations described by Fischer *et al.*¹, or analysis of a small number of high-resolution structures of the ribosome that clarify the transition state of EF-G-catalysed translocation? Either way, Fischer and colleagues' seminal cryo-EM work has revealed important and intriguing aspects of the interplay between rapidly inter-converting ribosomal conformations and elongation-factor-driven ribosome function.

Måns Ehrenberg is in the Department of Cell and Molecular Biology, BMC, Uppsala University, Uppsala S-75124, Sweden.

e-mail: ehrenberg@xray.bmc.uu.se

1. Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V. & Stark, H. *Nature* **466**, 329–333 (2010).
2. Schmeing, T. M. & Ramakrishnan, V. *Nature* **461**, 1234–1242 (2009).
3. Schmeing, T. M., Huang, K. S., Kitchen, D. E., Strobel, S. A. & Steitz, T. A. *Mol. Cell* **20**, 437–448 (2005).
4. Trobro, S. & Åqvist, J. *Proc. Natl Acad. Sci. USA* **102**, 12395–12400 (2005).
5. Ogle, J. M. & Ramakrishnan, V. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
6. Weixlbaumer, A. *et al. Science* **322**, 953–956 (2008).
7. Laurberg, M. *et al. Nature* **454**, 852–857 (2008).
8. Sund, J., Andér, M. & Åqvist, J. *Nature* **465**, 947–950 (2010).
9. Valle, M. *et al. Cell* **114**, 123–134 (2003).
10. Moazed, D. & Noller, H. *J. Nature* **342**, 142–148 (1989).
11. Rodnina, M. V. *et al. Nature* **385**, 37–41 (1997).
12. Kamerlin, S. C. L. & Warshel, A. *Proteins* **78**, 1339–1375 (2010).

EARTH SCIENCE

Proposal with a ring of diamonds

David A. D. Evans

Plate reconstructions show that ancient eruptions of diamond-bearing rocks occurred consistently above a ring-like region of plume-generation zones deep in Earth's mantle. Do such zones remain stationary?

Diamonds are formed only at great depths within Earth's mantle, enduring there for billions of years until they are brought to the surface in explosive 'kimberlite' volcanic eruptions¹ — named after the spectacularly diamond-rich Kimberley region of South Africa (Fig. 1). Kimberlite emplacement has been relatively rare during the past 50 million years, and no modern kimberlite eruption has ever

been observed. But such events were relatively common during the Mesozoic era of 250 million to 65 million years ago, when the Pangaea supercontinent was breaking apart. On page 352 of this issue, Torsvik *et al.*² report an intriguing pattern of ancient kimberlite fields in time and space, a pattern that they show has persisted for at least 200 million years, and is rooted to the bottom of Earth's mantle, nearly 3,000 km beneath the planet's surface.

Torsvik *et al.* combine state-of-the-art tectonic-plate reconstructions, using palaeomagnetic data³, with a stacked suite of tomographic images of the planet's deep interior⁴. They show that kimberlite eruptions have been consistently located above the edges of an enormous body of chemically distinct rock in the lowest 500 km of the mantle. This large, low-shear-wave-velocity province (LLSVP) located under Africa is one of two such regions on opposite sides of the world that dominate lowermost-mantle global seismic structure. Most of the world's large igneous provinces (LIPs) of the past 200 million years are localized along their edges^{5,6}. Torsvik *et al.* refer to those edges as 'plume-generation zones' in reference to the narrow mantle upwellings⁷ that give rise to many LIPs.

The LLSVPs are not completely understood. Nonetheless, Torsvik *et al.*² consider the hypothesis that they can be used as fixed markers of absolute palaeolongitude in the mantle. Their success in reconstructing LIPs and kimberlites to the LLSVP edges for the past 200 million to 250 million years shows that the hypothesis is viable. Extending the



Figure 1 | The Big Hole at Kimberley, South Africa. Here, one hundred years of mining for diamonds, largely by hand pick and shovel, have hollowed out the original funnel-like geometry of the kimberlite eruptive pipe that extends deep below the present water table.

R. DELA HARPE/PHOTOLIBRARY.COM

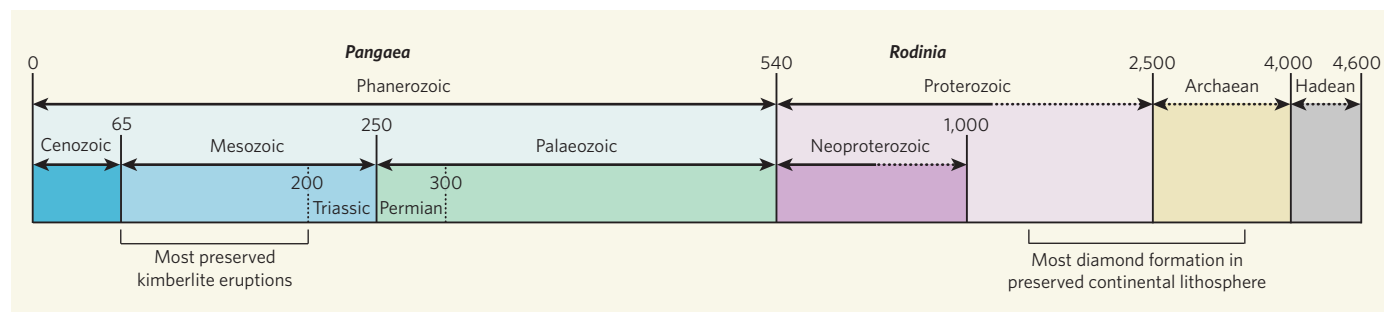


Figure 2 | Kimberlites, diamonds and supercontinents in time. Numbers are ages in millions of years. Most of the world's natural diamonds were formed deep within the continental lithosphere during Archaean and early Proterozoic time, and were brought to the surface much later by kimberlite volcanic eruptions. Torsvik *et al.*² show that the most common incidence of preserved kimberlites, which erupted during an interval starting about 200 million years ago, can be linked to plume-generation zones arising from the ring-shaped margin of a large,

low-shear-wave-velocity province (LLSVP) lying deep in the mantle. Torsvik and colleagues' proposal is that the African and Pacific LLSVPs have been stationary in the mantle for at least the past 540 million years, if not longer. An alternative explanation of this pattern, discussed here, is that supercontinents such as Pangaea and Rodinia caused a reorganization of mantle flow, and that LLSVPs are created and disintegrate in each supercontinental cycle, with no necessary palaeogeographic connection from one to the next.

hypothesis to greater ages, Torsvik *et al.* use the assumed-fixed LLSVPs to likewise reconstruct Palaeozoic continents into an absolute palaeogeographic framework, constrained in both palaeolatitude and palaeolongitude throughout the past 540 million years of Phanerozoic time (see Fig. 2 for a timeline).

Another possible interpretation of the LLSVPs, however, is that they are a legacy of the most recent supercontinent, Pangaea, and were created only about 200 million years ago when most of the preserved kimberlites and LIPs began erupting. Such a scenario is consistent with models of mantle convection that include supercontinents^{8,9} as large caps at Earth's surface that organize subduction zones to their peripheries and create stable, long-wavelength upwellings on the same scale as upwellings of the Mesozoic–Cenozoic Earth. Torsvik and colleagues' palaeogeographic correlations of LIPs, kimberlites and LLSVPs for ages greater than 200 million years are not particularly strong; the Permian–Triassic Siberian traps (one of the largest preserved LIPs on Earth) reconstruct to a position more than 3,000 km from the nearest edge of an LLSVP, and the early–middle Palaeozoic reconstructions to fixed-longitude LLSVPs are tenuous, with only a handful of data constraining the models.

The almost complete lack of known LIPs between 500 million and 300 million years old can be neatly explained by the fact that Pangaea hadn't yet formed. However, Torsvik *et al.* present an alternative explanation — that the LIP record is biased by preservation. They argue that the early–middle Palaeozoic continents were simply not well situated over the LLSVPs, and that any oceanic LIPs formed above the plume-generation zones would now be lost from the geological record by subduction.

However, large oceanic volcanic plateaux can resist subduction and become preserved¹⁰, so the remarkable gap in the current global LIP database between 500 million and 300 million years ago probably reflects a true paucity of such magmatism. Furthermore, despite the numerous late–Palaeozoic traverses of the

LLSVP edges by the supercontinent Gondwanaland's richest diamond-bearing areas (such as western Africa and the Congo region), there is an impressive dearth of kimberlites and LIPs in those regions through that time interval. The same observation holds for the middle–late Cenozoic era, which is almost completely lacking in kimberlites, even in Africa, lying astride its namesake LLSVP.

By contrast, the LIP record from about 800 million to 550 million years ago is globally abundant (although kimberlites of that antiquity have only rarely been found). This can readily be interpreted as the legacy of LLSVPs formed by organization of mantle convection under the earlier supercontinent Rodinia, and associated with its break-up in Neoproterozoic time. Those more ancient LLSVPs might have existed, but there is no guarantee that they occupied the same locations — in terms of an 'absolute' mantle reference frame — as their post-Pangaeian counterparts.

Whether the LLSVP-related plume-generation zones, like the proverbial diamonds they exhume, are 'forever' fixed in the mantle has implications for understanding planetary convection through the aggregation and break-up of supercontinents. If today's LLSVPs formed only 200 million years ago, resulting from Pangaea's organization of global mantle flow, then they can be interpreted as the loci of large-scale upwellings complementary to long-lived subduction systems¹¹. However, if the African LLSVP preceded formation of Pangaea, as interpreted by Torsvik *et al.*², then it seems, oddly, to have attracted the supercontinent's lateral motion so that, just before it broke up, the landmass lay precisely atop the mantle anomaly.

Torsvik *et al.* instead attribute that apparent attraction to coincidence, and speculate that the LLSVPs remain isolated from mantle convection on billion-year timescales. However, viscous entrainment should have eroded their edges through the aeons, creating the conundrum of why they are so large today. In the more conventional 'Pangaea-legacy' model of LLSVPs, there

is no need for such isolation; they are an integral part of a two-upwelling convection system that is expected to be stable until assembly of the next supercontinent around Eurasia, creating new LLSVPs in perhaps entirely different locations in the mantle^{12,13}.

Torsvik and colleagues² have thus made a provocative proposal — that of the existence of a stationary palaeogeographic 'ring' of regions in the mantle containing most of the world's diamond-bearing kimberlites. They propose that mantle convection's engagement with such a ring has endured for as long as 500 million years, and perhaps even longer, through Earth's history. The alternative proposal, espoused here, is that the global convection system breaks its engagement with the ring of plume-generation zones every 500 million years or so, through each supercontinental cycle. Regardless of which model survives further scrutiny, global palaeogeography is advancing to a new stage of kinematics — and consequent implications for dynamics — marrying palaeolatitude with palaeolongitude in deep time. ■

David A. D. Evans is in the Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06511, USA.
e-mail: dai.evans@yale.edu

1. Wilson, L. & Head, J. W. III *Nature* **447**, 53–57 (2007).
2. Torsvik, T. H. *et al.* *Nature* **466**, 352–355 (2010).
3. Torsvik, T. H., Müller, R. D., Van der Voo, R., Steinberger, B. & Gaina, C. *Rev. Geophys.* doi:10.1029/2007RG000227 (2008).
4. Becker, T. & Boschi, L. *Geochem. Geophys. Geosyst.* doi:10.1029/2001GC000168 (2002).
5. Torsvik, T. H., Smethurst, M. A., Burke, K. & Steinberger, B. *Geophys. J. Int.* **167**, 1447–1460 (2006).
6. Burke, K., Steinberger, B., Torsvik, T. H. & Smethurst, M. A. *Earth Planet. Sci. Lett.* **265**, 49–60 (2008).
7. Montelli, R., Nolet, G., Dahlen, F. A. & Masters, G. *Geochem. Geophys. Geosyst.* doi:10.1029/2006GC001248 (2006).
8. Zhong, S., Zhang, N., Li, Z.-X. & Roberts, J. H. *Earth Planet. Sci. Lett.* **261**, 551–564 (2007).
9. Phillips, B. R. & Coltice, N. J. *Geophys. Res.* doi:10.1029/2009JB006600 (2010).
10. Richards, M. A., Jones, D. L., Duncan, R. A. & DePaolo, D. J. *Science* **254**, 263–267 (1991).
11. Richards, M. A. & Engebretson, D. C. *Nature* **355**, 437–440 (1992).
12. Evans, D. A. D. *Tectonophysics* **362**, 303–320 (2003).
13. Li, Z.-X. & Zhong, S. *Phys. Earth Planet. Inter.* **176**, 143–156 (2009).

Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy

Niels Fischer¹, Andrey L. Konevega^{2,3}, Wolfgang Wintermeyer², Marina V. Rodnina² & Holger Stark¹

The translocation step of protein synthesis entails large-scale rearrangements of the ribosome–transfer RNA (tRNA) complex. Here we have followed tRNA movement through the ribosome during translocation by time-resolved single-particle electron cryomicroscopy (cryo-EM). Unbiased computational sorting of cryo-EM images yielded 50 distinct three-dimensional reconstructions, showing the tRNAs in classical, hybrid and various novel intermediate states that provide trajectories and kinetic information about tRNA movement through the ribosome. The structures indicate how tRNA movement is coupled with global and local conformational changes of the ribosome, in particular of the head and body of the small ribosomal subunit, and show that dynamic interactions between tRNAs and ribosomal residues confine the path of the tRNAs through the ribosome. The temperature dependence of ribosome dynamics reveals a surprisingly flat energy landscape of conformational variations at physiological temperature. The ribosome functions as a Brownian machine that couples spontaneous conformational changes driven by thermal energy to directed movement.

During the translocation step of protein synthesis, two tRNAs move by distances of 20 Å or more to adjacent sites in the ribosome, with the coupled movement of messenger RNA (mRNA) by one codon. Translocation is catalysed by elongation factor G (EF-G), which hydrolyses GTP in the process. At certain conditions, translocation can proceed spontaneously in both forward and backward directions, indicating that it is a function inherent to the ribosome itself^{1–4}. Initially, the peptidyl-tRNA and the deacylated tRNA move from the canonical A and P sites of the ribosome, respectively, into so-called hybrid states⁵ (Fig. 1a). In hybrid states, the anticodon stem loops (ASL) of tRNAs reside in the A (aminoacyl-tRNA) and P (peptidyl-tRNA) sites of the small 30S ribosomal subunit, whereas the acceptor ends interact with the P and E (exit) sites of the large 50S subunit. At the same time, the 30S subunit undergoes a ratchet-like movement relative to the 50S subunit^{6–11}. Single-molecule fluorescence experiments have established that, on the pre-translocation ribosome and in the absence of EF-G, tRNAs exist in a dynamic equilibrium between the classical and hybrid configurations^{12–14}. Upon translocation on the 30S subunit, the ASLs of the tRNAs move together with the mRNA and the tRNAs assume the canonical P/P and E/E states, while the ratcheting of the 30S subunit is reversed^{6,11}, followed by the dissociation of the deacylated tRNA from the E site.

Cryo-EM and crystallographic studies visualized various ribosome complexes with tRNAs bound in classical pre- and post-translocation and hybrid positions^{9,10,15–18}. Kinetic and single-molecule data revealed the sequence of events during translocation and provided a quantitative description of some of the tRNA and ribosome dynamics^{6,11,12,19–26}. However, to understand how the tRNAs move from one state to another and how ribosome motions might facilitate the movement, structures of a series of intermediate states, including weakly populated ones, along the reaction coordinate are required. Although cryo-EM in principle allows for time-resolved structure determination, obtaining such sets of structures is challenging, as many different states will coexist in solution in population sizes that vary with time and experimental conditions. Thus, solving structures from such ensembles requires large numbers of images and computational

sorting to resolve the heterogeneity inherent in the mixed population of complexes. Here, we have used time-resolved cryo-EM to obtain a set of ribosome structures that provide trajectories of tRNA movement and reveal sequential rearrangements in the network of ribosome–tRNA interactions which are coupled with global conformational changes of the 30S subunit during translocation.

Structure determination and tRNA trajectories

To monitor tRNA movement and conformational dynamics of the ribosome in parallel, we took advantage of the fact that translocation can happen spontaneously, both in forward and backward directions (Fig. 1a). Unlike the forward reaction, which is rather slow and inefficient¹, the reverse reaction (‘retro-translocation’) takes place at a tractable rate (minutes) and can proceed almost to completion (Fig. 1b). Here we have used a post-translocation complex containing fMetVal-tRNA^{Val} (a peptidyl-tRNA carrying the dipeptide fMetVal) in the P site and induced retro-translocation by adding deacylated tRNA^{fMet} for binding to the E site. Samples for cryo-EM were taken at different time points along the reaction coordinate; EM grids were prepared at 18 °C (Methods and Supplementary Methods).

In total, about 2,000,000 particle images were recorded and analysed (Supplementary Table 1). We applied an unbiased hierarchical image classification strategy to computationally sort the heterogeneous ribosome particles into homogenous sub-states of the ribosome that differed in tRNA position and/or in ribosome conformation (Supplementary Methods, Supplementary Figs 1–6 and Supplementary Movie 1). The 50 structures of these ribosomal sub-states were then refined by projection matching, allowing us to visualize the trajectories of the two tRNAs as they move through the ribosome (Supplementary Methods and Supplementary Movies 2–4). The eight snapshots of the movie—hereafter described and shown in the pre-to-post translocation direction—represent the eight most distinct states of tRNA movement into which the 50 ribosome sub-states can be classified (Fig. 1c and Supplementary 3D PDFs). These states consist of ensembles of up to 11 sub-states that share similar tRNA positions and ribosome–tRNA contacts, but vary in ribosome conformation (Supplementary Table 2).

¹3D Electron Cryomicroscopy Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. ²Department of Physical Biochemistry, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. ³Petersburg Nuclear Physics Institute, 188300 Gatchina, Russia.

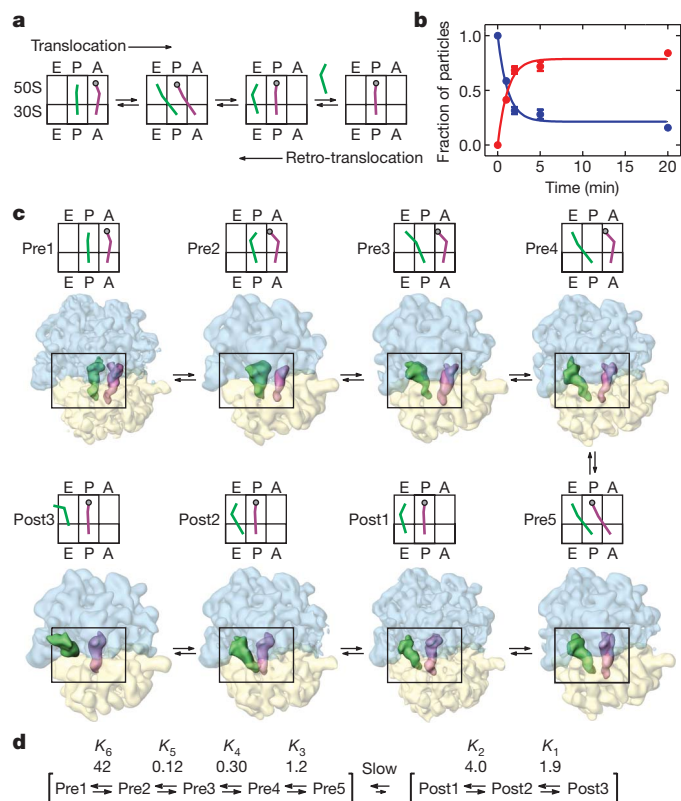


Figure 1 | Trajectories of tRNA movement during (retro-)translocation.

a, Schematic of tRNA movement through the ribosome. Ribosome complexes were prepared by adding a stoichiometric amount of deacylated tRNA^{fMet} (green) that was cognate to the E-site codon, AUG, to ribosome complexes with fMetVal-tRNA^{Val} (purple) in the P site. Samples for cryo-EM were taken before adding tRNA^{fMet}, or after incubation for 1 min, 2 min, 5 min and 20 min at 37 °C. EM grids were prepared at 18 °C. The tRNA binding sites are indicated by E (exit site), P (peptidyl-tRNA site) and A (aminoacyl-tRNA site). **b**, Time course of retro-translocation determined from cryo-EM images. Plotted is the sum of all pre states (red symbols) or of all post states (blue symbols) versus time. Error bars represent standard deviations of values obtained from up to four independent experiments. **c**, Cryo-EM reconstructions of ribosome-tRNA complexes classified by tRNA positions. A representative sub-state from the ensemble of sub-states for each pre- and post-translocational state (pre1 to pre5 and post1 to post3) is shown. In the boxes, the positions of the tRNAs are depicted schematically. **d**, Equilibrium constants (K_1 to K_6) for the transitions between various post and pre states. The transition between the ensembles of post and pre states (indicated by square brackets) is slow.

Retro-translocation time course monitored by cryo-EM

The change in distribution over time for each sub-state of the complex allowed us to estimate the overall rate of the reaction and the internal equilibria between the eight major states of tRNA movement. By summing over all pre and all post sub-state ribosome particles, respectively, we can determine the overall rate of retro-translocation, $0.8 \pm 0.1 \text{ min}^{-1}$ (Fig. 1b), comparable to the rate determined biochemically¹. The fraction of each pre state increases and the fraction of each post state decreases exponentially with time (Supplementary Fig. 7a). However, the ratio of particles in any of the five pre states does not change over time, indicating that they are in rapid equilibrium; the same is true for the three post states (Supplementary Fig. 7b). Thus, the rate-limiting step of retro-translocation is the transition between the post1 and pre5 state, which entails the movement of the ASLs of the tRNAs on the 30S subunit (Fig. 1d). Most likely, this step in the reverse direction is also rate-limiting during translocation in the forward direction, as the reported rates of the overall reaction¹ are several orders of magnitude lower than the rates of spontaneous hybrid-state formation^{12,14} and it is this step on which EF-G acts to accelerate translocation²⁰.

Dynamics of ribosome-tRNA contacts during tRNA movement

The pattern of interactions between the tRNAs and the ribosome changes along the tRNA trajectory, defining a constrained passage of the tRNAs through the ribosomal intersubunit space. By docking atomic models of ribosomal elements^{27,28} into the cryo-EM densities, the changes in the dynamic network of interactions were followed (Figs 2 and 3a and Supplementary Movie 5).

In the first steps of hybrid-state formation, the acceptor end of the deacylated tRNA detaches from the P site of the peptidyl transferase centre (pre2) and moves into the 50S E site (pre3), while the elbow region gradually moves with protein L5 towards the L1 stalk, to which it finally binds. After release from helix 69 (H69) of 23S rRNA and L5 in the pre3 state, the deacylated tRNA assumes the P/E hybrid state (pre4 and pre5). The peptidyl-tRNA in the A site is anchored to H89 during the first steps of hybrid-state formation (pre1 to pre4). Subsequently, the elbow region of the peptidyl-tRNA moves together with H38 towards the P site of the 50S subunit and establishes, in the pre5 state, an interaction with L5, followed by the release from H38. L5 seems to act as a swinging arm that escorts the tRNAs over large distances ($\sim 40 \text{ \AA}$ in total) from the pre to the post states until the

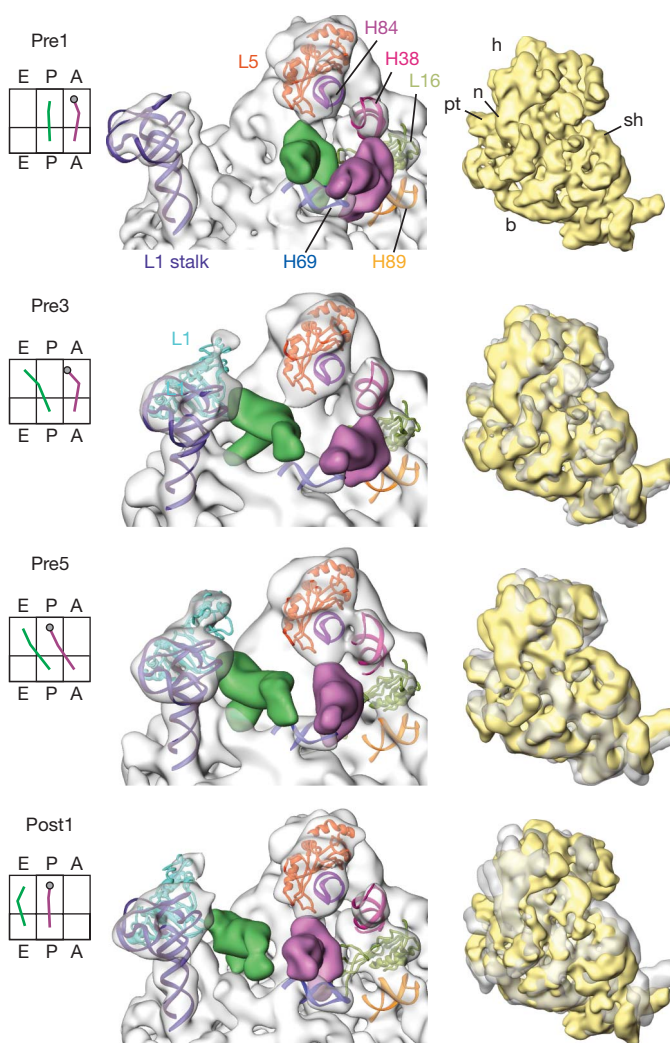


Figure 2 | Dynamics of ribosome-tRNA interactions and of the 30S subunit. tRNA positions in selected states of (retro-)translocation (left-hand boxes) and contacts with 50S subunit regions for representative sub-states (middle panels) are depicted along with the actual state of the 30S subunit (right-hand panels) in yellow, overlaid with the preceding state in grey. Note the anticlockwise 30S body rotation from pre1 to pre5 and the switch back to the non-rotated conformation upon transition to post1. 30S landmarks are: b, body; h, head; n, neck; pt, platform; sh, shoulder. Protein L1 in the L1 stalk is depicted in light blue (no clear density for L1 in pre1).

deacylated tRNA exits the ribosome. Movements of the 30S subunit accompanying hybrid-state formation (pre1 to pre5) result in a coupled shift of the ASLs of the two tRNAs and the apical part of helix 44 (h44) of 16S rRNA and H69 of 23S rRNA (the latter two forming intersubunit bridge 2a). Simultaneously, the gate between the P and E site of the 30S subunit widens²⁷. During the transition from pre5 to post1, the tRNA–mRNA complex moves from the A and P sites into the P and E sites on the 30S subunit, respectively, and the conformational changes of the 30S subunit and H69 that characterized the pre1 to pre5 transition are reversed (Fig. 2). The transition involves an additional shift of the ASLs with respect to the 50S subunit and multiple changes of the interactions between the tRNA–mRNA complex and the 30S subunit as well as of the ribosomal intersubunit contacts²⁹.

In all post states, the peptidyl-tRNA is bound to the P site in a stable position, whereas the deacylated tRNA assumes variable positions in the E site (Fig. 3 and Supplementary Fig. 8). Starting from the canonical E-site position¹⁸, the ASL of the tRNA first moves upwards towards the 30S head into a position above the P-site ASL that has not been observed before and that we term E' (post2 state). Upon further movement, the deacylated tRNA dissociates from the 30S subunit to assume a position where it contacts mainly the L1 stalk of the 50S subunit (post3). The latter 'L1 site' represents the outermost exit (during translocation) or entry (during retro-translocation) state of deacylated tRNA revealed by our analysis. The different positions the tRNA assumes while occupying the E-site region are probably related to the rearrangements of the E-site-bound tRNA observed previously^{30,31}.

The conformational dynamics of the L1 stalk can be described by narrow distributions around three distinct positions: open (prevailing in pre1, pre2 and post3), half-closed (post1, post2 and post3) and closed (pre3, pre4 and pre5). In agreement with previous single-molecule^{23–26} and cryo-EM data^{8–10}, the movements of the L1 stalk are coupled with the movements of the deacylated tRNA.

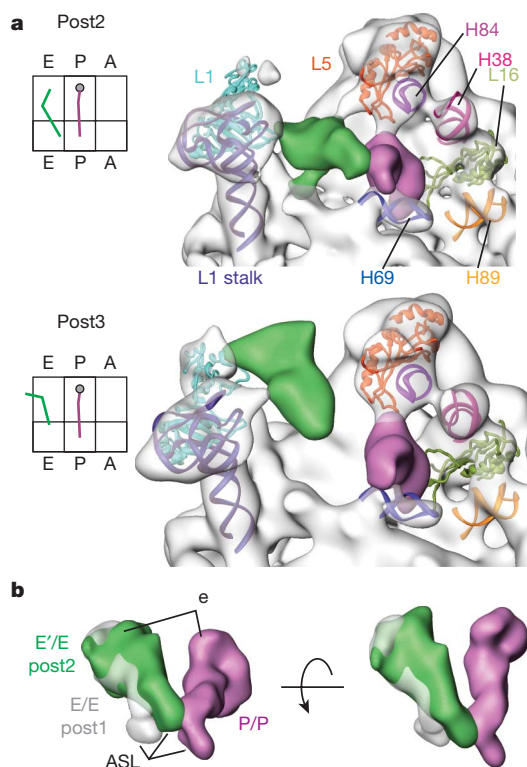


Figure 3 | Positions of the E-site tRNA. **a**, E-site tRNA in post-translocation states on the 50S subunit. One sub-state of state post2 and one sub-state of state post3 each are shown and depicted schematically in the boxes. Note the 'handover' of the deacylated tRNA from protein L5 to the L1 stalk. **b**, Arrangements of the E-site tRNA with the anticodon stem loop (ASL) in E and E' sites on the 30S subunit in different post-translocation states. **e**, tRNA elbow.

In summary, tRNA movement during hybrid-state formation and dissociation of the deacylated tRNA from the E site proceeds via several intermediate states that entail the sequential step-by-step rupture and formation of ribosome–tRNA contacts. Furthermore, several ribosomal elements, such as H38, H69, L5 and the L1 stalk move over long distances (between 8 Å and 40 Å) together with the tRNAs, thereby confining the space accessible for tRNA movement and reducing the need to form or break too many contacts at a time.

Coupling between tRNA movement and ribosome dynamics

For each of the eight states of tRNA movement, we observe a characteristic spectrum of nearly continuous 30S body rotations and 30S head movements, revealing an extended dynamic landscape of ribosome conformations that changes each time the tRNA positions change (Fig. 4). To analyse the coupling between tRNA and global ribosome dynamics in detail, we plotted the observed frequency of particles with certain 30S body rotation angles and 30S head positions for each of the eight intermediate states of tRNA movement (Fig. 1c and Fig. 4a). The propensity for higher degrees of 30S body rotation and head movement strongly increases when the tRNAs enter hybrid states (Fig. 4a, panels 1 to 5). The movements of the 30S subunit body and head are reversed upon transition from pre to post states (Fig. 4a, panels 6 to 8), as previously suggested by results from single-molecule experiments^{6,11}. Apparently, the peptidyl-tRNA in the P site stabilizes the ground state of the ribosome and restricts the rearrangements of the 30S body and head. However, it is also possible to observe a particular tRNA state over a wide range of different 30S conformations and, vice versa, to observe ribosomes in similar overall conformations that differ significantly in their tRNA positions. Furthermore, we also detect large 30S body rotation angles in sub-states that do not show any 30S head movements. Thus, although 30S body and head movements as well as tRNA movement and global changes in ribosome conformation are correlated, coupling is loose in either case.

The number of ribosome particles (n) found in each sub-state at equilibrium relative to the number for the most populated 30S conformation (n_0) can be converted into standard free energy differences [$\Delta G^\circ/k_B T = -\ln(n/n_0)$]; a plot representing the combined free energy landscape of ribosome conformations during hybrid-state formation is shown in Fig. 4d. Each tRNA state is characterized by an ensemble of global conformational states of the 30S subunit that fluctuate around a state with minimum free energy. The fluctuations seem to promote changes in tRNA positions on the 50S subunit, in particular during the initial stages of hybrid state formation (pre1 to pre4). Changes in ribosome–tRNA interactions, in turn, shift the local energy minimum of ribosome conformation. In this way, the tRNAs transiently rectify the structural fluctuations of the ribosome to a new space, thereby increasing the probability of further tRNA movement. The conformational landscapes of pre4 and pre5 states largely overlap, indicating that transitions between the two states do not require global conformational changes of the ribosome. The rapid equilibria between the various tRNA states indicate low kinetic barriers that allow the ribosome–tRNA complexes to sample rapidly all sub-states of either pre or post states and to follow multiple alternative pathways of 30S and tRNA movement. By coupling thermally activated fluctuations of the 30S subunit and of the tRNAs, the ribosome promotes tRNA movement over long distances and large-scale global conformational changes. Directional net movement during retro-translocation results from the thermodynamic gradient between post and pre states. Thus, the ribosome acts as a Brownian machine^{32,33} that harnesses thermal fluctuations into directed tRNA motion.

Temperature-dependence of ribosome dynamics

The ribosome as a Brownian machine should be strongly affected by temperature. To analyse the effect of temperature on the conformational dynamics of the ribosome, cryo-EM samples of a post-translocation complex were prepared at 4 °C, 18 °C and 37 °C (Fig. 5). Three dimensional reconstructions were initially determined without computational

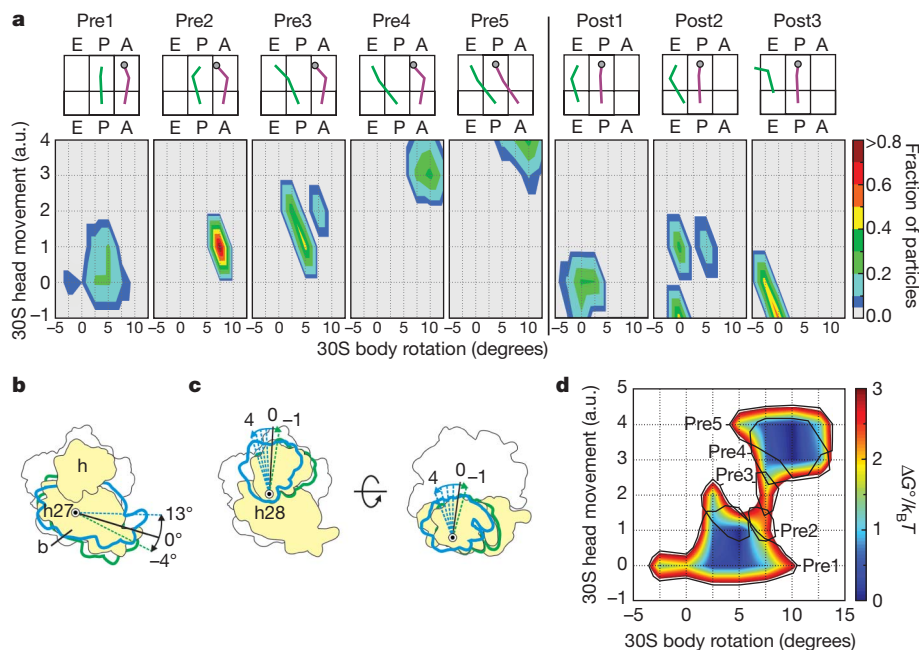


Figure 4 | Coupling between tRNA movement and 30S dynamics.

a, Distribution of 30S sub-states for different tRNA positions (boxes) in pre and post states. The population of sub-states according to 30S body rotation and 30S head movement is plotted for pre and post states. The heat map indicates the fraction of particles relative to the total number of particles in the respective state. **b**, Schematic of 30S body rotation. The 30S body (b) rotates around a pivot point at helix 27 (h27) of 16S rRNA independent of the 30S head (h). In the ground state of the ribosome the rotation angle is 0°.

sorting (Fig. 5a and Supplementary Movie 6). At 4 °C, the ribosome structure shows well-resolved details at ~10 Å resolution. At higher temperature, the overall resolution decreases strongly (Supplementary

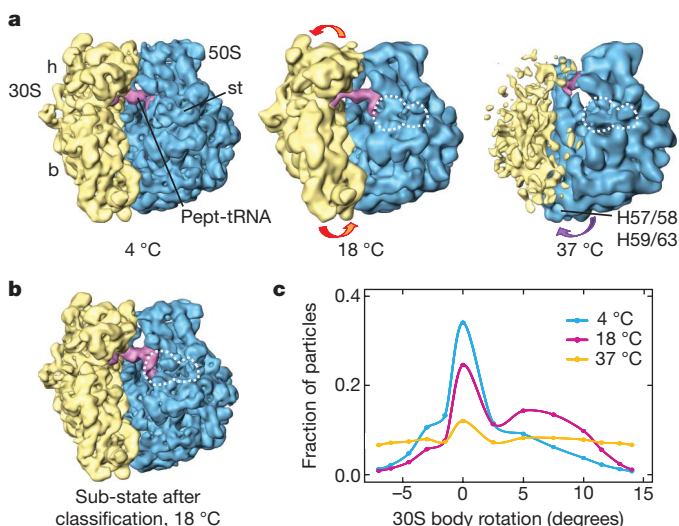


Figure 5 | Increase of ribosome dynamics with temperature. **a**, Three-dimensional reconstructions of post-translocation complexes from 25,000 unsorted cryo-EM images at various temperatures. Red arrows indicate head (h) movement (top) and body (b) rotation (bottom) of the 30S subunit and the blue arrow a displacement of 23S rRNA helices (H), as indicated. st designates the L12 stalk base, which averages out at the two higher temperatures (area indicated by dotted line). Note the scattered density of the 30S subunit and the partial disappearance of the P-site peptidyl-tRNA (purple) at 37 °C. **b**, Three-dimensional structure of the post-translocation complex after classification. The reconstruction was obtained from a sub-population of 25,000 particles at 9 Å resolution. **c**, Temperature dependence of 30S body rotation in the post-translocation ribosome. The fraction of particles is plotted against the rotation angle of the 30S body.

c, Schematic of 30S head movement. 30S head movement comprises a tilt (left) and a swivelling motion (right) around the neck region (h28), depicted in arbitrary units (a.u.) from -1 to 4 corresponding to an overall amplitude of 30 Å. Zero corresponds to the ground state of the ribosome. **d**, Free energy landscape of global ribosome conformation during hybrid-state formation. Black contour lines indicate the conformational space occupied by the different tRNA states (pre1 to pre5). The heat map indicates the free energy for the respective 30S sub-state.

Table 3 and Supplementary Figs 9 and 10). At 18 °C, the 30S subunit rotation angle is increased, whereas at 37 °C only scattered density for the 30S subunit and tRNA can be observed, indicating that the ribosome samples many different sub-states. A quantitative analysis of the 30S body rotation angle shows that the distribution of rotation angles broadens with increasing temperature and becomes entirely flat at 37 °C, indicating that there is almost no preference for any 30S body rotation angle (Fig. 5c and Supplementary Fig. 11). Thus, at elevated temperature the thermal energy is sufficient for the ribosome to sample any possible 30S body rotation with comparable probability. Most likely, such thermally induced transient states are important not only for ribosome function but also for a number of other macromolecular reactions. Whereas current standard protocols working at low temperatures (~4 °C) decrease thermal fluctuations and facilitate high-resolution structure determination of the most highly populated states, cryo-EM sample preparation at elevated temperatures increases the population sizes of the weakly populated intermediate states, thereby making these states accessible to three-dimensional structure determination.

The present results show that at physiologically relevant temperatures the ribosome samples a broad range of sub-states. In particular, during translocation the tRNAs can adopt a number of intermediate states in addition to the classical and hybrid states defined previously. For each tRNA state, the ribosome can assume conformational sub-states that differ in the position of the 30S subunit body and head as well as the conformation of various elements of the 50S subunit, such as the L1 stalk, protein L5, H38 and H69. Loose coupling between thermally activated motions of the tRNAs and the ribosome appears to promote long-distance tRNA movement which can proceed via multiple pathways in the forward and backward direction. Future work will show whether EF-G influences the conformational landscape of the ribosome-tRNA complex to accelerate translocation by stabilizing pre-existing ribosome conformations^{6,8,34}, inducing novel transient states that facilitate detachment of the tRNA-mRNA complex from the 30S subunit^{20,35,36}, preventing backward movement of the tRNAs¹, or by combinations of these effects.

Time-resolved cryo-EM offers the unique opportunity to study the dynamics of macromolecular complexes at physiological temperature, which is very likely to be instrumental to obtain a detailed understanding of how macromolecular machines work in a cell.

METHODS SUMMARY

Biochemical materials were prepared as described¹. Cryo-EM grids were prepared³⁷ under controlled environmental conditions³⁸ at 18 °C (4 °C, 18 °C and 37 °C for the zero and 20 min time point). For each condition, images were recorded from two to five independent experiments under cryo conditions on Philips CM200 FEG/FEI Titan Krios electron microscopes at $\times 161,000/\times 75,000$ magnification on 4,096 $\times$ 4,096 pixel CCD cameras (TVIPS GmbH/FEI Eagle) using twofold pixel binning³⁹, resulting in final pixel sizes of $1.87 \pm 0.02 \text{ \AA}/2.00 \pm 0.02 \text{ \AA}$. Particles were selected semi-automatically using Boxer⁴⁰ and corrected locally for the contrast transfer function⁴¹, resulting in a total of 2,004,547 particles (Supplementary Table 1).

A hierarchical strategy was applied to sort images into sub-populations according to: (1) 30S subunit body rotation relative to the 50S subunit; (2) 30S subunit head movement; and (3) tRNA positions (Supplementary Methods, Supplementary Figs 1, 3 and 4 and Supplementary Table 4). To avoid model bias during alignment, only the 50S subunit (Supplementary Fig. 2) was used to assign projection angles to all images within a group of sorted images, which is similar to the concept of using omit maps in X-ray crystallography⁴². We obtained 50 distinct three-dimensional maps of the ribosome which were refined at resolutions in the range of $\sim 9\text{--}20 \text{ \AA}$ (Supplementary Table 3 and Supplementary Fig. 9) using IMAGIC-5⁴³ and exhaustive alignment⁴⁴.

Received 8 December 2009; accepted 25 May 2010.

- Konevega, A. L. *et al.* Spontaneous reverse movement of mRNA-bound tRNA through the ribosome. *Nature Struct. Mol. Biol.* **14**, 318–324 (2007).
- Shoji, S., Walker, S. E. & Fredrick, K. Reverse translocation of tRNA in the ribosome. *Mol. Cell* **24**, 931–942 (2006).
- Gavrilova, L. P. *et al.* Factor-free (“non-enzymic”) and factor-dependent systems of translation of polyuridylic acid by *Escherichia coli* ribosomes. *J. Mol. Biol.* **101**, 537–552 (1976).
- Fredrick, K. & Noller, H. F. Catalysis of ribosomal translocation by sparsomycin. *Science* **300**, 1159–1162 (2003).
- Moazed, D. & Noller, H. F. Intermediate states in the movement of transfer RNA in the ribosome. *Nature* **342**, 142–148 (1989).
- Cornish, P. V., Ermolenko, D. N., Noller, H. F. & Ha, T. Spontaneous intersubunit rotation in single ribosomes. *Mol. Cell* **30**, 578–588 (2008).
- Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
- Valle, M. *et al.* Locking and unlocking of ribosomal motions. *Cell* **114**, 123–134 (2003).
- Agirrezabala, X. *et al.* Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol. Cell* **32**, 190–197 (2008).
- Julian, P. *et al.* Structure of ratcheted ribosomes with tRNAs in hybrid states. *Proc. Natl Acad. Sci. USA* **105**, 16924–16927 (2008).
- Marshall, R. A., Dorywalska, M. & Puglisi, J. D. Irreversible chemical steps control intersubunit dynamics during translation. *Proc. Natl Acad. Sci. USA* **105**, 15364–15369 (2008).
- Blanchard, S. C. *et al.* tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101**, 12893–12898 (2004).
- Kim, H. D., Puglisi, J. D. & Chu, S. Fluctuations of transfer RNAs between classical and hybrid states. *Biophys. J.* **93**, 3575–3582 (2007).
- Munro, J. B., Altman, R. B., O'Connor, N. & Blanchard, S. C. Identification of two distinct hybrid state intermediates on the ribosome. *Mol. Cell* **25**, 505–517 (2007).
- Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
- Selmer, M. *et al.* Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313**, 1935–1942 (2006).
- Korostelev, A., Trakhanov, S., Laurberg, M. & Noller, H. F. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**, 1065–1077 (2006).
- Jenner, L., Rees, B., Yusupov, M. & Yusupova, G. Messenger RNA conformations in the ribosomal E site revealed by X-ray crystallography. *EMBO Rep.* **8**, 846–850 (2007).
- Rodnina, M. V., Savelsbergh, A., Katunin, V. I. & Wintermeyer, W. Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome. *Nature* **385**, 37–41 (1997).
- Savelsbergh, A. *et al.* An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell* **11**, 1517–1523 (2003).
- Savelsbergh, A. *et al.* Control of phosphate release from elongation factor G by ribosomal protein L7/12. *EMBO J.* **24**, 4316–4323 (2005).
- Pan, D., Kirillov, S. V. & Cooperman, B. S. Kinetically competent intermediates in the translocation step of protein synthesis. *Mol. Cell* **25**, 519–529 (2007).
- Cornish, P. V. *et al.* Following movement of the L1 stalk between three functional states in single ribosomes. *Proc. Natl Acad. Sci. USA* **106**, 2571–2576 (2009).
- Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez, R. L. Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation. *Mol. Cell* **30**, 348–359 (2008).
- Fei, J. *et al.* Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl Acad. Sci. USA* **106**, 15702–15707 (2009).
- Munro, J. B. *et al.* Spontaneous formation of the unlocked state of the ribosome is a multistep process. *Proc. Natl Acad. Sci. USA* **107**, 709–714 (2010).
- Schuwirth, B. S. *et al.* Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310**, 827–834 (2005).
- Villa, E. *et al.* Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *Proc. Natl Acad. Sci. USA* **106**, 1063–1068 (2009).
- Zhang, W., Dunkle, J. A. & Cate, J. H. D. Structures of the ribosome in intermediate states of ratcheting. *Science* **325**, 1014–1017 (2009).
- Robertson, J. M., Paulsen, H. & Wintermeyer, W. Pre-steady-state kinetics of ribosomal translocation. *J. Mol. Biol.* **192**, 351–360 (1986).
- Paulsen, H. & Wintermeyer, W. tRNA topography during translocation: steady-state and kinetic fluorescence energy-transfer studies. *Biochemistry* **25**, 2749–2756 (1986).
- Cordova, N. J., Ermentrout, B. & Oster, G. F. Dynamics of single-motor molecules—the thermal ratchet model. *Proc. Natl Acad. Sci. USA* **89**, 339–343 (1992).
- Astumian, R. D. Thermodynamics and kinetics of a Brownian motor. *Science* **276**, 917–922 (1997).
- Ermolenko, D. N. *et al.* Observation of intersubunit movement of the ribosome in solution using FRET. *J. Mol. Biol.* **370**, 530–540 (2007).
- Lancaster, L. E. *et al.* Colicin E3 cleavage of 16S rRNA impairs decoding and accelerates tRNA translocation on *Escherichia coli* ribosomes. *Mol. Microbiol.* **69**, 390–401 (2008).
- Taylor, D. J. *et al.* Structures of modified eEF2•80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J.* **26**, 2421–2431 (2007).
- Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21**, 129–228 (1988).
- Bellare, J. R., Davis, H. T., Scriven, L. E. & Talmon, Y. Controlled environment vitrification system—an improved sample preparation technique. *J. Electron Microsc. Tech.* **10**, 87–111 (1988).
- Sander, B., Golas, M. M. & Stark, H. Advantages of CCD detectors for *de novo* three-dimensional structure determination in single-particle electron microscopy. *J. Struct. Biol.* **151**, 92–105 (2005).
- Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
- Sander, B., Golas, M. M. & Stark, H. Automatic CTF correction for single particles based upon multivariate statistical analysis of individual power spectra. *J. Struct. Biol.* **142**, 392–401 (2003).
- Bhat, T. N. & Cohen, G. H. Omitmap—an electron-density map suitable for the examination of errors in a macromolecular model. *J. Appl. Crystallogr.* **17**, 244–248 (1984).
- van Heel, M. *et al.* A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
- Sander, B., Golas, M. M. & Stark, H. Corrim-based alignment for improved speed in single-particle image processing. *J. Struct. Biol.* **143**, 219–228 (2003).

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

Acknowledgements We thank C. Blau for technical help in preparing Fig. 4. Research in the laboratory of H.S. was supported by grants from the Federal Ministry of Education and Research (BMBF), Germany, the Sixth Framework Programme of the European Union via the Integrated Project 3DRepertoire, by the state of Lower-Saxony and the Volkswagen Foundation, Hannover, Germany. N.F. was supported by a Boehringer-Ingelheim fellowship. M.V.R. and W.W. were supported by grants from the Deutsche Forschungsgemeinschaft.

Author Contributions N.F. conceived and performed the cryo-EM experiments and data analysis with mentoring by H.S.; A.L.K. prepared complexes for the analysis; N.F., A.L.K., W.W., M.V.R. and H.S. discussed results and wrote the manuscript.

Author Information The three-dimensional density maps have been deposited in the Electron Microscopy Data Bank at the European Biology Laboratory—European Bioinformatics Institute—under accession codes EMD-1716 to EMD-1720 (pre1–5), EMD-1721 to EMD-1723 (post1–3), EMD-1724 (sub-state of post-translocation complex at 18 °C), and EMD-1725 to EMD-1727 (post-translocation complex at 18 °C, 4 °C and 37 °C), respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.S. (hstark1@gwdg.de).

ARTICLES

Viruses in the faecal microbiota of monozygotic twins and their mothers

Alejandro Reyes¹, Matthew Haynes², Nicole Hanson², Florent E. Angly^{2,3}, Andrew C. Heath⁴, Forest Rohwer² & Jeffrey I. Gordon¹

Viral diversity and life cycles are poorly understood in the human gut and other body habitats. Phages and their encoded functions may provide informative signatures of a human microbiota and of microbial community responses to various disturbances, and may indicate whether community health or dysfunction is manifest after apparent recovery from a disease or therapeutic intervention. Here we report sequencing of the viromes (metagenomes) of virus-like particles isolated from faecal samples collected from healthy adult female monozygotic twins and their mothers at three time points over a one-year period. We compared these data sets with data sets of sequenced bacterial 16S ribosomal RNA genes and total-faecal-community DNA. Co-twins and their mothers share a significantly greater degree of similarity in their faecal bacterial communities than do unrelated individuals. In contrast, viromes are unique to individuals regardless of their degree of genetic relatedness. Despite remarkable interpersonal variations in viromes and their encoded functions, intrapersonal diversity is very low, with >95% of virotypes retained over the period surveyed, and with viromes dominated by a few temperate phages that exhibit remarkable genetic stability. These results indicate that a predatory viral–microbial dynamic, manifest in a number of other characterized environmental ecosystems, is notably absent in the very distal intestine.

The diversity of viruses in the gut, and their role in the assembly, maintenance and adaptations of the microbiota and its pool of genes (microbiome), remains unclear. In many environments, the dominant ecological relationship between viruses and their microbial hosts is predatory and follows Lotka–Volterra (or ‘kill the winner’) dynamics. This dynamic is characterized by top-down control of microbial communities (that is, microbial biomass is significantly below the carrying capacity of the habitat), rapid microbial and viral population shifts, and evidence of Red Queen coevolution (in other words, escape strategies in the prey population are countered by predator adaption). One manifestation of this ‘arms race’ is positive selection on loci such as bacterial O antigens and clustered regularly interspaced short palindromic repeats (CRISPR) elements^{1,2}, and viral tail fibres³. In contrast to this predator–prey dynamic, there is another viral life cycle where temperate rather than lytic viruses are longer-term contributors to microbial host phenotypes through provision of adaptive genes. This dynamic can change the metabolic capacities of free-living bacteria and obligate intracellular mutualists⁴, as well as the lifestyles of pathogens⁵. In fact, many of the differences between closely related microbial strains arise from prophage insertions^{6,7}.

Sampling the DNA virome

Recent studies of the human gut virome have focused on pathogen discovery⁸ or have analysed a few individuals without defining the microbial composition of their gastrointestinal tracts^{9–11}. In this Article, we characterize the faecal viromes of four pairs of adult female monozygotic twins and their mothers. All were healthy and none had received antibiotics in the six-month period before sampling. Faecal samples were obtained at the beginning of the study, two months later (50 ± 9 d) and one year later (364 ± 10 d). Virus-like particles (VLPs) were purified from 32 frozen faecal specimens. Because the yield of VLP DNA from 2–5 g of faeces averaged 500 ng, we performed random

amplification of the purified viromes to obtain sufficient material for shotgun 454 GS FLX pyrosequencing. After quality filtering, our final data set contained 280,625,127 nucleotides (Supplementary Table 1). We verified the reproducibility of the protocol by sequencing replicates from five samples (Supplementary Table 1 and Methods). One additional sample was subjected to deeper sequencing (70,157,333 nucleotides). We characterized bacterial taxa represented in faecal samples by pyrosequencing their 16S rRNA genes (Supplementary Table 2). We had previously performed¹² shotgun sequencing of total faecal DNA isolated from the 12 frozen samples obtained at the first time point that in this work were used to prepare purified VLPs (Supplementary Table 3).

Prominence of prophages and phages

We generated a custom non-redundant viral database (NR_Viral_DB) to facilitate analysis of VLP-derived metagenomic data sets (Methods). The NR_Viral_DB contained 4,193 genomic sequences (96.2 megabases): 73.3% were eukaryotic viral genomes and 25.8% were phages and prophages; 76.9% of the phages and prophages were double-stranded-DNA phages, mostly members of the order Caudovirales. The bacterial hosts of these known bacteriophages are principally members of Proteobacteria (54%), Firmicutes (32%) and Actinobacteria (7%).

A relaxed search against the NR_Viral_DB (tBLASTX; *E* value, $<10^{-3}$) showed that $81 \pm 6\%$ (mean $\pm$ s.d.) of reads generated in this study did not match to any known viruses (Supplementary Fig. 1). However, most of the identifiable viruses in the 32 VLP-derived viromes were prophages or phages generally classified as temperate (Fig. 1). The Podoviridae illustrate this point: this family consists of both lytic and temperate members and in the faecal viral community its dominant representatives were temperate (for example coliphage P22-like). The predicted hosts of the identifiable phages and prophages were members

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA. ²Department of Biology, San Diego State University, San Diego, California 92182, USA. ³Advanced Water Management Centre, The University of Queensland, Brisbane, Queensland 4072, Australia. ⁴Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri 63108, USA.

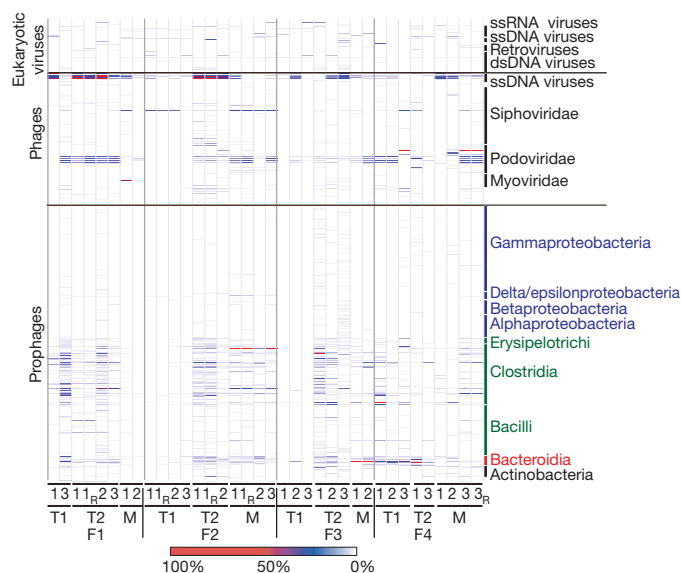


Figure 1 | Classification of viruses present in VLP preparations generated from faecal samples collected from four families of monozygotic twins and their mothers. Prophages are classified on the basis of their bacterial host taxonomy. Prominent bacterial phyla are represented by different colours (Proteobacteria, blue; Firmicutes, green; Bacteroidetes, red; Actinobacteria, black). Class-level taxa within these phyla are indicated. Phages and eukaryotic viruses are sorted according to taxonomy. Nomenclature used for VLP preparations from faecal biospecimens is as follows: F, family; T, co-twin; M, mother of co-twins. Time points (1–3), and technical replicates (R) produced from a given sample are noted. The colour bar at the bottom of the figure provides a reference key for the percentage coverage of a viral genome in the NR_Viral_DB by reads from the given VLP virome data set (data normalized using 14,000 randomly selected reads per data set). ds, double-stranded; ss, single-stranded.

of the Firmicutes and Bacteroidetes (Fig. 1); these phyla, and in particular the families Ruminococcaceae, Lachnospiraceae and Bacteroidaceae, constituted the most abundant bacterial taxa in the sampled faecal microbial communities, as defined by 16S rRNA gene analysis (Supplementary Fig. 2).

Intra- and interpersonal variation in viromes

Two different approaches were used to analyse ‘within VLP sample’ diversity (alpha-diversity estimates). First we used CD-HIT-EST¹³ to cluster reads with $\geq 90\%$ sequence identity over 85% of their length. This allowed us to calculate a cluster-level Shannon index for each VLP sample (Supplementary Table 4) and to define the expected number of virotypes per VLP preparation using procedures described in Methods (median, 346; range, 52–2,773). Second, PHACCS (phage communities from contig spectrum) analysis (Methods) indicated that each sample contained a median of 35 predicted virotypes (range, 10–984; average Shannon index, 3.32 ± 0.71 ; Supplementary Table 5). Analysis of 16S rRNA data sets, where noise due to pyrosequencing artefacts and chimaeras had been removed, indicated that there were ~ 800 species-level bacterial phylotypes per sample in the first-time-point faecal communities¹⁴. Recent results obtained from deep shotgun sequencing of other faecal microbiomes suggest that the number of species may be < 200 (ref. 15). Thus, the ratio of virotype to bacterial phylotype in the faecal microbiota of these healthy adults seems to approach one.

We used two complementary methods to define the percentage of shared viral sequences between faecal samples obtained at different time points from the same individual and from different individuals (beta-diversity estimates). In the first method, we used CD-HIT-EST to cluster pooled reads from all of the VLP viromes. Hellinger distances were subsequently calculated on the basis of a matrix of the number of CD-HIT clusters versus VLP samples, to determine relationships

between samples. The second approach was based on contig assemblies generated from the pooled VLP viromes (MAXIPHI¹⁶; see Methods). Both beta-diversity estimates showed that the same individual harbours very similar faecal viral communities over at least a one-year period: within an individual, $> 95\%$ of the viral genotypes were present and their relative abundances showed minimal variation ($< 8\%$ permutation of the rank abundance). This is different from other comparably characterized ecosystems where almost daily changes in beta diversity occur¹⁷.

Although intrapersonal variation in viromes was minimal, interpersonal variation was very high (Supplementary Fig. 3). To analyse the clustering patterns generated by the distance matrices, 100 random subsamplings of equal numbers of sequences per sample were used to generate the distance matrices and a consensus UPGMA (unweighted pair group method with arithmetic mean) tree. These trees showed that the main branching pattern clusters samples from the same individual, but there was no significant clustering of samples from the same family (Supplementary Figs 4 and 5).

To establish further that temperate phages are dominant members of faecal VLP preparations, we searched for sequence identity between VLP viromes and 121 sequenced human gut microbial genomes. We identified 13 different bacterial genomes containing prophages present at high abundance in at least one VLP sample (Supplementary Table 6). For example, Supplementary Fig. 6 shows a region of the *Ruminococcus torques* ATCC 27756 genome that contains a predicted ~ 60 -kilobase (kb) prophage. Sequence identity plots demonstrated that this prophage was prominently represented in a VLP sample from the mother of family two (F2M.1 and F2M.1_R), where it constituted 18–20% of reads at the first time point; at the two later time points (two months and one year), the phage was still present albeit at lower abundance (0.8 and 1.4%, respectively). This phage was not detectable in the mother’s twin daughters or in other individuals in our study.

Given the low diversity of individual faecal viromes and the apparent representation of a few high-abundance phages, we attempted to assemble partial or complete phage genomes using high stringency conditions (Methods). This effort yielded 5,004 contigs ≥ 500 base pairs long, 88 of which were 10–71.4 kb long (Supplementary Fig. 7). All VLP-derived pyrosequencing reads from all 32 data sets were then aligned against these large contigs. The results revealed that virotypes represented by the large contigs were present mainly in only one individual where their abundance varied over time (Supplementary Table 7). The nucleotide sequence conservation of these contigs (expressed as percentage identity between reads from a given VLP sample and the contig) during the one-year period was remarkably high within an individual ($99.74 \pm 0.26\%$). Only eight of the 88 contigs were present in more than one individual from different families, with an average percentage identity of $88.23 \pm 1.4\%$ between subjects (Supplementary Table 8).

Functional variation among viromes

We next defined functions encoded by the 32 faecal VLP-derived viromes by querying the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the NCBI Clusters of Orthologous Groups (COG) database. The same procedure for functional assignments was used for the 12 shotgun faecal microbiome data sets generated from the first-time-point sample from each individual in the four families. The percentages of faecal VLP-derived reads with significant hits in the COG and KEGG databases (BLAST cut-off, $E < 10^{-5}$) were only $3.2 \pm 2.8\%$ (mean $\pm$ s.d.) and $1.7 \pm 1.9\%$ (mean $\pm$ s.d.), respectively, compared with $36.0 \pm 6.9\%$ (mean $\pm$ s.d.) and $23.3 \pm 3.1\%$ (mean $\pm$ s.d.) for reads obtained from the 12 total-faecal-community DNA samples (Supplementary Fig. 8). Comparison of VLP-derived viromes and the 12 microbiomes revealed significant functional differences (Supplementary Fig. 9) in agreement with previous studies of aquatic ecosystems where viruses were significantly enriched for genes related to DNA, RNA synthesis and replication and the corresponding microbial communities were enriched for nitrogen and carbohydrate

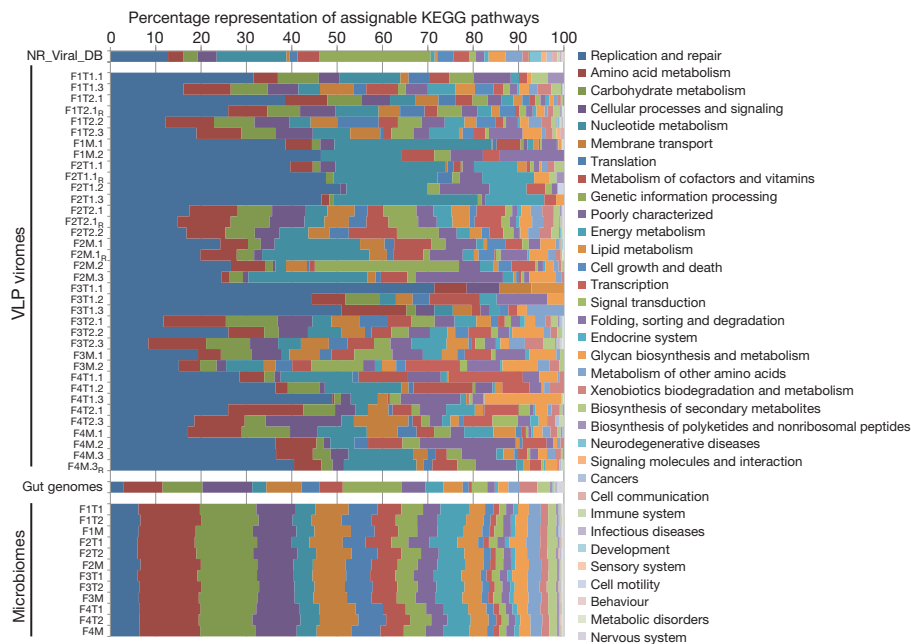


Figure 2 | Sample-by-sample view of the proportional representation of KEGG second-level pathways in sequenced VLP-associated viromes and gut microbiomes. Known or predicted proteins encoded by viruses in the

NR_Viral_DB, faecal VLP-derived viromes, 121-sequenced reference human gut-associated microbial genomes and faecal microbiomes are shown. See Fig. 1 for sample nomenclature.

metabolism and for membrane transport¹⁸. Figure 2 and Supplementary Fig. 10 provide a sample-by-sample view of the proportional representation of KEGG and COG categories in sequenced, purified VLP-derived viromes and in faecal microbiomes. There is only modest interpersonal variation in the distribution of KEGG and COG pathways in the microbiomes ($R^2 = 0.993 \pm 0.005$ (coefficient of determination) for KEGG, $R^2 = 0.984 \pm 0.013$ for COG). Moreover, the distribution of these functions is similar to their distribution in 121 sequenced human gut genomes ($R^2 = 0.82$ for KEGG, $R^2 = 0.95$ for COG; Supplementary Table 9). In contrast, there is marked variation in these functional categories in the sequenced VLP-associated viromes (Fig. 2 and Supplementary Fig. 10), although further and deeper analysis of these differences was limited by the low percentage of viral reads that were classifiable.

The 88 VLP-derived large contigs encode 2,440 predicted proteins, 830 of which have significant similarity to viral or bacterial proteins present in the NR_Viral_DB and/or in the 121 gut genomes (BLASTX; $E < 10^{-5}$). Statistical analysis using METASTATS software¹⁸ identified significant differences in the representation of KEGG and COG functions associated with the large contigs, relative to the NR_Viral_DB, including pathways related to ‘glycan metabolism’, ‘cell wall biosynthesis’ and ‘transcription’ (Supplementary Fig. 11). To identify genes that may confer new and potentially advantageous functions on viruses present in the distal gut microbiota and/or to their microbial hosts, we searched this list of 2,440 proteins, eliminating those with homologues in the NR_Viral_DB, as well as all others whose putative functions suggested a viral origin (for example polymerases, capsid proteins, holins and so on). We were left with 23 proteins belonging to 16 protein families (Table 1). These proteins, seven of which use iron or sulphur in their reactions, are involved in a number of processes associated with the anaerobic gut microbiota. Homologues of these proteins present in 121 gut genomes were subsequently retrieved and aligned¹⁹, and approximate maximum-likelihood trees were generated using FASTTREE²⁰. The results (Supplementary Fig. 12) indicate that some of the VLP virome-associated proteins are evolving in ways that are distinguishable from homologues present in known, sequenced bacterial genomes, as is the case in the few environmental communities that have been subjected to comparable metagenomic studies²¹.

Evidence of a temperate life cycle

Integrases are markers of temperate phages. We identified ten ORFs with homology to integrases in the 88 contigs, and 8,955 reads with significant similarity (BLASTP; $E < 10^{-4}$) to 785 different integrases among the 1,386,331 reads constituting the entire VLP data set.

Table 1 | Proteins encoded by 88 large viral contigs whose functions are involved in processes associated with anaerobic gut microbiota

No. of ORFs	Name
5	<i>N</i> -acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)
3	Thymidylate synthase (EC 2.1.1.148)
2	6-pyruvoyltetrahydropterin synthase (EC 4.2.3.12)
1	Anaerobic nitric oxide reductase transcription regulator (NifA)
1	Fe-S oxidoreductase
1	Anaerobic ribonucleoside-triphosphate reductase activating protein
1	ExsB
1	Phosphoadenosine phosphosulphate reductase family member (EC 1.8.4.8)
1	Ferritin Dps-family protein
1	Glycosyl transferase family 25
1	Glycosyl transferase family 2
1	Methylglyoxal synthase (EC 4.2.3.3)
1	Iron/manganese superoxide dismutases, carboxy-terminal domain (EC 1.15.1.1)
1	Thioredoxin
1	S-adenosylmethionine decarboxylase (EC 4.1.1.50)
1	Cysteine desulphurase (EC 2.8.1.7)

This list of proteins without homologues in the NR_Viral_DB includes (i) two transcriptional regulators (a homologue of ExsB involved in regulation of succinoglycan levels²⁸ and an anaerobic nitric oxide reductase regulator belonging to the sigma 54 family); (ii) an anaerobic ribonucleoside triphosphate reductase activating protein that uses S-adenosyl-L-methionine (SAM), an Fe-S cluster and a reductant for the *de novo* anaerobic synthesis of nucleotides^{29–31}; (iii) other SAM-related proteins (Fe-S oxidoreductase and a SAM-decarboxylase that uses SAM for synthesis of spermidine and spermine, which in turn stimulate RNA polymerases and stabilize the DNA helix, respectively³²); (iv) three oxidative-stress-related proteins (an iron/manganese superoxide dismutase, thioredoxin and a ferritin Dps-family protein); (v) a methylglyoxal synthase homologue involved in pyruvate metabolism; (vi) three thymidylate synthases and two 6-pyruvoyltetrahydropterin synthases involved in folate metabolism; (vii) a member of the phosphoadenosine phosphosulphate reductase family that participates in cysteine biosynthesis and uses thioredoxin as an electron donor; (viii) cysteine desulphurase (NifS), which is important in Fe-S-cluster biosynthesis because it catalyses the removal of sulphur from cysteine to produce alanine; and (ix) a group of proteins involved in peptidoglycan synthesis (a member of CAZy glycosyl transferase family 2 (GT2), a member of GT25 and five *N*-acetylmuramoyl-L-alanine amidases; acquisition of this last group of enzymes is intriguing in light of evidence that some phages can subvert normal bacterial pathways for surface glycan biosynthesis⁵). EC, Enzyme Commission number; ORF, open reading frame.

The number of hits to different integrases per VLP sample was then used to construct a distance matrix which showed that the diversity among identified integrases was significantly lower within VLP viromes purified from the same individual over time than between individuals, and that there were no significant differences in integrase diversity between individuals, regardless of family relationships (Supplementary Fig. 13).

As noted above, in most ecosystems where phage–host interactions have been studied in detail, lytic life cycles and Red Queen dynamics seem to dominate²². Metagenomic studies of salterns and sludge ecosystems indicate that many of the most apparent genetic changes over time are at loci that prevent phage attachment (for example outer-membrane proteins and polysaccharides). Probably the clearest example of Red Queen dynamics between phages and their hosts are the CRISPR elements in sludge¹ and acid mine drainage systems². Similarly, phage genomes often show evidence of changes in their tail fibres over time³. It is therefore striking that we found essentially no evidence of this type of behaviour in faecal phages (see Supplementary Discussion for analysis of CRISPRs). In contrast, we found high abundances of dominant phages, present in the same individual for extended periods of time, with no significant divergence or mutations in their genomes. The presence of integrases in the assembled VLP contigs is also consistent with the notion that they represent prominent temperate phages in the faecal microbiota.

One potential explanation is that phage production occurs through induction of prophages caused by energy limitation in the faeces; at this point, faecal microbial hosts are effectively at a dead end for their associated phages, and the viruses may gain an advantage for transmission by exiting their host cells as VLPs. Experimental evidence for this explanation is provided by gnotobiotic mice co-colonized for two

weeks with *Marvinbryantia formatexigens*, a human gut acetogen that contains three predicted prophages, and *Bacteroides thetaiotaomicron*, a saccharolytic bacterium that harbours two predicted prophages. Normalized RNA-Seq counts, generated from caecal contents and faecal samples collected when the mice were killed ($n = 3$ mice)²³, revealed that one of the three prophages in *M. formatexigens* was completely activated (all ORFs transcribed) in all faecal samples and in a subset of caecal samples (Fig. 3). In the case of the remaining two *M. formatexigens* prophages, only a few genes were expressed, including a pair of adjacent ORFs encoding a HicA-family toxin (BRYFOR7601) and a HicB-family antitoxin (BRYFOR7602) in one of the prophages, and a pair of genes that specify a RelE-family toxin (BRYFOR9696) and a Phd-family antitoxin (BRYFOR9697) in the other prophage; these two gene pairs were constitutively expressed in all faecal samples, in all caecal samples and during *in vitro* growth in defined medium containing a variety of carbon sources (Fig. 3). Coexpression of toxin–antitoxin genes is known to maintain stable integration of phage DNA in bacterial chromosomes^{24,25}. Importantly, the one prophage that was fully activated (prophage 2 in Fig. 3) does not have a detectable toxin–antitoxin gene pair. Only two small (2–3-gene) clusters were expressed in the two *B. thetaiotaomicron* prophages *in vivo*; all of these clusters encode predicted toxin or antitoxin genes (Supplementary Fig. 14). These findings illustrate how a prophage may be liberated from its host cell when that cell is present in a faecal community.

Incorporating virome analyses into human microbiome projects

Human microbiome projects have been initiated throughout the world to define the interrelationships between physiological status and/or disease states and microbial community structure and function. Our

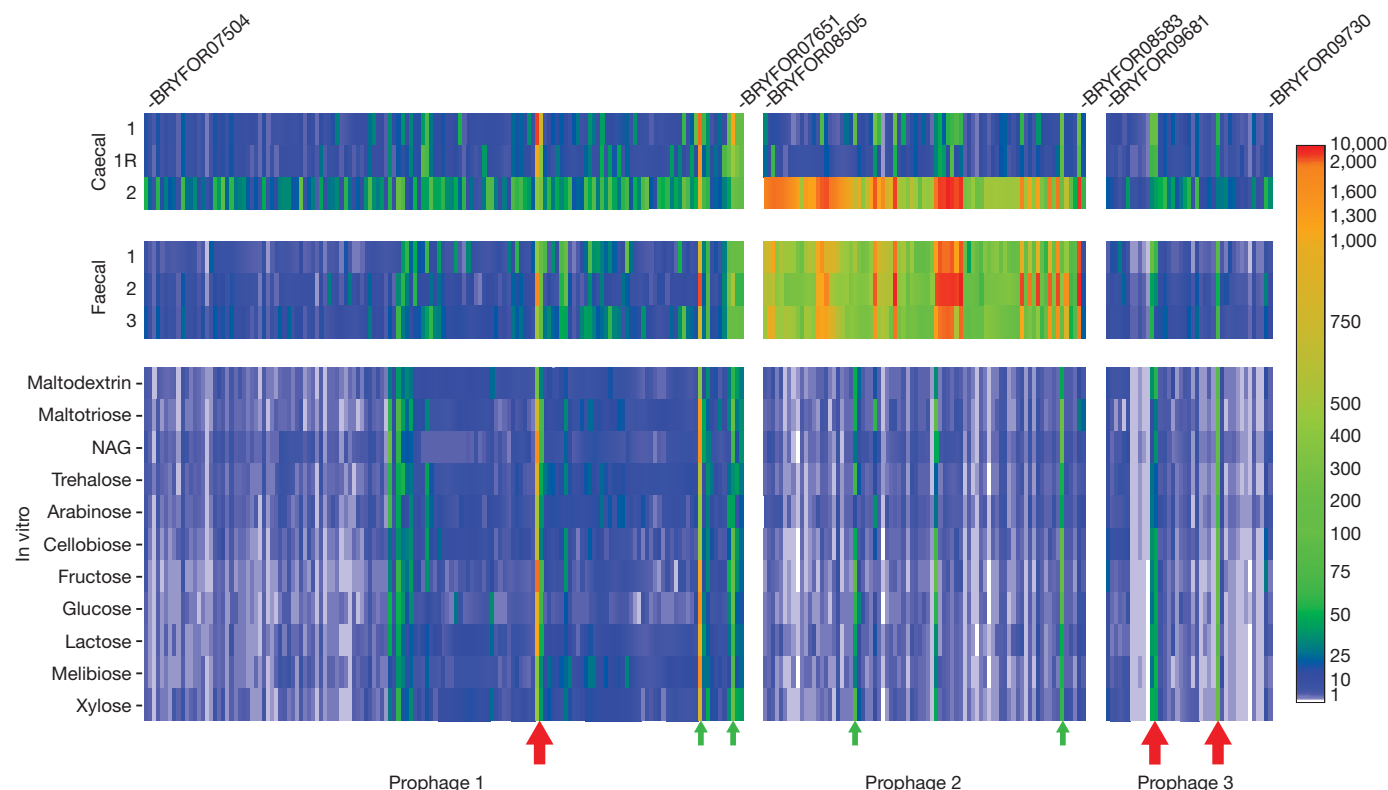


Figure 3 | Gnotobiotic mice reveal *in vivo* activation of the transcriptome of an *M. formatexigens* prophage. Shown are the three predicted prophages present in *M. formatexigens* and the levels of expression of their ORFs in caecal and faecal microbial communities collected from gnotobiotic mice co-colonized with *B. thetaiotaomicron*. Expression levels for genes in each prophage genome are from normalized RNA-Seq read count data (see colour key; normalization is based on sequencing effort and length of each predicted ORF). Active expression is defined as a normalized read count >100.

Technical replicate (R) shows the reproducibility of the method for performing RNA-Seq analysis. RNA-Seq data are also presented for each prophage genome in *M. formatexigens* during mid-log-phase growth in defined medium containing different carbon sources (NAG, *N*-acetyl-D-glucosamine). Red arrows indicate the positions of toxin–antitoxin gene pairs. Green arrows denote genes with hypothetical functions that are expressed under more than 50% of the conditions tested. ORF designations for the first and last genes in the predicted genomes of each prophage are provided.

results indicate that these metagenomic studies should also include VLPs recovered from various body-habitat microbiota (Supplementary Fig. 15 and Supplementary Discussion). Functions embedded in both dominant and subdominant phages may provide informative molecular signatures (biomarkers) of a microbiota and its human host, of microbial community responses to impending or fully manifest disease states, and of the extent to which community health or pathology endures after apparent recovery of the human host from a disease or therapeutic intervention. In addition, gnotobiotic mice harbouring defined collections of human gut symbionts inoculated with VLP-derived viromes should provide informative models for further investigation of various aspects of the interactions between phages and their microbial hosts in different regions of the gut, including an assessment of whether Lotka–Volterra dynamics operate in more proximal regions of the intestine where energy generated from dietary components may be more available.

METHODS SUMMARY

Sample collection. The Missouri Adolescent Female Twin Study²⁶ is composed of female twin pairs born in the state of Missouri between 1975 and 1986, and their mothers. Procedures for obtaining informed consent and sample collection were approved by the Washington University Human Studies Committee.

DNA extraction and 454 pyrosequencing. We processed aliquots of frozen faecal samples (2–5 g) for isolation of VLPs by serial filtration followed by caesium chloride gradient ultracentrifugation²⁷. VLPs were lysed in a solution containing proteinase K and 10% SDS. DNA was extracted with 10% cetyltrimethylammonium bromide and 0.7 M NaCl and amplified using the illustra GenomiPhi V2 kit (GE Healthcare). We used the resulting DNA for multiplex shotgun 454 GS FLX pyrosequencing. For further details about VLP purification, extraction of VLP DNA, assembly of pyrosequencer reads and data analysis, see Methods.

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

Received 23 March; accepted 25 May 2010.

- Kunin, V. *et al.* A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.* **18**, 293–297 (2008).
- Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* **10**, 200–207 (2008).
- Angly, F. *et al.* Genomic analysis of multiple Roseophage SIO1 strains. *Environ. Microbiol.* **11**, 2863–2873 (2009).
- Oliver, K. M., Degnan, P. H., Hunter, M. S. & Moran, N. A. Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **325**, 992–994 (2009).
- Schuch, R. & Fischetti, V. A. The secret life of the anthrax agent *Bacillus anthracis*: bacteriophage-mediated ecological adaptations. *PLoS ONE* **4**, e6532 (2009).
- Casjens, S. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49**, 277–300 (2003).
- Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
- Finkbeiner, S. R. *et al.* Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.* **4**, e1000011 (2008).
- Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).
- Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, e3 (2006).
- Breitbart, M. *et al.* Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).
- Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
- Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
- Turnbaugh, P. J. *et al.* Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl Acad. Sci. USA* **107**, 7503–7508 (2010).

- Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
- Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
- Rodriguez-Brito, B. *et al.* Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**, 739–751 (2010).
- White, J. R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* **5**, e1000352 (2009).
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
- Sharon, I. *et al.* Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**, 258–262 (2009).
- Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nature Rev. Microbiol.* **7**, 828–836 (2009).
- Rey, F. E. *et al.* Dissecting the *in vivo* metabolic potential of two human gut acetogens. *J. Biol. Chem.* **285**, 22082–22090 (2010).
- Magnuson, R. D. Hypothetical functions of toxin-antitoxin systems. *J. Bacteriol.* **189**, 6089–6092 (2007).
- DeShazer, D. Genomic diversity of *Burkholderia pseudomallei* clinical isolates: subtractive hybridization reveals a *Burkholderia mallei*-specific prophage in *B. pseudomallei* 1026b. *J. Bacteriol.* **186**, 3938–3950 (2004).
- Heath, A. C. *et al.* Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data. *Twin Res.* **5**, 107–112 (2002).
- Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nature Protocols* **4**, 470–483 (2009).
- Becker, A., Kuster, H., Niehaus, K. & Puhler, A. Extension of the *Rhizobium meliloti* succinoglycan biosynthesis gene cluster: identification of the *exsA* gene encoding an ABC transporter protein, and the *exsB* gene which probably codes for a regulator of succinoglycan biosynthesis. *Mol. Gen. Genet.* **249**, 487–497 (1995).
- Gon, S., Faulkner, M. J. & Beckwith, J. *In vivo* requirement for glutaredoxins and thioredoxins in the reduction of the ribonucleotide reductases of *Escherichia coli*. *Antioxid. Redox Signal.* **8**, 735–742 (2006).
- Padovani, D., Thomas, F., Trautwein, A. X., Mulliez, E. & Fontecave, M. Activation of class III ribonucleotide reductase from *E. coli*. The electron transfer from the iron-sulfur center to S-adenosylmethionine. *Biochemistry* **40**, 6713–6719 (2001).
- Garriga, X. *et al.* *nrdD* and *nrdG* genes are essential for strict anaerobic growth of *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **229**, 189–192 (1996).
- Tabor, C. W. & Tabor, H. 1,4-Diaminobutane (putrescine), spermidine, and spermine. *Annu. Rev. Biochem.* **45**, 285–306 (1976).

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

Acknowledgements We thank S. Wagoner and J. Manchester for technical assistance, J. Faith for help developing Phage_omics and, together with F. Rey, for microbial RNA-Seq data sets, P. Turnbaugh for assistance with faecal metagenomic studies, and B. Rodriguez-Mueller and D. Willner for valuable discussions. This work was supported in part by grants from the NIH (American Recovery and Reinvestment Act supplemental funding of DK78669), the Crohn's and Colitis Foundation of America and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation. A.R. is the recipient of an International Fulbright Science and Technology Program award.

Author Contributions A.R. and J.I.G. designed the experiments; A.C.H. recruited the patients; A.R., M.H. and N.H. generated the data; A.R., F.E.A., F.R. and J.I.G. interpreted the results; and A.R., F.R. and J.I.G. wrote the paper.

Author Information Virome data sets reported here are accessible in the NCBI Short Read Archive under accession number SRA012183. 16S rRNA data sets are available in GenBank under accession number SRA020605. RNA-Seq data have been deposited in the Gene Expression Omnibus as series GSE21906 (see Methods for further details). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and request for materials should be addressed to J.I.G. (jgordon@wustl.edu).

METHODS

Sample collection. The Missouri Adolescent Female Twin Study²⁶ is composed of female twin pairs born in the state of Missouri between 1975 and 1986, and their mothers. Procedures for obtaining informed consent and sample collection were approved by the Washington University Human Studies Committee.

Purification of VLPs. Viral purification was performed with minor modifications of the procedure described in an earlier publication²⁷. In brief, all samples were frozen at -20°C within 30 min after donation, placed at -80°C within 36 h and maintained at -80°C until they were used. For VLP purification, a 2–5-g aliquot of each faecal sample was resuspended in 25 ml SM buffer (100 mM NaCl, 8 mM MgSO_4 , 50 mM Tris (pH 7.5) and 0.002% gelatin (w/v)). Following centrifugation (2,500g for 10 min at room temperature (23°C)), the resulting supernatant was removed and passed sequentially through 0.45- μm and 0.22- μm Whatman filters to remove residual cells. The filtrate was then adjusted with CsCl to a density of 1.12 g ml^{-1} and deposited on top of a 3-ml step gradient prepared using 1-ml CsCl solutions with respective densities of 1.7 g ml^{-1} , 1.5 g ml^{-1} and 1.35 g ml^{-1} SM buffer. Samples were centrifuged for 2 h at 60,000g (4°C) in a SW41 swinging bucket rotor (Beckman). The 1.5 g ml^{-1} layer was recovered because material in this density range is known to be enriched for bacteriophages²⁷. At each step of the purification procedure, an aliquot of the sample was viewed under an epifluorescence microscope after viral particles had been stained with SYBR-gold; this allowed us to document the presence of VLPs and note whether a decrease in the representation of bacterial and eukaryotic cellular elements had occurred.

Extraction of viral DNA. After the 1.5 g ml^{-1} layer was collected from the step gradient, chloroform was added (0.2 volumes) and the solution was centrifuged for 5 min at 2,500g. The aqueous phase was treated with DNase (Sigma Aldrich; final concentration, 2.5 U ml^{-1}) to remove residual host and bacterial DNA. To extract the virions, 0.1 volumes of 2 M Tris HCl/0.2 M EDTA, 1 volume of formamide and 100 μl of a 0.5 M EDTA solution were added per 10 ml of sample, and the resulting mixture was incubated at room temperature for 30 min. The sample was subsequently washed with 2 volumes of ethanol and pelleted by centrifugation for 20 min at 8,000g at 4°C . The pellet was washed twice with 70% ethanol and resuspended in 567 μl of TE buffer, followed by 30 μl of 10% SDS and 3 μl of a 20 mg ml^{-1} solution of proteinase K (Fisher Scientific; cat no. AC61182-0500). The mixture was incubated for 1 h at 55°C , and 100 μl of 5 M NaCl and 80 μl of a solution of 10% cetyltrimethylammonium bromide/0.7 M NaCl were subsequently introduced. After a 10-min incubation at 65°C , an equal volume of chloroform was added and the mixture was centrifuged (5 min at 8,000g at room temperature). The resulting supernatant was transferred to a new tube and an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added, followed by centrifugation (5 min at 8,000g at room temperature). The supernatant was recovered and an equal volume of chloroform was introduced. Following centrifugation, the supernatant was collected and 0.7 volumes of isopropanol used to precipitate the DNA. After another centrifugation step (15 min at 13,000g at 4°C), the material was washed (500 μl of cold 70% ethanol), air-dried and resuspended in 50 μl TE. An aliquot of the purified DNA was used as a template in polymerase chain reactions that contained universal primers directed at bacterial 16S rRNA and eukaryotic 18S rRNA genes; this assay was used to confirm the absence of detectable, contaminating, non-viral DNA.

Amplification of VLP-associated DNA. Shotgun 454 pyrosequencing requires 3–5 μg of DNA for library preparation. The typical yield from our faecal VLP-DNA isolation procedure was 500 ng per sample. Therefore, whole-genome amplification (WGA) was performed using reagents and protocols in the illustra GenomiPhi V2 kit (GE Healthcare) to generate sufficient material for library construction. Purified VLP DNA (10–50 ng) was mixed with 9 μl 'Sample Buffer' from the kit and heat-denatured at 95°C for 3 min. Nine microlitres of 'Reaction Buffer' and 1 μl of 'Enzyme Mix' were then added and the solution incubated for 90 min at 30°C . Three separate WGA reactions were performed for each viral DNA preparation to minimize potential bias in amplification. The amplified products from each sample were subsequently pooled and purified (QIAGEN DNeasy kit).

To test for bias in the amplification and sequencing of VLP-DNA preparations that had been subjected to WGA, we analysed the VLP sample from individual F3T1.1, where the yield of DNA was sufficient to perform shotgun pyrosequencing with unamplified as well as with amplified subsamples. We pooled 16,567 reads derived from WGA DNA and 18,845 reads from the unamplified aliquot, and clustered them using the procedures followed for alpha-diversity calculations (see 'CD-HIT clustering', below). We found that 98.4% of the sequences from the unamplified DNA were also present in the WGA reads whereas 91.96% of the WGA sequences were represented in the unamplified sample data set. This difference could be due to sequencing of amplified low-abundance viral DNAs that were not sequenced in the unamplified sample. WGA is also known to amplify small, single-stranded DNA viruses preferentially¹⁷.

Multiplex shotgun pyrosequencing of VLP viromes. DNAs, purified from each of 12 VLP preparations, were labelled with a different MID (multiplex identifiers; Roche). Equivalent amounts of the bar-coded samples were then pooled before a run of 454 FLX pyrosequencing. Shotgun reads were filtered by removing (i) all duplicates (defined as sequences whose initial 20 nucleotides are identical and that share an overall identity of $>97\%$ throughout the length of the shortest read; duplicates are a known pyrosequencing artefact³³), (ii) reads with degenerate bases ('N's) and (iii) sequences with significant similarity to human reference genomes (BLASTN (v2.2.20); $E < 10^{-5}$) to ensure the de-identification of samples.

Bacterial 16S rRNA gene amplification and sequencing. An aliquot (500 mg) of each frozen pulverized faecal sample was resuspended in a solution containing 500 μl of extraction buffer (200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA, 210 μl of 20% SDS, 500 μl of phenol/chloroform/isoamyl alcohol (25:24:1) and 500 μl of a slurry of 0.1-mm-diameter zirconia/silica beads (BioSpec Products). Cells were mechanically disrupted using a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol/chloroform/isoamyl alcohol and precipitation with isopropanol. DNA obtained from three separate aliquots of each faecal sample were pooled and used for amplification of bacterial 16S rRNA genes.

Approximately 330-base-pair (bp) amplicons, spanning variable region 2 (V2) of bacterial 16 rRNA genes were generated by using PCR, modified primer 8F (5'-GCCTTGCCAGCCCGCTCAGTCAGAGTTTGATCCTGGTCTAG-3'), which consisted of 454 primer B (underlined) and the universal bacterial primer 8F (italics), and modified primer 338R (5'-GCCTCCCTCGCGCCATCAGNNNNNNNNNNCATGCTGCCTCCCGTAGGAGT-3'), which contained 454 primer A (underlined), a sample-specific, error-correcting 12-mer bar code³⁴ ('N's) and the bacterial primer 338R (italics).

Four replicate polymerase chain reactions were performed for each pooled faecal DNA sample. Each 20- μl reaction contained 100 ng of gel-purified DNA (Qiaquick, QIAGEN), 8 μl 2.5X HotMaster PCR Mix (Eppendorf) and 0.3 μM of each primer. The PCR program consisted of initial denaturation at 95°C for 2 min followed by 30 cycles of denaturation (95°C for 20 s), annealing (52°C for 20 s) and amplification (65°C for 1 min). Replicate PCRs were subsequently pooled and purified using Ampure magnetic purification beads (Agencourt). DNA was quantified using Picogreen (Invitrogen) and an equimolar amount of each sample was used for multiplex 454 FLX amplicon pyrosequencing.

Bacterial 16S rRNA data processing and analysis. 16S rRNA reads were analysed using QIIME (v1.0)³⁵: fasta, quality files and a mapping file indicating the bar-code sequence corresponding to each sample were used as inputs. The QIIME pipeline takes this input information and split reads by samples according to the bar code, and classifies reads into OTUs on the basis of sequence similarity. It also performs taxonomical classification using the RDP-classifier (v2.2)³⁶, builds a *de novo* taxonomic tree of the sequences on the basis of sequence similarity and creates a table of samples versus OTUs that can be used, together with the tree, to calculate alpha and beta diversity.

Custom non-redundant viral database (NR_Viral_DB). All complete viral and bacteriophage genomes available in the European Bioinformatics Institute (EBI) database were downloaded, as were all complete prophage genomes present in the SEED database and all entries for the taxid 10239 (Viruses) in RefSeq between 1 kb and 500 kb. The database was complemented with prophage sequences identified from a survey of 396 sequenced microbial genomes (Supplementary Table 10) using the software tool PHAGEFINDER (v1.9beta)³⁷. To make the database non-redundant, all sequences were compared against each other and only those with $<95\%$ identity throughout their length were retained.

CD-HIT clustering. CD-HIT-EST (v3.1) is a software tool designed for clustering nucleotide sequences by similarity¹³. We used CD-HIT-EST to cluster the pooled reads obtained from all viral samples. Hierarchical clustering was performed on the basis of continuous reduction of the required percentage overlap between reads (from 99% to 85%) while maintaining a sequence identity of $\geq 90\%$. A table of sample versus CD-HIT clusters was subsequently generated, analogous to an OTU table. The table was analysed using QIIME to generate alpha-diversity estimates (as Shannon indices) as well as beta-diversity matrices based on Hellinger distances.

Viral alpha and beta diversity. These estimates were based on a pipeline composed of several software programs: GAAS (v0.14)³⁸, CIRCONSPCT (v0.2.1)¹⁶, PHACCS (v1.0.7)³⁹ and MAXIPHI¹⁶. CIRCONSPCT (<http://sourceforge.net/projects/circonspect/>) was used to form cross-contig spectra of $\times 3$ coverage. To ensure stringent assembly, contigs in CIRCONSPCT were determined by MINIMO (v2.0.5) (available in the AMOS package at <http://sourceforge.net/projects/amos/>) instead of TIGR ASSEMBLER. Contig assembly parameters were 98%-similar sequences overlapping by at least 35 bp. Average viral genome length was then estimated using GAAS (<http://sourceforge.net/projects/gaas/>)

(tBLASTX against complete NCBI RefSeq viral genomes; minimum 30% similarity, minimum 70% relative length).

For each pair of samples to compare, the input for MAXIPHI was their cross-contig spectrum and average genome length. Building on PHACCS (<http://sourceforge.net/projects/phaccs/>), MAXIPHI ran a Monte Carlo simulation to determine how many virotype samples had in common (percentage shared), and how many of the most abundant ones changed their abundance rank (percentage permuted). Using VLP reads from each individual sample assembled against themselves as internal controls, the best average genome length for each beta-diversity computation was found as the length within 20% of the input value that produced the percentage shared and percentage permuted closest to 100% and 0%, respectively.

The entire viral diversity analysis was done at several levels: between time points for each individual, between twins for each family, between twin and mother for each family, and between all families. Input viral metagenomes were pooled as necessary; for example, for the co-twin comparison, sequences from all three time points from each twin were merged.

PHACCS calculates the expected number of virotypes by estimating Shannon indexes from contig spectra. To compare the results using an independent method, Shannon indexes derived from CD-HIT clustering were determined: the expected number of clusters per sample was divided by the expected number of clusters per virotype, as determined by the average genome size given by GAAS and the mapping of clusters per kilobase of viral contigs.

Assembly and analysis of viral genomes. The 454 Newbler Assembler (software release 2.0.01.14) was used for assembly of viral genomes. Default parameters were used except for minimum identity between the sequences (98%) and minimum overlap (100 bp). These stringent conditions decrease the risk of false assembly between reads from different viruses⁴⁰.

We created an online tool for visualizing assembled viral contigs (http://gordonlab.wustl.edu/phage_omics/). Each contig with a length >10 kb generated from the assembly was BLASTed against a non-redundant set of proteins encoded by viruses present in our NR_Viral_DB (which contains entries deposited in public databases as of May 2009), as well as translated ORFs present in 121 sequenced microbial genomes representing cultured representatives of the human gut microbiota. ORF prediction was performed using GLIMMER3 (v3.02)⁴¹. ORFs were subsequently annotated on the basis of BLASTX searches ($E < 10^{-5}$) of the KEGG (v51), COG/String (v8⁴²), PFAM (v23⁴³) and TIGRFAM (v7⁴⁴) databases. All features and annotations for each contig were included in a MySQL database and displayed using LWGV⁴⁵ (lightweight genome viewer, v0.4).

All processed pyrosequencing reads from each VLP sample were BLASTed (BLASTN; $E < 10^{-5}$) against each of the contigs. Significant hits were recorded and the positions used to plot cumulative coverage. The length of the alignment was used to calculate a normalized coverage value⁴⁶. The percentage similarity of each read to the contig was also calculated and averaged for calculation of total percentage identity.

Functional assignment of reads and statistical analyses. All available pyrosequencing reads from faecal microbiomes and VLP-derived viromes were used to query (BLASTX, $E < 10^{-5}$) the KEGG (v51) and COG/String (v8⁴²) databases. The same databases were queried (BLASTP, $E < 10^{-5}$) with known and predicted proteins encoded by the 121 reference-sequenced human gut microbial genomes, and with all viral genomes (excluding prophages) in our NR_Viral_DB. After best BLAST hits were assigned to COG categories or KEGG second-level pathways, METASTATS¹⁸ was used to identify significant functional differences ($P < 0.05$) between faecal virome and microbiome data sets.

Prophage coverage plots. Prophages present in the 121 gut microbial genomes were identified using PHAGEFINDER. Each identified prophage was then extracted *in silico* together with 50 kb of flanking bacterial genomic sequences. NUCMER (v3.06)⁴⁷ was subsequently used to map all VLP pyrosequencer reads (default settings) onto this set of extracted sequences. MUMMERPLOT (v3.5), which like NUCMER is part of the MUMMER package⁴⁷, was used to generate sequence identity plots (threshold, >80% similarity). The prophage genome coordinates of the matches and the sequence identity with VLP reads were used to generate tables of 'percentage coverage' and 'fold-coverage'.

Search for integrase genes. All integrase protein sequences were extracted from the NR_Viral_DB and from the 121 sequenced human gut-associated microbial genomes. VLP pyrosequencing reads were BLASTed against this database of extracted sequences (BLASTP, $E < 10^{-4}$) and the best BLAST hit stored for every read. The number of reads from a given sample that were similar to a given integrase in the database was recorded and used to generate a Hellinger-based distance matrix between samples (QIIME).

CRISPR spacers represented in viral metagenomes. Seventy-four available human gut microbial genomes, representing members of the most predominant bacterial

families present in the human faecal microbiota, were used to search for CRISPR elements with CRISPRFINDER⁴⁸. CRISPRs were identified in 48 of these genomes: they contained a total of 95 different repeat sequences and 2,196 spacers. The direct repeats were subsequently compiled into a database, which in turn was used to search each of our faecal microbiome data sets (using CROSS_MATCH (v1.090518)⁴⁹; parameters: -minmatch 7 -maxmatch 12 -gap1_only -screen -minscore 10). Spacers were then extracted from all microbiome pyrosequencer reads where at least two matches for the same CRISPR repeat were identified and the intervening spacer was ≥ 10 nucleotides. All of these spacers from the microbiomes and 121 sequenced-reference genomes were pooled together and used to screen all VLP-derived pyrosequencing reads using CROSS_MATCH (parameters: -minmatch 14 -maxmatch 14 -gap1_only -screen -minscore 10). Virome reads with hits to microbiome or reference-genome CRISPR spacers over >90% of the length of their spacers were recorded.

Gnotobiotic mouse experiments. All studies with mice used protocols approved by the Washington University Animal Studies Committee. Methods for co-colonization of germ-free adult male C57Bl/6J mice with *Marvinbryantia formatexigens* and *Bacteroides thetaiotaomicron*, collection of their caecal contents, preparation of rRNA-depleted RNA from caecal contents and faecal samples for subsequent cDNA synthesis, and Illumina GA-IIx sequencing of complementary DNA, plus mapping and normalization of the resulting reads, are described in another publication²³. RNA-Seq data sets used for our analysis of prophage gene expression can be found under GEO accession numbers GSM544893, GSM544900, GSM544858, GSM544866, GSM544940 and GSM544944 (*in vivo* data), and GSM544856, GSM544873, GSM544835, GSM544947, GSM544872, GSM544917, GSM544931, GSM544871, GSM544883, GSM544863 and GSM544865 (*in vitro* data).

Statistical tests. Statistical tests were performed and heat maps were produced using the R package (v2.9.2)⁵⁰.

33. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* **3**, 1314–1317 (2009).
34. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5**, 235–237 (2008).
35. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
36. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
37. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
38. Angly, F. E. *et al.* The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.* **5**, e1000593 (2009).
39. Angly, F. E. *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41 (2005).
40. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
41. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
42. Jensen, L. J. *et al.* STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412–D416 (2009).
43. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
44. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
45. Faith, J. J., Olson, A. J., Gardner, T. S. & Sachidanandam, R. Lightweight genome viewer: portable software for browsing genomics data in its chromosomal context. *BMC Bioinformatics* **8**, 344 (2007).
46. Rohwer, F., Seguritan, V., Choi, D. H., Segall, A. M. & Azam, F. Production of shotgun libraries using random amplification. *Biotechniques* **31**, 108–112, 114–106 (2001).
47. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
48. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52–W57 (2007).
49. Gordon, D., Desmarais, C. & Green, P. Automated finishing with Autofinish. *Genome Res.* **11**, 614–625 (2001).
50. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna) (<http://www.r-project.org>) (2009).

A hot compact dust disk around a massive young stellar object

Stefan Kraus¹, Karl-Heinz Hofmann², Karl M. Menten², Dieter Schertl², Gerd Weigelt², Friedrich Wyrowski², Anthony Meilland^{2,3}, Karine Perraut⁴, Romain Petrov³, Sylvie Robbe-Dubois³, Peter Schilke⁵ & Leonardo Testi⁶

Circumstellar disks are an essential ingredient of the formation¹ of low-mass stars. It is unclear, however, whether the accretion-disk paradigm can also account for the formation of stars more massive than about 10 solar masses², in which strong radiation pressure might halt mass infall^{3,4}. Massive stars may form by stellar merging⁵, although more recent theoretical investigations suggest that the radiative-pressure limit may be overcome by considering more complex, non-spherical infall geometries^{6,7}. Clear observational evidence, such as the detection of compact dusty disks⁸ around massive young stellar objects, is needed to identify unambiguously the formation mode of the most massive stars. Here we report near-infrared interferometric observations that spatially resolve the astronomical-unit-scale distribution of hot material around a high-mass (~20 solar masses) young stellar object. The image shows an elongated structure with a size of $\sim 13 \times 19$ astronomical units, consistent with a disk seen at an inclination angle of $\sim 45^\circ$. Using geometric and detailed physical models, we found a radial temperature gradient in the disk, with a dust-free region less than 9.5 astronomical units from the star, qualitatively and quantitatively similar to the disks observed in low-mass star formation. Perpendicular to the disk plane we observed a molecular outflow and two bow shocks, indicating that a bipolar outflow emanates from the inner regions of the system.

We investigated the massive young stellar object IRAS 13481-6124, which harbours a central object of about 20 solar masses ($20M_\odot$) (ref. 9) embedded in a cloud with a total gas mass of $\sim 1,470 M_\odot$ (ref. 10). Our imaging observations covered wavelengths from 2 to 870 μm and spatial scales from several parsecs to astronomical units (AU). In archival Spitzer Space Telescope images¹¹ (Fig. 1a), we discovered two faint bow shocks, which are oriented along position angle $\text{PA} = 31 \pm 6^\circ$ and separated by $7''$, corresponding to a physical scale of ~ 6.5 pc at a distance of 3,500 pc (ref. 12). We also observed the source with the Atacama Pathfinder Experiment (APEX) 12-m submillimetre telescope, located at the Llano de Chajnantor Observatory in Chile. Using observations in emission lines of molecular gas, we detected a bipolar outflow, which is oriented along the same axis ($\text{PA} = 26 \pm 9^\circ$, Fig. 1b) as the bow shocks. The detected outflow signatures have a relatively narrow bow-shock opening angle of $\sim 6^\circ$ and show similarities with the bipolar collimated jets observed in low-mass star formation. Outflows from young stars are powered by accretion¹³ and require a magnetized, compact accretion disk to collimate the ejected material¹⁴, so the detected outflow provides some indirect evidence that there is a disk around IRAS 13481-6124.

In order to directly detect this disk and to characterize its inner structure, we used the Very Large Telescope Interferometer (VLTI) of the European Southern Observatory (ESO), located on Cerro Paranal

in Chile. Using the near-infrared beam combination instrument AMBER¹⁵, we combined the light of three of the VLTI 1.8-m telescopes and measured the interferometric observables (visibilities and closure phases) in 17 spectral channels in the K-band (wavelength, $\lambda = 1.95\text{--}2.55 \mu\text{m}$). We obtained the VLTI observations using three different three-telescope array configurations and complemented them with speckle-interferometric observations from the New Technology Telescope, providing precise spatial information for baseline lengths of less than 3.5 m.

Using our extensive VLTI + speckle data set, we reconstructed¹⁶ a model-independent interferometric image from the measured visibility amplitudes and closure phases. This technique, which is routinely applied in radio interferometry, has recently been demonstrated with the VLTI¹⁷. We applied it in our study to image the circumstellar environment around a young star. With baseline lengths B up to 85 m, our image provides an angular resolution of $\lambda/2B = 2.4$ milliarcsecond (mas) or ~ 8.4 AU, which is at least one order of magnitude higher than conventional infrared imaging techniques at 10-m-class telescopes; the gain compared to state-of-the-art submillimetre disk studies is about two orders of magnitude. The reconstructed image (Fig. 1c) clearly resolves the inner environment around IRAS 13481-6124 and reveals a compact elongated structure with a size of $5 \text{ mas} \times 8 \text{ mas}$. The disk orientation ($\text{PA} = 120^\circ$) is perpendicular to the determined outflow axis ($26 \pm 9^\circ$ and $31 \pm 6^\circ$), suggesting that the AU-scale disk resolved by VLTI/AMBER is indeed the driving engine of the detected collimated outflow.

To characterize further the detected elongated structure, we fitted geometric and detailed physical models to our data, allowing us to deduce object information smaller than the diffraction-limited resolution of our reconstructed image. Our measurements show that the visibility function drops rapidly at baseline lengths of less than 3 m, but then stays nearly constant up to baseline lengths of about 12 m, followed by a uniform, nearly linear decline at baseline lengths up to 85 m (Supplementary Fig. 10), providing clear evidence that there are at least two spatial components. We used a Gaussian model to estimate the characteristic size (full-width at half-maximum diameter) of the two components and found that the first, extended component has a size of about 108 mas and contributes approximately 15% of the total flux. On the basis of the discovery of scattered-light envelopes around other massive stars¹⁸, we suggest that this extended centrosymmetric component is the scattered-light contribution of a natal envelope. The second, compact component has a size of 5.4 mas, accounts for 85% of the total K-band flux and is significantly elongated along $\text{PA} = 114^\circ$. Furthermore, we found that the size of the compact component increases significantly towards longer wavelengths, as shown by the measured equivalent widths in Fig. 2a. This effect, which was found

¹Department of Astronomy, University of Michigan, 500 Church Street, Ann Arbor, Michigan 48103, USA. ²Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany. ³Laboratoire Hippolyte Fizeau, UMR 6525, Université de Nice Sophia-Antipolis CNRS, Observatoire de la Côte d'Azur, Parc Valrose, 06108 Nice cedex 2, France.

⁴Laboratoire d'Astrophysique de Grenoble, UMR 5571, Université Joseph Fourier CNRS, BP 53, 38041 Grenoble cedex 9, France. ⁵Physikalisches Institut, Universität zu Köln, Zùlpicher Strasse 77, 50937 Cologne, Germany. ⁶INAF-Osservatorio Astrofisico di Arcetri, Largo Fermi 5, 50125 Firenze, Italy.

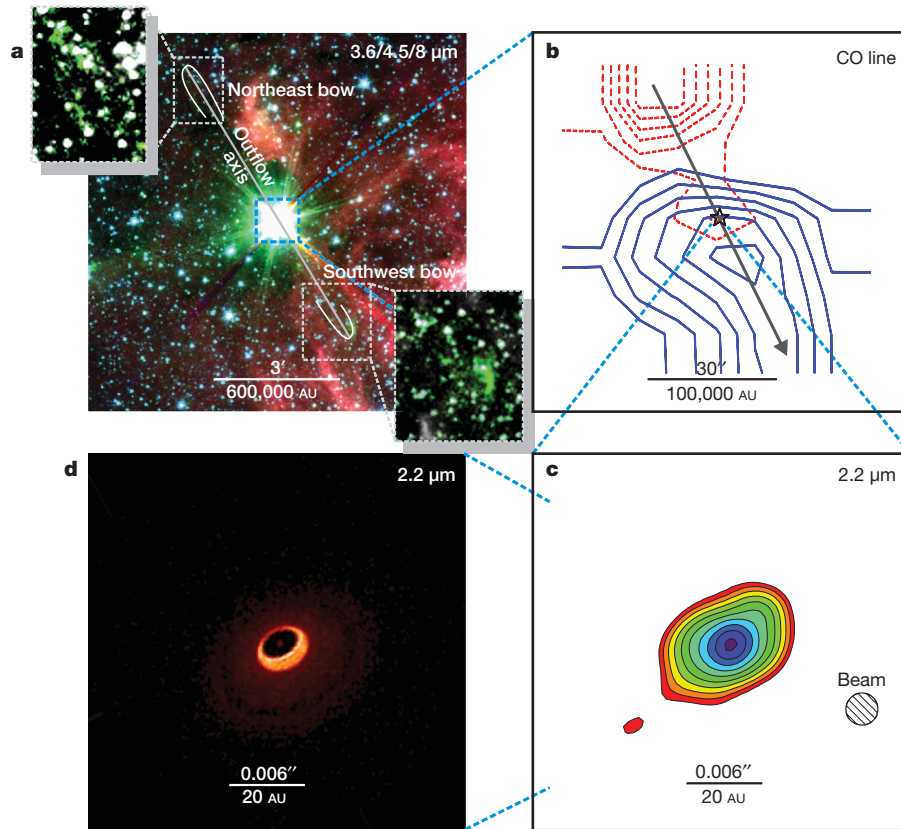


Figure 1 | Zoom in on IRAS 13481-6124, covering structures over more than five orders of magnitude. **a**, In Spitzer/Infrared Array Camera (IRAC) images, we detect two bow-shock structures, indicating a collimated bipolar outflow. The bows (insets) are separated by 7' and appear as excess emissions in the IRAC 4.5- μm band (green), probably tracing shocked molecular hydrogen gas³⁰. The colours in the composite image correspond to wavelength-bands around 3.6 μm (blue), 4.5 μm (green), and 8 μm (red). **b**, The outflow is also detected in molecular line emissions using the APEX/Swedish Heterodyne Facility Instrument on scales of a few 10,000 AU, with the approaching (blueshifted) lobe southwest of IRAS 13481-6124 (solid contours show a blueshifted velocity of -10 km s^{-1} and dashed contours a redshifted velocity of $+5 \text{ km s}^{-1}$). We used the ^{12}CO (3–2) emission line at 867 μm as the outflow tracer line. **c**, VLT/AMBER aperture-synthesis imaging reveals an elongated structure perpendicular to the outflow direction (contours decrease from peak intensity by factors of $\sqrt{2}$). The

structure has a size of $5 \text{ mas} \times 8 \text{ mas}$, as measured at 10% of the peak flux intensity, and contributes about 80% of the total flux in the image; the remaining flux elements are spread uniformly over the image and correspond to the extended component detected in our model-fits. The image was reconstructed from an extensive data set of 33 independent VLT/AMBER three-telescope observations obtained around a wavelength of 2.2 μm and using three different three-telescope array configurations. **d**, The best-fit radiative-transfer model image as constrained by the measured wavelength-dependent visibilities and closure-phases and the SED from near-infrared to millimetre wavelengths (see also Supplementary Figs 13 and 14). The image was computed for a wavelength of 2.2 μm ; brighter colours indicate higher flux. This modelling allows us to determine the intensity profile on scales smaller than the diffraction-limited resolution, and strongly suggests that the dust disk is truncated at a radius of $6.2 \pm 1.2 \text{ AU}$.

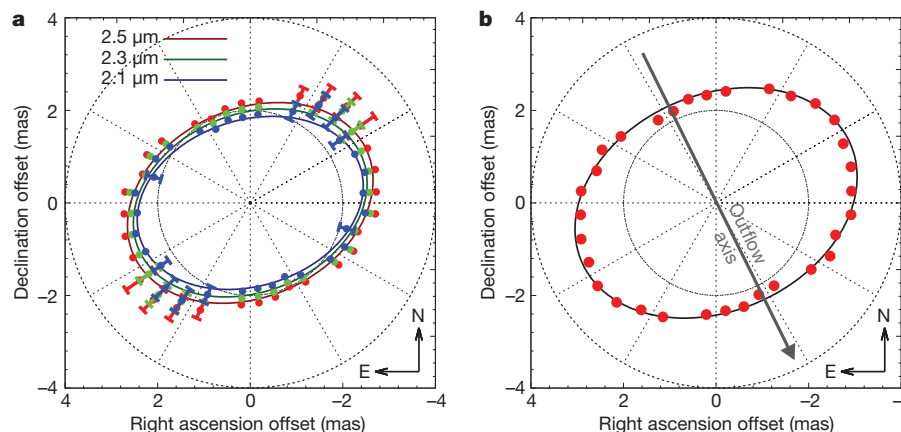


Figure 2 | Elongation of the compact emission component, as determined with a Gaussian and a temperature-gradient disk model. **a**, Using a Gaussian model, we determined the object size for different position-angle bins (each covering 10°) and wavelength channels centred on 2.1 μm (blue), 2.3 μm (green) and 2.5 μm (red), revealing strong object elongation and indications of a temperature gradient. The error bars give the standard deviation. **b**, Using a temperature power-law disk-model $T(r) = 2,000 \text{ K} (r/r_{\text{in}})^{-q}$, we determined

the temperature power-law index q to be ~ 0.4 and found an ellipsoidal source geometry, corresponding to a disk seen under an inclination angle of $\sim 45^\circ$. The truncated dust-disk model yields a significantly better representation of the measured wavelength-dependent visibilities ($\chi^2/\text{d.f.} = 1.1$) than standard geometric models (for example, Gaussians, $\chi^2/\text{d.f.} = 2.5$, or disks of uniform brightness, $\chi^2/\text{d.f.} = 4.9$; see Supplementary Information Section 4).

in many low- and intermediate-mass young stellar object disks^{19,20}, probably indicates a temperature gradient in the circumstellar material, in which hotter material (radiating more effectively at shorter wavelengths) is located closer to the star.

Motivated by these indications for an internal temperature structure, we fitted analytical disk models with a radial-temperature power-law $T(r) = T_{\text{in}}(r/r_{\text{in}})^{-q}$, where $T(r)$ is the temperature at radius r , T_{in} is the inner-disk temperature, r_{in} is the inner-disk radius and q is the temperature power-law index, to our data. Assuming reasonable values for T_{in} of 1,500–2,000 K, we found that this model could reproduce our interferometric data with an r_{in} of 3 mas (9.5 AU) and a q of 0.4, which is consistent with the theoretical temperature gradient²¹ of flared irradiated disks ($q = 0.43$). This disk model provides a significantly better representation of our data than do other geometric models or disk models with emissions extending to the star (Fig. 2b and Supplementary Information Section 4). The derived inner-disk radius of 9.5 AU agrees with the expected location of dust sublimation in an irradiated circumstellar disk: 6.2–10.9 AU, assuming grey dust and dust-sublimation temperatures between 1,500 and 2,000 K. Therefore, IRAS 13481-6124 follows the well established size–luminosity relation²² for low- to intermediate-mass young stellar objects, suggesting that the near-infrared emission mainly traces material at the dust sublimation radius, similarly to the disks around T Tauri and Herbig Ae/Be stars.

In addition to the rather general arguments presented above, we applied detailed physical models using multi-dimensional continuum radiative-transfer simulations. Encompassing the central star, disk and envelope, these models can provide a much more comprehensive and consistent picture of the IRAS 13481-6124 system than purely geometric models. Because these radiative-transfer models depend on many free parameters, we used a precomputed model grid²³ and an associated spectral energy distribution (SED)-fitting tool²⁴ to find a good initial set of model parameters. Using an adapted version of the radiative-transfer modelling code by Whitney *et al.*²⁵ we further refined the model parameters and found a model that could reproduce the SED and the VLTI + speckle observations simultaneously, yielding the model image shown in Fig. 1d (see Supplementary Figs 13 and 14 for visibilities, closure phases and the SED, and Supplementary Table 4 for the complete set of model parameters). Our global best-fit model parameters suggest a distance of 3,260 pc and include a central star with a mass of $18M_{\odot}$, a massive ($1,000M_{\odot}$) circumstellar envelope with a bipolar cone-shaped outflow cavity and a compact (~ 130 AU) flared disk with a mass similar to that of the central star ($20 \pm 8M_{\odot}$). Many model parameters, such as those concerning the envelope geometry and the disk mass, may still suffer from the parameter ambiguities that are inherent to SED-fitting results²⁴. Other parameters, such as the inner-disk radius and the disk orientation, are constrained by spatially resolved observations. The resolved disk lacks asymmetry, as is evident from the measured small closure phases and the symmetrical reconstructed image. For our radiative-transfer models, this low degree of asymmetry puts direct constraints on the vertical disk structure near the dust-sublimation region and suggests that the inner dust rim is smooth. As with findings for intermediate-mass stars, this probably indicates that the inner dust rim around IRAS 13481-6124 is curved (see also Supplementary Information Sections 3 and 5), although further theoretical and observational work is needed to characterize fully the detailed inner-rim geometry. In particular, future observations using mid-infrared and submillimetre interferometers could resolve the inner-disk structure over a very wide wavelength range, putting more constraints on the radial density structure and the more extended disk regions.

The measured low degree of asymmetry is also interesting in the context of numerical simulations of self-gravitating disks, which predict the formation of large-scale spiral density waves^{8,26} and a highly non-axisymmetric disk structure. Given that these asymmetries are not detected by our VLTI phase-closure observations, we conclude

that these processes are not dominant in the probed inner regions of the IRAS 13481-6124 disk, or are not sufficiently resolved. Our speckle and VLTI data also rule out the presence of nearby stellar companions (down to flux ratios of $\sim 1:40$ and separations exceeding 10 mas), contrasting with the prediction of compact stellar clusters in competitive accretion and stellar merger models²⁷. Our observations suggest that IRAS 13481-6124 has already gone through the main accretion phase in its evolution, and that the disk is heated mostly by stellar irradiation. The derived disk structure, which closely resembles the irradiated dust disks around Herbig Ae/Be stars but is distinctly different from the actively accreting disks observed around FU Orionis stars, for instance, supports this conclusion. According to our radiative-transfer model and the underlying stellar evolutionary tracks²⁸, the system has an evolutionary age of about 60,000 years, roughly matching the expected age for the start of disk evaporation, 100,000 years (ref. 29). Therefore, we are probably observing the system in a short-lived phase, in which the strong stellar winds, radiation pressure and photoevaporation have already stopped the mass infall, and are just starting to dissipate the circumstellar disk.

Received 15 April; accepted 11 May 2010.

- McKee, C. F. & Ostriker, E. C. Theory of star formation. *Annu. Rev. Astron. Astrophys.* **45**, 565–687 (2007).
- Zinnecker, H. & Yorke, H. W. Toward understanding massive star formation. *Annu. Rev. Astron. Astrophys.* **45**, 481–563 (2007).
- Kahn, F. D. Cocoons around early-type stars. *Astron. Astrophys.* **37**, 149–162 (1974).
- Wolfire, M. G. & Cassinelli, J. P. Conditions for the formation of massive stars. *Astrophys. J.* **319**, 850–867 (1987).
- Bonnell, I. A., Bate, M. R. & Zinnecker, H. On the formation of massive stars. *Mon. Not. R. Astron. Soc.* **298**, 93–102 (1998).
- McKee, C. F. & Tan, J. C. Massive star formation in 100,000 years from turbulent and pressurized molecular clouds. *Nature* **416**, 59–61 (2002).
- Krumholz, M. R. *et al.* The formation of massive star systems by accretion. *Science* **323**, 754–757 (2009).
- Cesaroni, R., Galli, D., Lodato, G., Walmsley, C. M. & Zhang, Q. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) 197–212 (University of Arizona Press, 2007).
- Grave, J. M. C. & Kumar, M. S. N. Spitzer-IRAC GLIMPSE of high mass protostellar objects. II. SED modelling of a bona fide sample. *Astron. Astrophys.* **498**, 147–159 (2009).
- Beltrán, M. T. *et al.* Search for massive protostellar candidates in the southern hemisphere. II. Dust continuum emission. *Astron. Astrophys.* **447**, 221–233 (2006).
- Benjamin, R. A. *et al.* GLIMPSE. I. An SIRTf legacy project to map the inner Galaxy. *Publ. Astron. Soc. Pacif.* **115**, 953–964 (2003).
- Fontani, F. *et al.* Search for massive protostellar candidates in the southern hemisphere. I. Association with dense gas. *Astron. Astrophys.* **432**, 921–935 (2005).
- Cabrit, S., Edwards, S., Strom, S. E. & Strom, K. M. Forbidden-line emission and infrared excesses in T Tauri stars — evidence for accretion-driven mass loss? *Astrophys. J.* **354**, 687–700 (1990).
- Blandford, R. D. & Payne, D. G. Hydromagnetic flows from accretion discs and the production of radio jets. *Mon. Not. R. Astron. Soc.* **199**, 883–903 (1982).
- Petrov, R. G. *et al.* AMBER, the near-infrared spectro-interferometric three-telescope VLTI instrument. *Astron. Astrophys.* **464**, 1–12 (2007).
- Hofmann, K.-H. & Weigelt, G. Iterative image reconstruction from the bispectrum. *Astron. Astrophys.* **278**, 328–339 (1993).
- Kraus, S. *et al.* Tracing the young massive high-eccentricity binary system θ^1 Orionis C through periastron passage. *Astron. Astrophys.* **497**, 195–207 (2009).
- Tobin, J. J. Imaging scattered light from the youngest protostars in L1448: signatures of outflows. *Astrophys. J.* **659**, 1404–1419 (2007).
- Eisner, J. A. *et al.* Spectrally dispersed K-band interferometric observations of Herbig Ae/Be sources: inner disk temperature profiles. *Astrophys. J.* **657**, 347–358 (2007).
- Kraus, S., Preibisch, T. & Ohnaka, K. Detection of an inner gaseous component in a Herbig Be star accretion disk: near- and mid-infrared spectrophotometry and radiative transfer modeling of MWC 147. *Astrophys. J.* **676**, 490–508 (2008).
- Chiang, E. I. & Goldreich, P. Spectral energy distributions of T Tauri stars with passive circumstellar disks. *Astrophys. J.* **490**, 368–376 (1997).
- Monnier, J. D. *et al.* The near-infrared size–luminosity relations for Herbig Ae/Be disks. *Astrophys. J.* **624**, 832–840 (2005).
- Robitaille, T. P., Whitney, B. A., Indebetouw, R., Wood, K. & Denzmore, P. Interpreting spectral energy distributions from young stellar objects. I. A grid of 200,000 YSO model SEDs. *Astrophys. J.* **167**, 256–285 (2006).

24. Robitaille, T. P., Whitney, B. A., Indebetouw, R. & Wood, K. Interpreting spectral energy distributions from young stellar objects. I. Fitting observed SEDs using a large grid of precomputed models. *Astrophys. J.* **169**, 328–352 (2007).
25. Whitney, B. A., Wood, K., Bjorkman, J. E. & Wolff, M. J. Two-dimensional radiative transfer in protostellar envelopes. I. Effects of geometry on class I sources. *Astrophys. J.* **591**, 1049–1063 (2003).
26. Bodenheimer, P. Angular momentum evolution of young stars and disks. *Annu. Rev. Astron. Astrophys.* **33**, 199–238 (1995).
27. Bonnell, I. A. & Bate, M. R. Binary systems and stellar mergers in massive star formation. *Mon. Not. R. Astron. Soc.* **362**, 915–920 (2005).
28. Bernasconi, P. A. & Maeder, A. About the absence of a proper zero age main sequence for massive stars. *Astron. Astrophys.* **307**, 829–839 (1996).
29. Hollenbach, D. J., Yorke, H. W. & Johnstone, D. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 401–428 (Univ. Arizona Press, 2000).
30. Smith, H. A. *et al.* Outflows from massive young stellar objects as seen with the infrared array camera. *Astrophys. J.* **645**, 1264–1271 (2006).

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

Acknowledgements This work was done in part under contract with the California Institute of Technology (Caltech), funded by NASA through the Sagan Fellowship

Program (S.K. is a Sagan Fellow). We thank the ESO Paranal staff for support and their efforts in improving the VLT. This paper is based on observations made with ESO telescopes at the La Silla Paranal Observatory and archival data obtained with the Spitzer Space Telescope, operated by the Jet Propulsion Laboratory, Caltech, under a contract with NASA. We also used data acquired with APEX, a collaboration between the Max-Planck-Institut für Radioastronomie, ESO, and the Onsala Space Observatory.

Author Contributions S.K. worked on the AMBER data reduction and data interpretation, model fitting and image reconstruction, made some of the observations and wrote the telescope proposals and the initial paper manuscript. K.-H.H. worked on the image reconstruction. D.S. worked on the speckle data reduction. G.W. made some of the observations. F.W. worked on the APEX data reduction and data interpretation. K.M.M. and P.S. were co-authors on the telescope proposal. A.M., K.P., R.P., S.R.-D. and L.T. represent the AMBER instrument consortium, which provided some observation time. All authors commented on the paper.

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

Transmission of topological surface states through surface barriers

Jungpil Seo¹, Pedram Roushan¹, Haim Beidenkopf¹, Y. S. Hor², R. J. Cava² & Ali Yazdani¹

Topological surface states are a class of novel electronic states that are of potential interest in quantum computing or spintronic applications^{1–7}. Unlike conventional two-dimensional electron states, these surface states are expected to be immune to localization and to overcome barriers caused by material imperfection^{8–14}. Previous experiments have demonstrated that topological surface states do not backscatter between equal and opposite momentum states, owing to their chiral spin texture^{15–18}. However, so far there is no evidence that these states in fact transmit through naturally occurring surface defects. Here we use a scanning tunnelling microscope to measure the transmission and reflection probabilities of topological surface states of antimony through naturally occurring crystalline steps separating atomic terraces. In contrast to non-topological surface states of common metals (copper, silver and gold)^{19–23}, which are either reflected or absorbed by atomic steps, we show that topological surface states of antimony penetrate such barriers with high probability. This demonstration of the extended nature of antimony's topological surface states suggests that such states may be useful for high current transmission even in the presence of atomic-scale irregularities—an electronic feature sought to efficiently interconnect nanoscale devices.

Topological surface states occur on compounds for which strong spin–orbit interaction influences the bulk electronic band structure in a fundamental way^{1–7}. The key experimental signatures of topological surface states are an odd number of crossings of their energy–momentum dispersion at the Fermi level and chiral spin texture that locks the surface electrons' momenta to their spin angular momenta^{8–11,15–18}. Angle-resolved photoemission spectroscopy (ARPES) measurements have observed both signatures of topological surface states for a number of compounds with a bulk gap, the so-called topological insulators such as Bi_{1–x}Sb_x, Bi₂Se₃ and Bi₂Te₃, and on the (111) surface of the semimetal antimony^{8–11} (Sb). The two features of topological surface states alter their response to crystalline imperfections, first by eliminating 180° backscattering, owing to scattering restriction imposed by their spin texture, and second by the virtue of the odd number of band crossings, which prevents the surface states from being fully gapped when perturbed. The combination of these features makes the topological surface states robust against localization; hence, they are predicted to wrap the surface of the topological bulk sample regardless of the presence of surface defects. Previous scanning tunnelling microscopy (STM) experiments have provided evidence for the absence of 180° backscattering of these surface states by examining the scattering from substitutional impurities and atomic step edges^{15–18}. However, these experiments do not demonstrate whether topological surface states are in fact extended, as they do not probe the transmission properties of these states. In fact, it is possible that in the presence of spin–orbit interaction, 180° backscattering is forbidden but electronic states can still

be localized by surface defects. An example of this situation is the spin-split surface states of Au(111) that become localized by scattering from atomic step edges^{22,23}. Topological surface states must be immune to such localization. To test this unique feature of topological surface states, we have used STM to examine antimony's surface states and to measure their reflection by and transmission through atomic steps.

We have studied the (111) surface of antimony in a home-built cryogenic STM that operates at 4 K. Single crystals of antimony were cleaved *in situ* in ultrahigh vacuum to expose a pristine surface. The STM topographic image shows nanometre-sized atomic terraces that are separated by single atom-height steps along the [110] crystallographic direction (Fig. 1a). We have examined the spatial and energy dependence of the local density of states (LDOS) within the terraces by performing differential conductance (dI/dV) measurements along a line perpendicular to the step edges. In Fig. 1b, the spatially averaged spectroscopic measurements for terraces of different widths show peaks that signal the occurrence of quantized resonances. These spectroscopic features start at around –225 meV, which is the bottom of the topological surface state bands as measured by ARPES studies^{8,18}. Resolving the spatial variation of the LDOS within each terrace (Fig. 1c), we find clear evidence for quantized resonances, with nearly equal spacing in energy (see also Supplementary Information). The quantized resonances and the standing-wave patterns in the LDOS are caused by scattering of antimony's topological surface states from the atomic step edges and can be analysed to obtain reflection properties of these steps.

We first demonstrate that the standing-wave patterns and the nearly linear quantized resonances in LDOS can be understood on

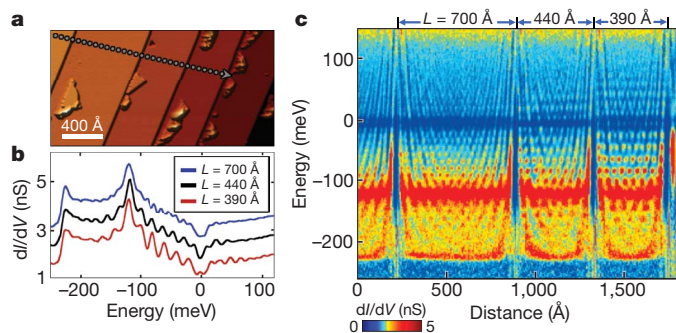


Figure 1 | The topological surface states of Sb(111) on atomic terraces. **a**, STM topographic image ($V_{\text{bias}} = 1$ V, $I = 8$ pA) of a 2,500 Å-by-1,250 Å area showing terraces of various widths separated by 3.7 Å-high single atomic steps. **b**, The dI/dV measurement averaged over each terrace shows quantized peaks. The spectra are offset vertically for clarity. **c**, Spatially and energetically resolved dI/dV measurements taken along the dotted arrow in **a** demonstrate the interference in space and the quantization in energy.

¹Joseph Henry Laboratories & Department of Physics, Princeton University, Princeton, New Jersey 08544, USA. ²Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA.

the basis of the ARPES band structure of antimony's surface states and the spin selection rules that must be obeyed for scattering of topological surface states. The Fourier transform (Fig. 2b) of the dI/dV measurement on the 390 Å-wide terrace (Fig. 2a) reveals two different quantized wavevectors, q_A and q_B , for the standing-wave pattern of the LDOS. Unlike conventional confined states that involve the superposition of states with equal and opposite momenta^{24–27} (k), the chiral spin texture of antimony eliminates the possibility of superposition of states with orthogonal spins. Consequently, the wavevectors (q) observed correspond to superposition between states with different values of k but similar spin states. From the schematic ARPES measurements of antimony's constant-energy contours (Fig. 2c) and their spin texture (Fig. 2d), we identify q_A with scattering between adjacent hole pockets, which accordingly vanishes once the hole pockets overlap each other (around -110 meV in Fig. 2c). Similarly, q_B corresponds to superposition between the electron and hole pockets with oppositely oriented momenta and parallel spins¹⁸ (Fig. 2d). The scattering from the step edges and the geometric constraint set by the terrace width, L , result in quantization of these allowed superpositions of momentum states, such that $q_n = 2\pi n/L$, where n is a positive integer and q_n is the corresponding scattering wavevector. Despite their quantization, the energy dependences of q_A and q_B are related to the dispersion measured by ARPES bands because $q_n(E_n) = k_i(E_n) - k_f(E_n)$, where k_i and k_f are respectively the initial and final momentum states in an elastic scattering with energy E_n . The nearly linear energy dependences of q_A and q_B have slopes (~ 1.2 eV Å) that are in excellent agreement with ARPES measurements.

The energy widths of the quantized resonances in the LDOS contain information on the scattering properties of antimony's topological surface states. In general, energy-level broadening, Γ , is inversely proportional to the lifetime of a state, τ : $\Gamma \propto \hbar/\tau$ ($\hbar$, Planck's constant divided by 2π). Examining the broadening of the resonances as a function of energy for different terraces (Fig. 3), we find that $\Gamma(E) = \Gamma_L + \gamma(E - E_F)^2$, where E_F is the Fermi energy (with $\hbar/\gamma = 0.37 \pm 0.01$ fs eV², for terraces where level broadening is smaller than level spacing). This functional form illustrates the role of electron–electron scattering in the decoherence of topological

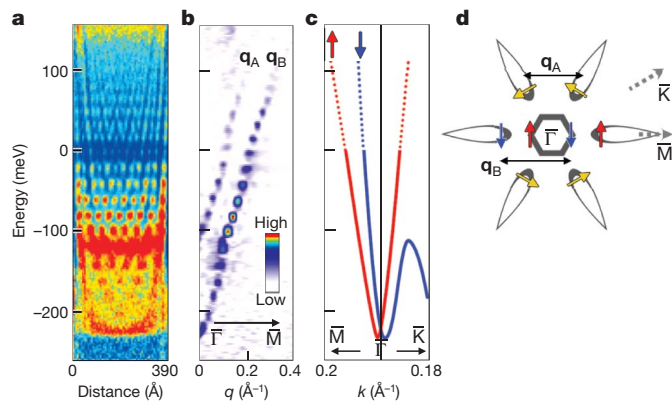


Figure 2 | Allowed scattering wavevectors and their quantization. **a**, The dI/dV measurement on a 390 Å-wide terrace. **b**, Energy-resolved Fourier transform of the spatial modulation of the data in **a** reveals the quantization of scattering wavevectors q_A and q_B . **c**, The dispersions of q_A and q_B match the dispersion of the surface bands as measured by ARPES along the high-symmetry directions (solid lines) and extend it above the Fermi level (dotted lines). **d**, Contours of constant energy of the antimony surface state (ARPES measurements from ref. 8). The contours consist of a central electron pocket and six surrounding hole pockets. The coloured arrows represent spin texture of the surface state. The scattering wavevectors q_A and q_B indicate allowed scattering processes. Γ , M and K are the high-symmetry points of the first Brillouin zone, with Γ located at the centre of the zone, M located in the middle of each side of this six-fold symmetric zone, and K located at the vertex.

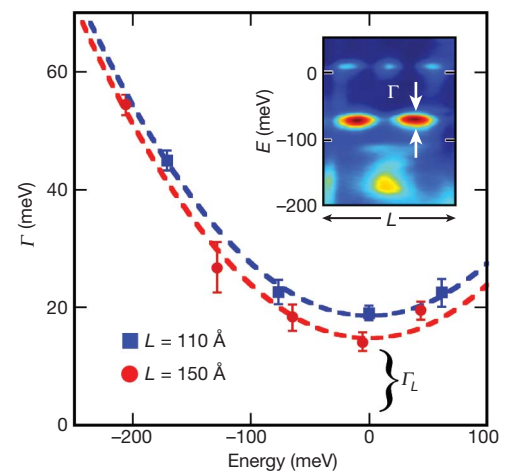


Figure 3 | Lifetime and leakage of quantized quasiparticles. Each resonant peak is fitted to a Lorentzian function to yield the full-width at half-maximum of the quantized energy peaks, Γ (inset). The plot shows the energy dependence of Γ for two different terraces. The dashed lines are parabolic fits, and Γ_L is the peak broadening at the Fermi energy. The error bars are from the fitting process of the resonant peaks and are mainly due to the measurement resolution.

surface states, similar to the case in other Fermi liquids²⁰. However, Γ_L , the residual finite resonance width at E_F , characterizes the degree to which step edges can reflect antimony's topological surface states in the same spin channel. Using a simple model of energy broadening based on a double-barrier potential, we find that the energy broadening corresponds to a $42 \pm 4\%$ reflection probability for this process (Supplementary Information). This reflectivity is much lower than those reported for free-electron-like surface states on copper, silver or gold from STM and other measurements; however, it can be a result of scattering of the surface states into the bulk states rather than their transmission. Indeed, previous studies show that non-topological noble-metal surface states are absorbed in the bulk states with 30–50% probability without transmission when scattered by step edges^{19–23}.

Finite transmission through the atomic step edges should give rise to coupling between electronic states on adjacent terraces. To search for evidence of such effects, we examine a configuration of atomic steps consisting of a narrow (110 Å) and a wide (2,500 Å) terrace, which resembles a nanoscale Fabry–Pérot resonator, as shown in Fig. 4. The narrow terrace shows the signature of quantized resonances according to the spin scattering rules described above (Fig. 4b). On the wide terrace, which is effectively semi-infinite, the LDOS oscillation forms an almost continuous pattern (Fig. 4c) with diverging wavelength towards the bottom of the surface state bands. The standing-wave patterns on the wide terrace also obey the chiral spin texture and are made up of the two wavevectors q_A and q_B described above, as illustrated in the Fourier transform in Fig. 4d. Any finite transmission through the step edges results in resonant tunnelling of surface state electrons from the wide terrace through the electron states of the narrow terrace.

Evidence for resonant tunnelling of the topological surface states can be seen both in the structure of energy-resolved modulations of the LDOS and in its Fourier transform (Fig. 4). At energies corresponding to the quantized resonances of the narrow terrace, there are clear suppressions of the modulation in the LDOS on the wide terrace. We illustrate this in Fig. 4e by plotting the Fourier intensity of the q_B peak as a function of energy. Once the overall background changes in the LDOS (Fig. 4e, dashed green line) are removed, the intensity of the q_B Fourier component displays characteristic signs of Fabry–Pérot resonant tunnelling (Fig. 4f) in which the surface state electrons pass through the narrow terrace without reflection. Deviation from perfect transmission is, however, a signature of scattering of the surface states into the antimony bulk states, from which

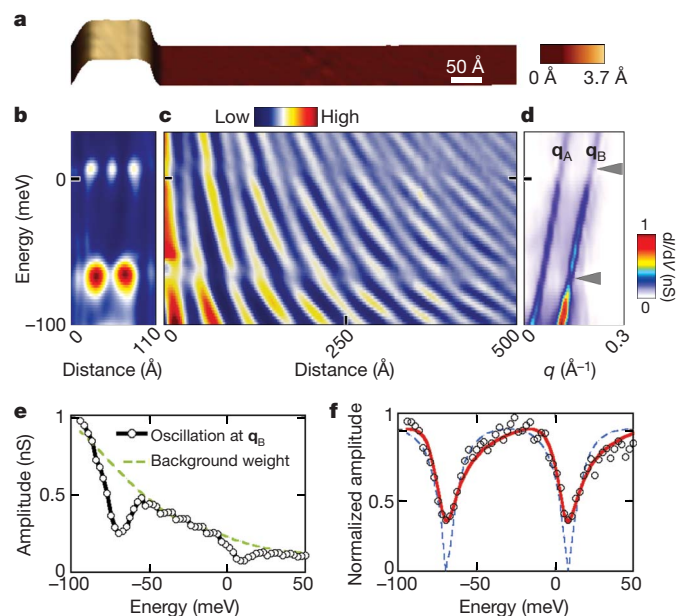


Figure 4 | Resonant tunnelling between adjacent terraces. **a**, STM topographic image of a narrow terrace ($L = 110 \text{ \AA}$) and an adjacent wide terrace ($L = 2,500 \text{ \AA}$). Only part of the wide terrace is displayed. The colour scale shows the height variations in the topographic image. **b**, On the narrow terrace, the dI/dV measurement shows the quantization of the energy levels. **c**, The dI/dV measurement on the wide terrace shows sudden phase shifts and suppressions of the dI/dV intensities at around $+5 \text{ meV}$ and -70 meV due to resonant tunnelling. The averaged background conductance at each energy has been subtracted. **d**, The Fourier transform of the spatial modulation of the data in **c**. The grey markers indicate suppression of modulation intensities. **e**, Spectral weight of the scattering wavevector q_B as extracted from **d**. The dashed line corresponds to the background spectral weight in the absence of resonance tunnelling. **f**, Spectral weight (circles) normalized by the background. The best fit (red line) yields $\sim 42\%$ reflection, $\sim 35\%$ transmission and $\sim 23\%$ bulk absorption from the scattering process at the boundary. The blue line, which is the fit based on a model without bulk absorption, is displayed for comparison.

we can deduce that the probability of this process is $23 \pm 7\%$ (Supplementary Information). Similarly, the width of the Fabry–Pérot resonance (Fig. 4f) can be used to extract the transmission and reflection probabilities and to show that they are $35 \pm 3\%$ and $42 \pm 4\%$, respectively. In contrast, STM data on Fabry–Pérot structures for non-topological surface states of Ag(111) show no evidence of resonant tunnelling and can be explained in detail by considering only reflections and absorption of the surface states at the step edges²¹.

The remarkable finding that topological surface states are as likely to be transmitted through an atomic step edge as reflected by it demonstrates an important difference between topological states and typical surface states in other materials. The transmissibility illustrates the extended nature of topological surface states even in the presence of strong surface barriers. The nearly equal probabilities of reflection and transmission at the step edges in antimony can also be understood on the basis of the topological surface state's band structure, which includes both forwards- and backwards-moving momenta with the same spin orientation (Fig. 2). The observation of high transmission also suggests excellent transport of topological surface states despite boundaries that would suppress the ability of other surface states to carry current. Such a feature may find applications in nanoscale devices, where there is a need for high surface conductivity because the surface-to-volume ratio is large.

Our experiments also show that nanoscale structures can be used to explore the properties of topological surface states both for fundamental studies and to evaluate their potential for device applications. Clearly, an extension of these studies to materials with a bulk

gap and simpler surface band structure (such as in Bi_2Te_3 or Bi_2Se_3) would allow further examination of these states without the complication of coupling to the bulk states. However, antimony's surface band structure allows STM characterization of both reflection and transmission scattering rates within the same spin channel, which would not be possible for a single Dirac-cone surface band structure. Nevertheless, our demonstrations of the transmission of topological surface states through barriers that fully reflect other surface states suggest an extraordinary insensitivity of transport through these states to the geometrical shape of the surface.

METHODS SUMMARY

Crystal growth. We cleaved a single crystal of antimony from a boule grown by the modified Bridgman method. Before the growth, high-purity antimony (99.999%) was sealed in a vacuum quartz tube and heated to 700°C . The crystal growth process took about a day at a slow rate of cooling from 700 to 500°C . We then kept the boule at 500°C for annealing over five days. The boule was then furnace-cooled to room temperature (25°C).

STM measurement. We performed the experiment using a home-built cryogenic STM that operates at 4 K in ultrahigh vacuum. The typical sample size used was about $2 \text{ mm} \times 2 \text{ mm} \times 1 \text{ mm}$. The sample was cleaved *in situ* at room temperature in ultrahigh vacuum before being measured at low temperature. The cleaved sample revealed a (111) surface orientation due to the rhombohedral antimony crystal structure. A mechanically sharpened platinum–iridium alloy wire was used as a STM tip, and the quality of the tip was checked by imaging the surface states of a copper sample. The STM topography was measured in a constant-current mode, and dI/dV signals were obtained by a standard lock-in technique with frequency $f = 757 \text{ Hz}$ and 1-mV root-mean-squared modulation.

Received 25 February; accepted 19 May 2010.

- Kane, C. L. & Mele, E. J. Z_2 topological order and the quantum spin Hall effect. *Phys. Rev. Lett.* **95**, 146802 (2005).
- Moore, J. E. & Balents, L. Topological invariants of time-reversal-invariant band structures. *Phys. Rev. B* **75**, 121306(R) (2007).
- Fu, L., Kane, C. L. & Mele, E. J. Topological insulators in three dimensions. *Phys. Rev. Lett.* **98**, 106803 (2007).
- Roy, R. Z_2 classification of quantum spin Hall systems: an approach using time-reversal invariance. *Phys. Rev. B* **79**, 195321 (2009).
- Bernevig, B. A., Hughes, T. L. & Zhang, S.-C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).
- König, M. *et al.* Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318**, 766–770 (2007).
- Zhang, H. *et al.* Topological insulators in Bi_2Te_3 , Bi_2Se_3 and Sb_2Te_3 with a single Dirac cone on the surface. *Nature Phys.* **5**, 438–442 (2009).
- Hsieh, D. *et al.* Observation of unconventional quantum spin textures in topological insulators. *Science* **323**, 919–922 (2009).
- Hsieh, D. *et al.* A topological Dirac insulator in a quantum spin Hall phase. *Nature* **452**, 970–974 (2008).
- Chen, Y. L. *et al.* Experimental realization of a three-dimensional topological insulator, Bi_2Te_3 . *Science* **325**, 178–181 (2009).
- Li, Y.-Y. *et al.* Growth dynamics and thickness-dependent electronic structure of topological insulator Bi_2Te_3 thin films on Si. Preprint at (<http://arxiv.org/abs/0912.5054v1>) (2009).
- Guo, H.-M. & Franz, M. Theory of quasiparticle interference on the surface of a strong topological insulator. *Phys. Rev. B* **81**, 041102(R) (2010).
- Biswas, R. R. & Balatsky, A. V. Scattering from surface step edges in strong topological insulators. Preprint at (<http://arxiv.org/abs/0912.4477v2>) (2010).
- Zhou, X., Fang, C., Tasi, W.-F. & Hu, J. Theory of quasiparticle scattering in a two-dimensional system of helical Dirac fermions: surface band structure of a three-dimensional topological insulator. *Phys. Rev. B* **80**, 245317 (2009).
- Roushan, P. *et al.* Topological surface states protected from backscattering by chiral spin texture. *Nature* **460**, 1106–1109 (2009).
- Alpichshev, Z. *et al.* STM imaging of electronic waves on the surface of Bi_2Te_3 : topologically protected surface states and hexagonal warping effects. *Phys. Rev. Lett.* **104**, 016401 (2010).
- Zhang, T. *et al.* Experimental demonstration of the topological surface states protected by the time-reversal symmetry. *Phys. Rev. Lett.* **103**, 266803 (2009).
- Gomes, K. K. *et al.* Quantum imaging of topologically unpaired spin-polarized Dirac fermions. Preprint at (<http://arxiv.org/abs/0909.0921v2>) (2009).
- Bürgi, L., Jeandupeux, O., Brune, H. & Kern, K. Probing hot-electron dynamics at surfaces with a cold scanning tunneling microscope. *Phys. Rev. Lett.* **82**, 4516–4519 (1999).
- Bürgi, L., Jeandupeux, O., Hirstein, A., Brune, H. & Kern, K. Confinement of surface state electrons in Fabry–Pérot resonators. *Phys. Rev. Lett.* **81**, 5370–5373 (1998).
- Crommie, M. F., Lutz, C. P. & Eigler, D. M. Imaging standing waves in a two-dimensional electron gas. *Nature* **363**, 524–527 (1993).

22. Avouris, P. & Lyo, I.-W. Observation of quantum size effects at room temperature on metal surfaces with STM. *Science* **264**, 942–945 (1993).
23. Mugarza, A. *et al.* Lateral quantum wells at vicinal Au(111) studied with angle-resolved photoemission. *Phys. Rev. B* **66**, 245419 (2002).
24. Pivetta, M., Silly, F., Patthey, F., Pelz, J. P. & Schneider, W.-D. Reading the ripples of confined surface-state electrons: profiles of constant integrated local density of states. *Phys. Rev. B* **67**, 193402 (2003).
25. Li, J., Schneider, W.-D., Berndt, R. & Crampin, S. Electron confinement to nanoscale Ag islands on Ag(111): a quantitative study. *Phys. Rev. Lett.* **80**, 3332–3335 (1998).
26. Heller, E. J., Crommie, M. F., Lutz, C. P. & Eigler, D. M. Scattering and absorption of surface electron waves in quantum corrals. *Nature* **369**, 464–466 (1994).
27. Fiete, G. A. & Heller, E. J. Theory of quantum corrals and quantum mirages. *Rev. Mod. Phys.* **75**, 933–948 (2003).

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

Acknowledgements We gratefully acknowledge discussions with B. A. Bernevig, B. Boyanov, M. Z. Hasan and N. P. Ong. This work was supported by grants from the NSF-MRSEC programme through the Princeton Center for Complex Materials, the ARO, the DOE, the NSF-DMR and the W. M. Keck Foundation. P.R. acknowledges support by a NSF graduate fellowship.

Author Contributions Y.S.H. and R.J.C. carried out the growth of the single crystals and characterized them; STM measurements and data analysis were done by J.S., P.R., H.B. and A.Y. All authors discussed the results and contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.Y. (yazdani@princeton.edu).

Intra-unit-cell electronic nematicity of the high- T_c copper-oxide pseudogap states

M. J. Lawler^{1,2*}, K. Fujita^{2,3,4*}, Jinhwan Lee^{2,3,5}, A. R. Schmidt^{2,3}, Y. Kohsaka⁶, Chung Koo Kim^{2,3}, H. Eisaki⁷, S. Uchida⁴, J. C. Davis^{2,3,8}, J. P. Sethna² & Eun-Ah Kim²

In the high-transition-temperature (high- T_c) superconductors the pseudogap phase becomes predominant when the density of doped holes is reduced¹. Within this phase it has been unclear which electronic symmetries (if any) are broken, what the identity of any associated order parameter might be, and which microscopic electronic degrees of freedom are active. Here we report the determination of a quantitative order parameter representing intra-unit-cell nematicity: the breaking of rotational symmetry by the electronic structure within each CuO_2 unit cell. We analyse spectroscopic-imaging scanning tunnelling microscope images of the intra-unit-cell states in underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ and, using two independent evaluation techniques, find evidence for electronic nematicity of the states close to the pseudogap energy. Moreover, we demonstrate directly that these phenomena arise from electronic differences at the two oxygen sites within each unit cell. If the characteristics of the pseudogap seen here and by other techniques all have the same microscopic origin, this phase involves weak magnetic states at the O sites that break 90° -rotational symmetry within every CuO_2 unit cell.

The pseudogap phase emerges in the CuO_2 plane (Fig. 1a) upon removal of electrons from the O atoms of the parent ‘charge-transfer’² Mott insulator (Fig. 1b). At low hole-density p , two fundamentally distinct types of electronic excitations are observed³ (Fig. 1c). Here the lower-energy branch of excitations Δ_0 appears to be associated with the maximum energy of Cooper pairing⁴. The higher-energy branch labelled Δ_1 is the spectroscopic ‘pseudogap’ because the same energy gap exists in both the superconducting and pseudogap phases (Fig. 1b). Moreover, Δ_1 tracks T^* , the temperature at which pseudogap phenomenology appears, as a function of doping, implying that Δ_1 is the energy gap associated with any electronic symmetry breaking at T^* . Therefore the question of which symmetries are broken at T^* is equivalent to that of which broken symmetries the excitations with energy Δ_1 exhibit.

The ‘pseudogap states’ at energy $\omega \approx \pm \Delta_1$ can be examined directly using tunnelling spectroscopy. Figure 1d shows the evolution of spatially averaged conductance spectra⁴ with the doping dependence of $\pm \Delta_1$ indicated by blue dashed curves. Figure 2a and b shows atomically resolved electronic structure images with $\omega \approx \Delta_1$ (refs 5 and 6) acquired at $T < T_c$ (Fig. 2a) and $T > T_c$ (Fig. 2b) for an underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ sample with $T_c \approx 35$ K. The insets show that their Fourier transforms are dominated by four inequivalent non-dispersive wavevectors labelled $\mathbf{Q}_x$, $\mathbf{Q}_y$, $\mathbf{S}_x$ and $\mathbf{S}_y$. These phenomena are indistinguishable above and below T_c (refs 5 and 6 and Fig. 2), so the spatial symmetries of the pseudogap states can be established to the highest precision by using the $T \approx 0$ experiments. A striking feature of

such pseudogap state images in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (refs 5–8) and $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ (ref. 5) is the apparent local breaking of 90° rotation (C_4) symmetry at the unit-cell scale. Lattice-translation symmetry is also broken locally at the nanoscale, yielding broad peaks around $\mathbf{S}_x$ and $\mathbf{S}_y$ (refs 5 and 6). Theoretical analyses of such images have posited disordered smectic electronic liquid crystals or stripes^{9,10}.

Indeed, from the perspective of symmetry, the spatial arrangements of the pseudogap states^{5–8} exhibit similarities to those of liquid crystals. Electronic liquid crystals have been defined^{11–17} to arise when the electronic structure breaks spatial symmetries of the crystal lattice (any resultant lattice symmetry changes may or may not be detectable¹⁶). Nematic phases preserve the (lattice) translational symmetry and break the rotational (point group) symmetry, whereas smectic and striped phases break both^{10–17}. Correlated electronic systems exhibiting evidence for electronic nematicity include $\text{Sr}_3\text{Ru}_2\text{O}_7$ (ref. 18) and $\text{CaFe}_{1.94}\text{Co}_{0.06}\text{As}_2$ (ref. 19). Microscopic models for both systems posit nematicity due to a C_4 symmetry breaking which renders the d_{xz} and d_{yz} orbitals of each Ru or Fe atom inequivalent^{20–23}. For copper oxides, the local 180° rotational (C_2) symmetry observed in the pseudogap states^{5,6,8} implied the possibility of electronic nematicity but with contributions from multiple sites within the CuO_2 cell (shown schematically in Fig. 2c).

Here we focus on C_4 to C_2 symmetry reduction in sub-unit-cell resolution images of the copper-oxide pseudogap states^{5,6,8}. We introduce a general measure of electronic nematicity in any image $\tilde{M}(\mathbf{r})$:

$$O_n[M] = \frac{1}{2} [\tilde{M}(\mathbf{Q}_y) - \tilde{M}(\mathbf{Q}_x) + \tilde{M}(-\mathbf{Q}_y) - \tilde{M}(-\mathbf{Q}_x)] \quad (1)$$

$$= \text{Re} \tilde{M}(\mathbf{Q}_y) - \text{Re} \tilde{M}(\mathbf{Q}_x)$$

Here $\tilde{M}(\mathbf{q}) = \text{Re} \tilde{M}(\mathbf{q}) + i \text{Im} \tilde{M}(\mathbf{q})$, the complex-valued two-dimensional Fourier transform of any $M(\mathbf{r})$ with its origin at a Cu site, is evaluated at inequivalent Bragg peaks $\mathbf{Q}_x = (2\pi/a_0)\mathbf{x}$ and $\mathbf{Q}_y = (2\pi/a_0)\mathbf{y}$, and a_0 is the unit-cell dimension. Figure 1a illustrates why only $\text{Re} \tilde{M}(\mathbf{Q})$ enters the definition of $O_n[M]$ by showing it in phase with the Cu lattice (even about each Cu site). By focusing on the Bragg peaks, $O_n[M]$ locks onto only those electronic phenomena with the same spatial periodicity as the crystal and then quantifies the C_4 symmetry breaking. Importantly, this generic order parameter in equation (1) will be widely applicable to data from other techniques. Successful application of equation (1) for spectroscopic-imaging scanning tunnelling microscopy (STS) images requires (1) a highly accurate registry of each Cu site location in $M(\mathbf{r})$ to satisfy the extreme sensitivity of $O_n[M]$ to the phase of $\tilde{M}(\mathbf{Q}_{x,y})$, and (2) that $M(\mathbf{r})$ has the sub-unit-cell

¹Department of Physics, Applied Physics and Astronomy, Binghamton University, Binghamton, New York 13902-6000, USA. ²Laboratory for Atomic and Solid State Physics, Department of Physics, Cornell University, Ithaca, New York 14853, USA. ³Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, New York 11973, USA. ⁴Department of Physics, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan. ⁵Department of Physics, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea. ⁶Magnetic Materials Laboratory, RIKEN, Wako, Saitama 351-0198, Japan. ⁷Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki 305-8568, Japan. ⁸School of Physics and Astronomy, University of St. Andrews, North Haugh, St Andrews, Fife KY16 9SS, Scotland.

*These authors contributed equally to this work.

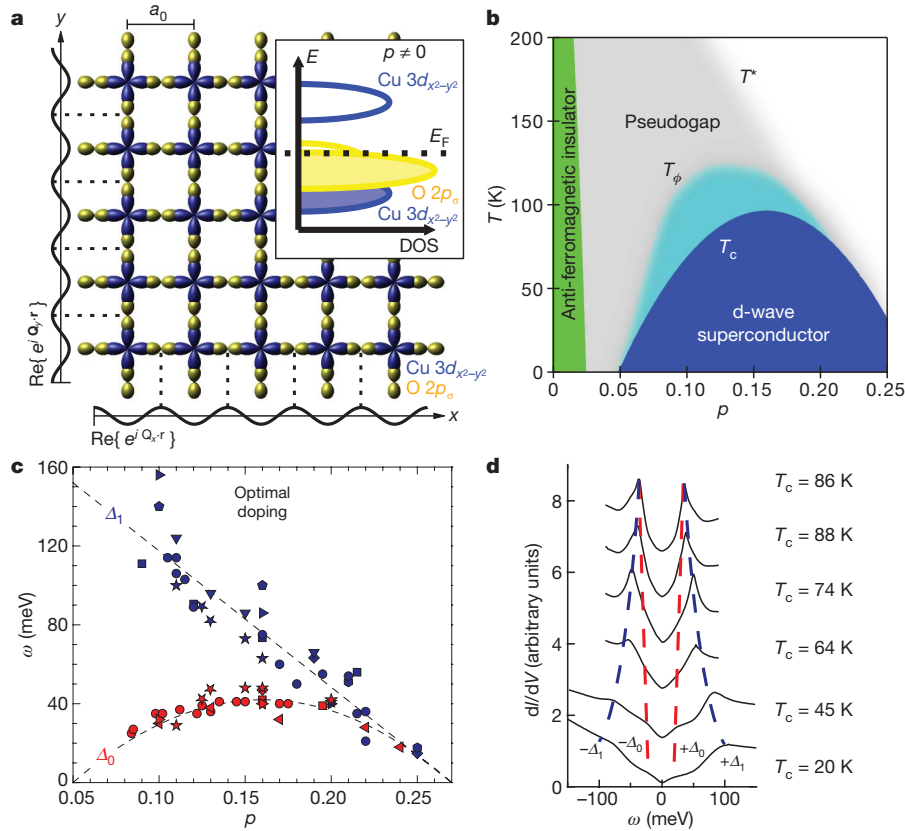


Figure 1 | CuO₂ electronic structure and $\omega \approx \Delta_1$ pseudogap states.

a, Schematic of the spatial arrangements of CuO₂ electronic structure with Cu sites and $d_{x^2-y^2}$ orbitals indicated in blue and O sites and $2p_\sigma$ orbitals in yellow. E_F , Fermi energy. The inset shows the approximate energetics of the band structure when such a charge-transfer insulator is doped by removing electrons from the O atoms (the respective bands are indicated by the same colours as in the CuO₂ schematic). The ‘real’ part of any $\mathbf{q}$ -space electronic structure at the Bragg wavevector $\text{Re}\tilde{M}(\mathbf{Q}, \omega)$ is defined throughout this paper as being in phase with the Cu lattice (and therefore even about each Cu site). This definition is shown schematically as modulations in $\tilde{M}(\mathbf{r}, \omega)$ which would contribute to $\text{Re}\tilde{M}(\mathbf{Q}, \omega)$ at the two Bragg wavevectors $\mathbf{Q}_x$ and $\mathbf{Q}_y$. DOS, density of states. **b**, Schematic copper-oxide phase diagram. Here T_c is the critical temperature circumscribing a ‘dome’ of superconductivity, T_ϕ is the maximum temperature at which phase fluctuations are detectable

resolution without which $O_n[M]$ will be identically zero. $O_n[M]$ would also be identically zero if only the Cu sites $\mathbf{R} = (ma_0, na_0)$ (with m, n integers) contribute to $M(\mathbf{r})$ because then $\tilde{M}(\mathbf{Q}_x) \equiv \tilde{M}(\mathbf{Q}_y)$. In contrast, if all three atomic sites within each CuO₂ unit cell contribute to $M(\mathbf{r})$, then:

$$\begin{aligned}\tilde{M}(\mathbf{Q}_x) &= \bar{M}_{\text{Cu}} - \bar{M}_{\text{O}_x} + \bar{M}_{\text{O}_y} \\ \tilde{M}(\mathbf{Q}_y) &= \bar{M}_{\text{Cu}} + \bar{M}_{\text{O}_x} - \bar{M}_{\text{O}_y} \\ O_n[M] &= 2(\bar{M}_{\text{O}_x} - \bar{M}_{\text{O}_y})\end{aligned}\quad (2)$$

where $\bar{M}_{\text{Cu}}$, $\bar{M}_{\text{O}_x}$ and $\bar{M}_{\text{O}_y}$ are the average of $M(\mathbf{r})$ at the Cu, O_x and O_y sites respectively. The last line follows from equation (1). Hence detection of $O_n[M] \neq 0$ would imply that an inequivalence exists between the electronic structure at the O_x and O_y sites, as shown schematically in Fig. 2c.

Standard differential conductance mapping of $dI/dV(\mathbf{r}, \omega = eV) \equiv g(\mathbf{r}, \omega)$ to image the local density of states $N(\mathbf{r}, \omega)$ is challenging in strongly underdoped copper-oxides^{5–8} because of the intense electronic heterogeneity. However, imaging the ratio $Z(\mathbf{r}, \omega) \equiv \frac{g(\mathbf{r}, +\omega)}{g(\mathbf{r}, -\omega)} \equiv \frac{N(\mathbf{r}, +\omega)}{N(\mathbf{r}, -\omega)}$ avoids these severe systematic errors^{5–8} (Supplementary Information section I). Another challenge is the

within the pseudogap phase, and T^* is the approximate temperature at which the pseudogap phenomenology first appears. **c**, The two distinct classes of excitations as identified by multiple spectroscopies in underdoped copper-oxides (reproduced with permission from ref. 3) as a function of hole-density p . The excitations to energies $\Delta_1(p)$ are referred to as the ‘pseudogap states’ both because $\Delta_1(p)$ tracks $T^*(p)$ and because they exist unchanged in both the pseudogap and superconducting phases. Those excitations to energies $\omega < \Delta_0(p)$ can be associated with the Bogoliubov quasiparticles^{6–8}. **d**, Evolution of the spatially averaged tunnelling spectra of Bi₂Sr₂CaCu₂O_{8+δ} with diminishing p , here characterized by $T_c(p)$. The energies $\Delta_1(p)$ (blue dashed line) are easily detected as the pseudogap edge while the energies $\Delta_0(p)$ (red dashed line) are more subtle but can be identified by the correspondence of the ‘kink’ energy⁴ with the extinction energy of Bogoliubov quasiparticles, following the procedure in ref. 8.

random nanoscale variations of $\Delta_1(\mathbf{r})$; this problem can be mitigated⁸ by defining a reduced energy scale $e \equiv \omega/\Delta_1(\mathbf{r})$ which measures the energy at each position $\mathbf{r}$ with respect to the pseudogap magnitude $\Delta_1(\mathbf{r})$ at that position, thus ensuring that the pseudogap states occur at $e \approx 1$ in all $Z(\mathbf{r}, e)$ data⁸. Using these techniques, a clearer picture of the two basic classes³ of electronic excitations in copper oxides has emerged^{5–8}. First, for all energies below $\omega \approx \Delta_0$ (a slowly doping-dependent energy⁸ indicated by the red dashed line in Fig. 1d) the electronic excitations are demonstrably^{5–8} Bogoliubov quasiparticles. Indeed, in both the superconducting^{5–8} and underdoped pseudogap⁶ phases, these dispersive low-energy states exhibit only the symmetries of a d -wave superconductor. The $\omega \approx \Delta_1$ pseudogap states are profoundly different, being non-dispersive^{5–8} and locally breaking crystal point-group symmetries. The challenge is therefore to identify first which symmetries are broken by the $\omega \approx \Delta_1$ pseudogap states and then the microscopic degrees of freedom involved in any such symmetry breaking.

We explore the intra-unit-cell electronic nematicity of Bi₂Sr₂CaCu₂O_{8+δ} by applying equation (1) to $\tilde{Z}(\mathbf{Q}_x, e)$ and $\tilde{Z}(\mathbf{Q}_y, e)$, the complex-valued two-dimensional Fourier transforms of $Z(\mathbf{r}, e)$ evaluated at $\mathbf{Q}_x$ and $\mathbf{Q}_y$ (red circles in Fig. 2). The sub-unit-cell resolution $Z(\mathbf{r}, e)$ imaging was performed on multiple different underdoped Bi₂Sr₂CaCu₂O_{8+δ} samples with values of T_c between

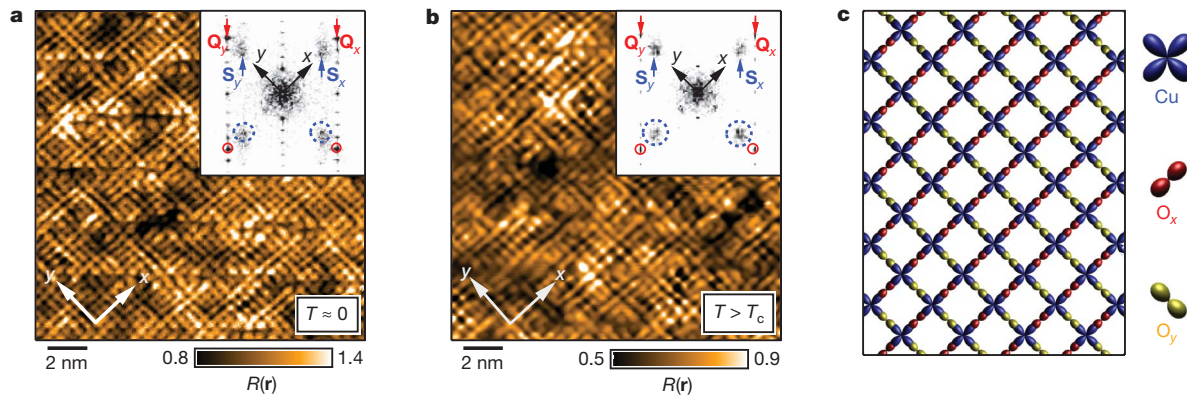


Figure 2 | Imaging the spatial symmetries of the $\omega \approx \Delta_1$ pseudogap states. **a**, Spatial image (R -map⁵) of the $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ pseudogap states $\omega \approx \Delta_1$ at $T \approx 4.3$ K for an underdoped sample with $T_c = 35$ K. The inset shows the Fourier transform upon which the inequivalent Bragg vectors $\mathbf{Q}_x = (1, 0)2\pi/a_0$ and $\mathbf{Q}_y = (0, 1)2\pi/a_0$ are identified by red arrows and circles. The inequivalent wavevectors $\mathbf{S}_x = (\sim 3/4, 0)2\pi/a_0$ and $\mathbf{S}_y = (0, \sim 3/4)2\pi/a_0$ are identified by blue arrows and circles. **b**, Spatial image (R -map⁵) of the $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ pseudogap states $\omega \approx \Delta_1$ at $T \approx 55$ K for the same sample with $T_c = 35$ K. Again, the inset shows the Fourier transform with the inequivalent Bragg

vectors $\mathbf{Q}_x = (1, 0)2\pi/a_0$ and $\mathbf{Q}_y = (0, 1)2\pi/a_0$ identified by red arrows and $\mathbf{S}_x = (\sim 3/4, 0)2\pi/a_0$ and $\mathbf{S}_y = (0, \sim 3/4)2\pi/a_0$ identified by blue arrows and circles. The phenomenology of the $\omega \approx \Delta_1$ pseudogap states, especially their broken spatial symmetries, appear indistinguishable whether in the superconducting phase (**a**) or in the pseudogap phase (**b**). **c**, A schematic representation of how electronic contributions from multiple sites within the CuO_2 unit cell could lead to global electronic nematicity in the copper oxides. Here the two O sites are labelled using different colours to represent the inequivalent electronic structure at those locations within each unit cell.

20 K and 55 K. The necessary registry of the Cu sites in each $Z(\mathbf{r}, e)$ is achieved by a picometre-scale realignment procedure that identifies a transformation to render the topographic image $T(\mathbf{r})$ perfectly a_0 -periodic (Supplementary Information section II). We then apply the same transformation to the simultaneously acquired $Z(\mathbf{r}, e)$ to register all the electronic structure data to this ideal lattice.

Such a topographic image $T(\mathbf{r})$ is shown in Fig. 3a; the location of every Cu site is known with a precision of about 10 picometres. The inset compares the Bragg peaks of its real (in phase) Fourier components $\text{Re}T(\mathbf{Q}_x)$ and $\text{Re}T(\mathbf{Q}_y)$ and demonstrates that $\text{Re}T(\mathbf{Q}_x)/\text{Re}T(\mathbf{Q}_y) = 1$. Therefore $T(\mathbf{r})$ preserves the C_4 symmetry of the crystal lattice. In contrast, Fig. 3b shows that $Z(\mathbf{r}, e = 1)$ determined simultaneously with Fig. 3a seems to break various crystal symmetries locally^{5–7}. The inset showing that $\text{Re}\tilde{Z}(\mathbf{Q}_x, e = 1)/\text{Re}\tilde{Z}(\mathbf{Q}_y, e = 1) \neq 1$ reveals a remarkable discovery: the pseudogap states break C_4 symmetry throughout Fig. 3b. Therefore, following equation (1), we define a normalized order parameter over the entire field of view (FOV) for intra-unit-cell electronic nematicity as a function of e :

$$O_n^Q(e) \equiv \frac{\text{Re}\tilde{Z}(\mathbf{Q}_y, e) - \text{Re}\tilde{Z}(\mathbf{Q}_x, e)}{\tilde{Z}(e)} \quad (3)$$

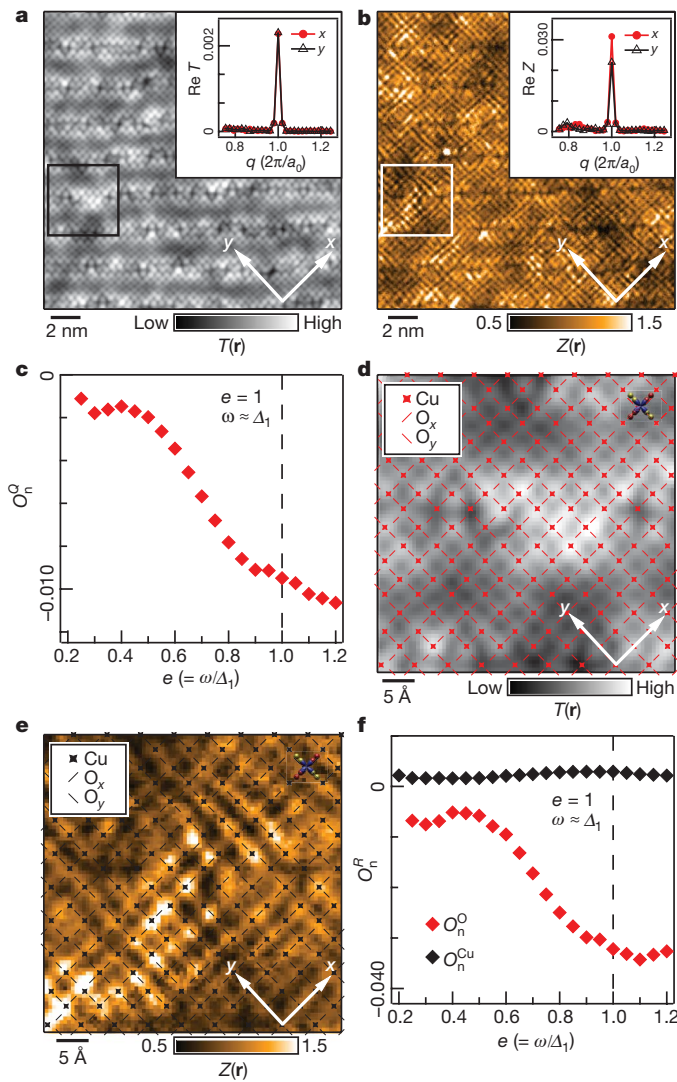
The plot of $O_n^Q(e)$ in Fig. 3c provides the second important discovery: the magnitude of $O_n^Q(e)$ is low for $e \ll 1$, only begins to grow near $e \approx \Delta_0/\Delta_1$, and becomes well defined near $e \approx 1$ ($\omega \approx \Delta_1 \approx 100$ meV in this sample). Hence, the electronic nematicity is associated with the pseudogap states. The low value of $|O_n^Q(e)|$ for small e is to be expected, because here the Bogoliubov quasiparticles are observed to respect the d -wave superconductor symmetries^{4,5,8}. The rapid increase in $|O_n^Q(e)|$ is primarily of electronic origin, because the normalization factor (the spatial average of $Z(\mathbf{r}, e)$) $\tilde{Z}(e) \approx 1$ for all e . Finally, equivalent $O_n^Q(e)$ analyses of all samples studied, measured using two different STMs, show results in agreement with Fig. 3 (Supplementary Information section III), with the exception that the opposite sign of $O_n^Q(e)$ can also be observed.

Equation (2) predicts that the predominant contributions to nematicity must come from the O sites (Fig. 2c). To test this idea directly, we examine $Z(\mathbf{r}, e)$ with sub-unit-cell resolution. Figure 3d shows the topographic image of a representative region from Fig. 3a; the locations of each Cu site $\mathbf{R}$ and of the two O sites within its unit cell are indicated. Figure 3e shows $Z(\mathbf{r}, e)$ measured simultaneously with Fig. 3d with the same Cu and O site labels. To quantify the $\mathbf{r}$ -space electronic nematicity we define the $\mathbf{r}$ -space equivalent of equation (3):

$$O_n^R(e) = \sum_{\mathbf{R}} \frac{Z_x(\mathbf{R}, e) - Z_y(\mathbf{R}, e)}{\tilde{Z}(e)N} \quad (4)$$

Here $Z_x(\mathbf{R}, e)$ is the magnitude of $Z(\mathbf{r}, e)$ at the O site $a_0/2$ along the x -axis from $\mathbf{R}$ while $Z_y(\mathbf{R}, e)$ is the equivalent along the y -axis, and N is the number of unit cells. Hence $O_n^R(e)$ counts only the O site contributions. Figure 3e shows the calculated value of $O_n^R(e)$ from the same FOV as Fig. 3a and b. It is obviously in good agreement with $O_n^Q(e)$, remaining at low amplitude for small e and then rapidly becoming established near the pseudogap states at $e \approx 1$. Thus, the electronic nematicity discovered using $O_n^Q(e)$ indeed derives primarily from the inequivalence of electronic structure at the two O sites within each unit cell. The energy dependence of the average unit-cell electronic structure (Supplementary Information section IV) shows directly that although electronic changes are undetectable at the Cu sites, they are observable at the O sites; this is an explicit demonstration that electronic nematicity in $O_n^Q(e)$ is associated with the O sites and the pseudogap energy. All these observations have been confirmed, using both independent measures $O_n^Q(e)$ and $O_n^R(e)$, in the other samples (Supplementary Information section III). Moreover, other possible sources of systematic error including tip-shape, STM head geometry, scan direction, and sub-atomic register of $Z(\mathbf{r}, e)$ to $T(\mathbf{r})$ have been ruled out positively for every tip/sample combination reported herein (Supplementary Information section V).

Evidence that a predominant symmetry of the pseudogap states of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ is that of an intra-unit-cell nematic motivates examination of any smectic characteristics in $Z(\mathbf{r}, e)$, because a smectic can melt into a nematic^{9–16}. Originally the searches for smectic patterns focused on $g(\mathbf{r}, \omega)$ modulations with $\mathbf{Q}_{1/4,x} \approx (1/4, 0)2\pi/a_0$, $\mathbf{Q}_{1/4,y} \approx (0, 1/4)2\pi/a_0$ (refs 9, 10 and 12). However, all $\mathbf{Q}_{1/4,x}$ and $\mathbf{Q}_{1/4,y}$ features at low energy are actually dispersive Bogoliubov quasiparticle interferences^{6–8} and non-dispersive modulations with these $\mathbf{Q}$ values become very weak in $Z(\mathbf{r}, e \approx 1)$ (Figs 2 and 3). Instead, the strong non-dispersive modulations at the pseudogap energy^{5–8} exhibit wavevectors $\mathbf{S}_x \equiv (\sim 3/4, 0)2\pi/a_0$, $\mathbf{S}_y \equiv (0, \sim 3/4)2\pi/a_0$ (blue circles in Fig. 2 insets) the magnitudes $|\mathbf{S}_x|$ and $|\mathbf{S}_y|$ of which are weakly doping-dependent⁸. There are two important points here: (1) our definition of intra-unit-cell nematicity using the Bragg peaks in equation (1) is independent of the $\mathbf{S}_x$, $\mathbf{S}_y$ (or any other) smectic modulations and, (2) although $\mathbf{S}_x$ and $\mathbf{S}_y$ might seem related to $\mathbf{Q}_{1/4,x}$ and $\mathbf{Q}_{1/4,y}$ by reciprocal lattice vectors, the two sets of wavevectors are actually inequivalent in a $Z(\mathbf{q}, e)$ derived from a $Z(\mathbf{r}, e)$ resolving the three atomic sites within the unit cell.



To examine smectic spatial organization, we next define a measure analogous to equation (3) of C_4 symmetry breaking in modulations with S_x, S_y as:

$$O_s^Q(e) = \frac{\text{Re}\tilde{Z}(S_y, e) - \text{Re}\tilde{Z}(S_x, e)}{\tilde{Z}(e)} \quad (5)$$

For all samples studied, $|O_s^Q(e)|$ is found to be very low and independent of energy (Fig. 4b). Obviously, $|O_s^Q(e)|$ is low at low e because these states are dispersive Bogoliubov quasiparticles^{6–8} but, more importantly, $|O_s^Q(e)|$ shows no tendency to become well established at the pseudogap energy (Fig. 4b).

To separate the nematic and smectic contributions in $Z(\mathbf{r}, e)$, we examine the spatial fluctuations of each ordering tendency by defining coarse-grained $\mathbf{r}$ -space order parameter fields $O_n^Q(\mathbf{r}, e)$ and $O_s^Q(\mathbf{r}, e)$ using the coarsening length scales $1/\Lambda_n$ and $1/\Lambda_s$ shown as red and blue circles in Fig. 2 (Supplementary Information section VI). The resulting $O_n^Q(\mathbf{r}, e)$ and $O_s^Q(\mathbf{r}, e)$ (movies in the Supplementary Information) show that $O_s^Q(\mathbf{r}, e)$ spatially fluctuates wildly in the entire energy range whereas the spatial fluctuation of $O_n^Q(\mathbf{r}, e)$ rapidly subsides as it approaches $e \approx 1$. This dramatic difference is summarized in plots of the correlation lengths $\xi_n(e)$ and $\xi_s(e)$ extracted from $O_n^Q(\mathbf{r}, e)$ and $O_s^Q(\mathbf{r}, e)$ (Fig. 4b), wherein $\xi_n(e)$ diverges to the FOV size at $e \approx 1$. The representative spatial images of $O_n^Q(\mathbf{r}, e=1)$ and $O_s^Q(\mathbf{r}, e=1)$ in Fig. 4c and d show how distinct are the spatial structures of $O_n^Q(\mathbf{r}, e)$ and $O_s^Q(\mathbf{r}, e)$.

Our results also provide a new perspective which, by using C_2 symmetry as a common theme, may help to unify the understanding of

Figure 3 | Nematic ordering and O-site specificity of

$\omega \approx \Delta_1$ pseudogap states. **a**, Topographic image $T(\mathbf{r})$ of the $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ surface. The inset shows that the real part of its Fourier transform $\text{Re}T(\mathbf{q})$ does not break C_4 symmetry at its Bragg points because plots of $T(\mathbf{q})$ show its values to be indistinguishable at $\mathbf{Q}_x = (1, 0)2\pi/a_0$ and $\mathbf{Q}_y = (0, 1)2\pi/a_0$. Importantly, this means that neither the crystal nor the tip used to image it (and its $Z(\mathbf{r}, \omega)$ simultaneously) exhibits C_2 symmetry (Supplementary Information section V). The bulk incommensurate crystal supermodulation is seen clearly here; as always, it is at 45° to, and therefore is the mirror plane between, the x and y axes. For this symmetry reason it has no influence on the electronic nematicity discussed in this paper. **b**, The $Z(\mathbf{r}, e=1)$ image measured simultaneously with $T(\mathbf{r})$ in **a**. The inset shows that the Fourier transform $\text{Re}Z(\mathbf{q}, e=1)$ does break C_4 symmetry at its Bragg points because $\text{Re}Z(\mathbf{Q}_x, e=1) \neq \text{Re}Z(\mathbf{Q}_y, e=1)$. This means that, on average throughout the FOV of **a** and **b**, the modulations of $Z(\mathbf{r}, \omega \approx \Delta_1)$ that are periodic with the lattice have different intensities along the x axis and along the y axis. This is a priori evidence for electronic nematicity in the pseudogap states $\omega \approx \Delta_1$. **c**, The value of $O_n^Q(e)$ defined in equation (3) computed from $Z(\mathbf{r}, e)$ data measured in the same FOV as **a** and **b**. Its magnitude is low for all $\omega < \Delta_0$ and then rises rapidly to become well established near $e \approx 1$ or $\omega \approx \Delta_1$. Thus the quantitative measure of intra-unit-cell electronic nematicity established in equations (1) and (3) reveals that the pseudogap states in this FOV of a strongly underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ sample are nematic. **d**, Topographic image $T(\mathbf{r})$ from the region identified by a small black box in **a**. It is labelled with the locations of the Cu atom plus both the O atoms within each CuO_2 unit cell (labels shown in the inset). Overlaid is the location and orientation of a Cu and four surrounding O atoms using a representation similar to that of Fig. 2c. **e**, The simultaneous $Z(\mathbf{r}, e=1)$ image in the same FOV as **d** (the region identified by small white box in **b**) showing the same Cu and O site labels within each unit cell (see inset). Thus the physical locations at which the nematic measure $O_n^R(e)$ of equation (4) is evaluated are labelled by the dashes. Overlaid is the location and orientation of a Cu atom and four surrounding O atoms using a representation similar to that of Fig. 2c. **f**, The value of $O_n^R(e)$ in equation (4) computed from $Z(\mathbf{r}, e)$ data measured in the same FOV as **a** and **b**. As in **c**, its magnitude is low for all $\omega < \Delta_0$ and then rises rapidly to become well established at $e \approx 1$ or $\omega \approx \Delta_1$. If the function in equation (4) is evaluated using the Cu sites only, the nematicity is about zero (black diamonds), as it must be. This independent quantitative measure of intra-unit-cell electronic nematicity $O_n^R(e)$ again shows that the pseudogap states are strongly nematic and, moreover, that the nematicity is due primarily to electronic inequivalence of the two O sites within each unit cell.

angle-resolved photoemission (ARPES), neutron scattering and spectroscopic-imaging STM studies of broken electronic symmetries within the pseudogap phase. ARPES reveals spontaneous dichroism of the $\mathbf{k} = (1, 0)\pi/a_0$ and $\mathbf{k} = (0, 1)\pi/a_0$ states²⁴, exhibiting C_2 symmetry because the opposite sign of the effect occurs at $\mathbf{k} = (1, 0)\pi/a_0$ and $\mathbf{k} = (0, 1)\pi/a_0$ (ref. 24). These excitations, which are probably magnetic, appear at T^* in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. The unusual magnetic order detected by polarized neutron diffraction at the Bragg peak^{25,26} consists of magnetic moments of about $0.1\mu_B$ (where μ_B is the Bohr magneton) exhibiting C_2 symmetry. These intra-unit-cell signals appear at T^* in both $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ (ref. 25) and $\text{HgBa}_2\text{CuO}_{4+\delta}$ (ref. 26). Our studies reveal intra-unit-cell, C_2 symmetric excitations at the pseudogap energy and that these effects are associated primarily with electronic inequivalence at the two O sites within the CuO_2 unit cell. Given the many common characteristics observed by these diverse techniques, it is reasonable to consider whether ARPES, neutron diffraction and spectroscopic-imaging STM are detecting the same excitations with the same broken symmetries. If so, the pseudogap excitations of underdoped copper oxides would represent weakly magnetic states at the O sites within each CuO_2 unit cell, the electronic structure of which breaks C_4 symmetry. Then, the electronic symmetry breaking that occurs on entering the pseudogap phase would be due to the electronic nematic state visualized here, for the first time to our knowledge (Figs 3 and 4). Finally, the nematicity found in electronic transport²⁷, thermal transport²⁸ and the spin excitation spectrum²⁹ of $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ could then occur because the Ising domains of $O_n^Q(\mathbf{r}, e)$ become aligned by the strong orthorhombicity of its crystal structure³⁰.

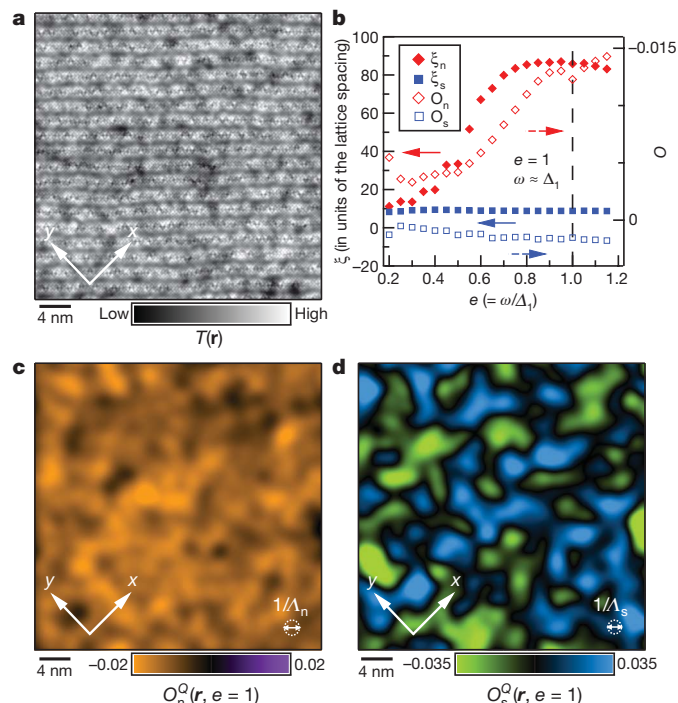


Figure 4 | Rapid increase of correlation length of nematicity at $\omega \approx \Delta_1$. **a**, A large FOV $T(r)$ image which preserves C_4 symmetry. **b**, Correlation lengths $\xi_n(e)$ for the nematic ordering $O_n^Q(r, e)$ (red solid diamonds), and $\xi_s(e)$ for the possibly smectic ordering $O_s^Q(r, e)$ (blue solid squares). (See Supplementary Information section VI for an evaluation of $O_n^Q(r, e)$ and $O_s^Q(r, e)$.) The coarsening length scales $1/\Lambda_n$ and $1/\Lambda_s$ are one-third that of typical spatial variations, as determined from the 3σ radius of the respective peaks. The correlation lengths are determined from the full-width at half-maximum of the spatial auto-correlation functions. Comparison of the nematic order parameter $O_n^Q(e)$ evaluated from equation (3) (open diamonds) and smectic order parameter $O_s^Q(e)$ from equation (6) (open squares) shows that $O_s^Q(e)$ has low magnitude and is energy independent, whereas $O_n^Q(e)$ rises rapidly to become well established at the pseudogap energy. **c**, Image of $O_n^Q(r, e=1)$ (Supplementary Information section VI) from the same FOV showing that $\xi_n(e)$ has diverged to the size of the image. Thus, if there are Ising nematic domains (as there should be), they must be larger than 0.025 micrometres square. We note that the spatial resolution is limited by the cut-off scales shown in the figure. **d**, Images of $O_s^Q(r, e=1)$ (Supplementary Information section VI) from the same FOV as a showing that $\xi_s(e)$ is spatially disordered with very short correlation length; this is equally true at all energies. The evolution of $O_n^Q(r, e)$ and $O_s^Q(r, e)$ as a function of e are available in Supplementary movies 1 and 2, respectively.

Received 5 March; accepted 4 May 2010.

- Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Science* **288**, 468–474 (2000).
- Zaanen, J., Sawatzky, G. A. & Allen, J. W. Band gaps and electronic structure of transition-metal compounds. *Phys. Rev. Lett.* **55**, 418–421 (1985).
- Hüfner, S., Hossain, M. A., Damascelli, A. & Sawatzky, G. A. Two gaps make a high-temperature superconductor? *Rep. Prog. Phys.* **71**, 062501 (2008).
- Allredge, J. W. *et al.* Evolution of the electronic excitation spectrum with strongly diminishing hole density in superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature Phys.* **4**, 319–326 (2008).
- Kohsaka, Y. *et al.* An intrinsic bond-centered electronic glass with unidirectional domains in underdoped cuprates. *Science* **315**, 1380–1385 (2007).
- Lee, J. *et al.* Spectroscopic fingerprint of phase-incoherent superconductivity in the cuprate pseudogap state. *Science* **325**, 1099–1103 (2009).
- McElroy, K. *et al.* Coincidence of checkerboard charge order and antinodal state decoherence in strongly underdoped superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **94**, 197005 (2005).
- Kohsaka, Y. *et al.* How Cooper pairs vanish approaching the Mott insulator in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature* **454**, 1072–1078 (2008).
- Maestro, A. D., Rosenow, B. & Sachdev, S. From stripe to checkerboard ordering of charge-density waves on the square lattice in the presence of quenched disorder. *Phys. Rev. B* **74**, 024520 (2006).

- Robertson, J. A., Kivelson, S. A., Fradkin, E., Fang, A. C. & Kapitulnik, A. Distinguishing patterns of charge order: stripes or checkerboards. *Phys. Rev. B* **74**, 134507 (2006).
- Kivelson, S. A., Fradkin, E. & Emery, V. J. Electronic liquid-crystal phases of a doped Mott insulator. *Nature* **393**, 550–553 (1998).
- Kivelson, S. A. *et al.* How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
- Sachdev, S. Colloquium: order and quantum phase transitions in the cuprate superconductors. *Rev. Mod. Phys.* **75**, 913–932 (2003).
- Vojta, M. Lattice symmetry breaking in cuprate superconductors: stripes, nematics, and superconductivity. *Adv. Phys.* **58**, 699–820 (2009).
- Kim, E.-A. *et al.* Theory of the nodal nematic quantum phase transition in superconductors. *Phys. Rev. B* **77**, 184514 (2008).
- Fradkin, E., Kivelson, S. A., Lawler, M. J., Eisenstein, J. P. & Mackenzie, A. P. Nematic Fermi fluids in condensed matter physics. *Annu. Rev. Condens. Matter Phys.* **1**, 7.1–7.26 (2010).
- Cvetkovic, V., Nussinov, Z., Mukhin, S. & Zaanen, J. Observing the fluctuating stripes in high- T_c superconductors. *Europhys. Lett.* **81**, 27001 (2008).
- Borzi, R. A. *et al.* Formation of a nematic fluid at high fields in $\text{Sr}_3\text{Ru}_2\text{O}_7$. *Science* **315**, 214–217 (2007).
- Chuang, T.-M. *et al.* Nematic electronic structure in the “parent” state of the iron-based superconductor $\text{Ca}(\text{Fe}_{1-x}\text{Co}_x)_2\text{As}_2$. *Science* **327**, 181–184 (2010).
- Raghu, S. *et al.* Microscopic theory of the nematic phase in $\text{Sr}_3\text{Ru}_2\text{O}_7$. *Phys. Rev. B* **79**, 214402 (2009).
- Lv, W. W., J. & Phillips, P. Orbital ordering induces structural phase transition and the resistivity anomaly in iron pnictides. *Phys. Rev. B* **80**, 224506 (2009).
- Lee, W.-C. & Wu, C. Theory of unconventional metamagnetic electron states in orbital band systems. *Phys. Rev. B* **80**, 104438 (2009).
- Krüger, F., Kumar, S., Zaanen, J. & van den Brink, J. Spin-orbital frustrations and anomalous metallic state in iron-pnictide superconductors. *Phys. Rev. B* **79**, 054504 (2009).
- Kaminski, A. *et al.* Spontaneous breaking of time-reversal symmetry in the pseudogap state of a high- T_c superconductor. *Nature* **416**, 610–613 (2002).
- Fauqué, B. *et al.* Magnetic order in the pseudogap phase of high- T_c superconductors. *Phys. Rev. Lett.* **96**, 197001 (2006).
- Li, Y. *et al.* Unusual magnetic order in the pseudogap region of the superconductor $\text{HgBa}_2\text{CuO}_{4+\delta}$. *Nature* **455**, 372–375 (2008).
- Sun, X. F., Segawa, K. & Ando, Y. Metal-to-insulator crossover in $\text{YBa}_2\text{Cu}_3\text{O}_y$ probed by low-temperature quasiparticle heat transport. *Phys. Rev. Lett.* **93**, 107001 (2004).
- Daou, R. *et al.* Broken rotational symmetry in the pseudogap phase of a high- T_c superconductor. *Nature* **463**, 519–522 (2010).
- Hinkov, V. *et al.* Electronic liquid crystal state in the high-temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{6.45}$. *Science* **319**, 597–600 (2008).
- Sun, K., Lawler, M. J. & Kim, E.-A. Spin-charge interplay in electronic liquid crystals: fluctuating spin stripe driven by charge nematic. *Phys. Rev. Lett.* **104**, 106405 (2010).

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

Acknowledgements We are grateful to P. Abbamonte, D. Bonn, J.C. Campuzano, D.M. Eigler, E. Fradkin, T. Hanaguri, W. Hardy, J. E. Hoffman, S. Kivelson, A.P. Mackenzie, M. Norman, B. Ramshaw, S. Sachdev, G. Sawatzky, H. Takagi, J. Tranquada and J. Zaanen, for discussions and communications. Theoretical studies were supported by NSF DMR-0520404 to the Cornell Center for Materials Research. Experimental studies are supported by the Center for Emergent Superconductivity, an Energy Frontier Research Center, headquartered at Brookhaven National Laboratory and funded by the US Department of Energy, under DE-2009-BNL-PM015, as well as by a Grant-in-Aid for Scientific Research from the Ministry of Science and Education (Japan) and the Global Centers of Excellence Program for Japan Society for the Promotion of Science. A.R.S. acknowledges support from the US Army Research Office. M.J.L., J.C.D. and E.-A.K. thank KITP for its hospitality. J.C.D. acknowledges gratefully the hospitality and support of the Physics and Astronomy Department at the University of British Columbia, Vancouver, Canada.

Author Contributions M.J.L. and E.-A.K. led the theoretical studies and initially observed the possibility of nematic order in electronic structure images; M.J.L., K.F. and J.P.S. developed the data analysis algorithms and carried out the analysis; K.F., J.L., C.K.K., A.R.S. and Y.K. carried out the experiments and systematic tests; K.F., H.E. and S.U. synthesized the sequence of samples; K.F. was responsible for the figures and SOM; E.-A.K., J.C.D. and M.J.L. supervised the investigation and wrote the paper. The manuscript reflects the contributions of all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.-A.K. (yun-ah.kim@cornell.edu).

LETTERS

Diamonds sampled by plumes from the core–mantle boundary

Trond H. Torsvik^{1,2,3}, Kevin Burke^{3,4}, Bernhard Steinberger^{1,2,5}, Susan J. Webb³ & Lewis D. Ashwal³

Diamonds are formed under high pressure more than 150 kilometres deep in the Earth's mantle and are brought to the surface mainly by volcanic rocks called kimberlites. Several thousand kimberlites have been mapped on various scales^{1–4}, but it is the distribution of kimberlites in the very old cratons (stable areas of the continental lithosphere that are more than 2.5 billion years old and 300 kilometres thick or more⁵) that have generated the most interest, because kimberlites from those areas are the major carriers of economically viable diamond resources. Kimberlites, which are themselves derived from depths of more than 150 kilometres, provide invaluable information on the composition of the deep subcontinental mantle lithosphere, and on melting and metasomatic processes at or near the interface with the underlying flowing mantle. Here we use plate reconstructions and tomographic images to show that the edges of the largest heterogeneities in the deepest mantle, stable for at least 200 million years and possibly for 540 million years, seem to have controlled the eruption of most Phanerozoic kimberlites. We infer that future exploration for kimberlites and their included diamonds should therefore be concentrated in continents with old cratons that once overlay these plume-generation zones at the core–mantle boundary.

Kimberlites are volatile-rich, potassic, ultramafic igneous rocks that vary enormously in chemical and isotopic composition, mineralogy and texture, and show evidence that they are derived from depleted, enriched and/or fertile mantle sources. The minimum depth of kimberlite generation, as inferred from diamond stability and experimental petrology, is about 150 km (refs 6 and 7), but some have suggested far deeper generation depths of 400–600 km (ref. 8) or even more than 660–1,700 km (refs 9 and 10). Here we put these results into wider perspective by demonstrating that most kimberlites generated during the past 540 million years (Myr) are probably related to plumes that had risen from the two plume generation zones¹¹ (PGZs) at the core–mantle boundary (CMB).

Large igneous provinces consist mainly of basaltic rock erupted relatively rapidly (in the course of 1–5 Myr) over great areas ($1\text{--}10 \times 10^6 \text{ km}^2$; ref. 12). Earlier work has shown that most large igneous provinces of the past 300 Myr (when rotated back to their eruption sites) and active deep-plume-sourced hotspots at the Earth's surface (Fig. 1) project radially down to lie on narrow, stable PGZs at the CMB at the edge of the hot and dense large low-shear-wave-velocity provinces (LLSVPs¹³) of the deep mantle^{11,14–19}, thus demonstrating the long-term stability of LLSVPs. The 1% low-velocity contour in the lowermost layer of the SMEAN tomography model²⁰ is a reasonable proxy for the PGZs, because most reconstructed large igneous province eruption sites and steep horizontal gradients in shear-wave anomalies in the SMEAN model fall close to that contour¹⁴. In Fig. 1 we show 12 hotspots that have been found, by seismic tomography¹⁸, to be sourced by deep plumes. There is evidence from other selection criteria that some further hotspots (for

example, Tristan da Cunha, Réunion, Afar and Hawaii) are also sourced from deep plumes; these are not shown on our map, but they are also almost vertically above the PGZs^{14,16}.

To find out whether kimberlites show an association with PGZs similar to that shown by large igneous provinces and hotspots, we used plate reconstructions^{21,22} to rotate kimberlites that are younger than the initial assembly of Pangaea (~320 Myr) to their original eruption sites. We find that 80% of kimberlites (1,112 out of 1,395) of the past 320 Myr were erupted when their eruption sites lay above a half-width of 15° on either side of the 1% slow contour of the SMEAN model in the lowermost mantle beneath Africa (Fig. 1). On average, this dominant part of the kimberlite population plots at a distance of $7 \pm 5^\circ$ from that contour (Supplementary Table 1). The most anomalous kimberlites younger than 320 Myr (17%) are in the Slave Province of Canada (Late Cretaceous/Early Tertiary period kimberlites²), which was close to a tectonically active continental margin at the time of their eruption.

A remarkable pattern is observed when we plot kimberlites on our series of plate reconstructions. At practically all times, eruption sites

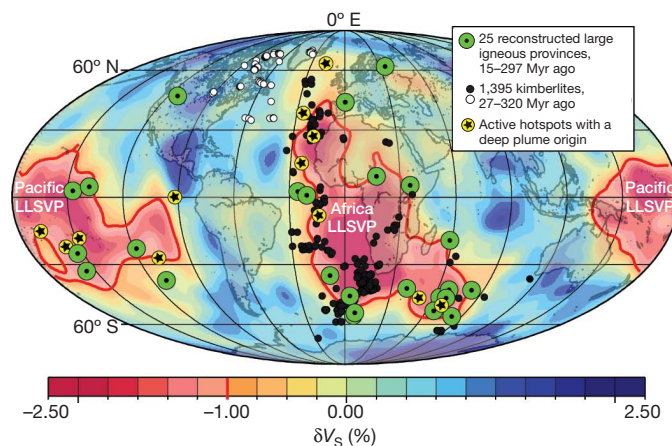


Figure 1 | Reconstructed large igneous provinces and kimberlites for the past 320 Myr with respect to shear-wave anomalies at the base of the mantle. The deep mantle (2,800 km on the SMEAN tomography model²⁰) is dominated by two LLSVPs beneath Africa and the Pacific. The 1% slow contour (approximating to the PGZs) is shown as a thick red line. 80% of all reconstructed kimberlite locations (black dots) of the past 320 Myr erupted near or over the sub-African PGZ. The most 'anomalous' kimberlites (17%) are from Canada (white dots). Present-day continents are shown as a background, to illustrate the distribution of hotspots classified as being of deep-plume origin¹⁸ and present-day shear-wave velocity anomalies (percentage δv_s), and bear no geographical relationship to reconstructed kimberlites or large igneous provinces.

¹Physics of Geological Processes and Geosciences, University of Oslo, Blindern, 0316 Oslo, Norway. ²Centre for Geodynamics, Geological Survey of Norway, Leiv Eirikssons vei 39, 7491 Trondheim, Norway. ³School of Geosciences, University of the Witwatersrand, Wits 2050, South Africa. ⁴Department of Geosciences, University of Houston, 312 Science and Research 1, Houston, Texas 77204-5007, USA. ⁵Helmholtz Centre Potsdam, German Research Centre for Geosciences, 14473 Potsdam, Germany.

plot close to the African PGZ (Fig. 2 and Supplementary Figs 2–5). For the past 320 Myr, Gondwana, with Africa at its heart, has drifted slowly northward over the African PGZ (Supplementary Figs 2–5), and this readily explains the dominance of African (Gondwanan) kimberlites in the global record if, as we suggest, their origin relates to heat from deep plumes. Globally, kimberlite activity peaked between 70 and 120 Myr ago (Supplementary Fig. 1), corresponding to the time of formation of some of the most economically viable diamond sources in southern Africa. This time interval overlaps with the most intense large igneous province activities in the Earth's history, and also with a major superchron⁹ of the magnetic field (~83–120 Myr ago; Cretaceous Normal Superchron). Almost 25% of all known kimberlites erupted between 80 and 90 Myr ago, when Africa was moving very slowly (~1 cm yr⁻¹) northeastward with respect to the mantle (Supplementary Fig. 1).

There are few Phanerozoic kimberlites with ages older than 320 Myr; only about 200 are known to have erupted between 540 and 320 Myr ago, and kimberlites were altogether absent from core Gondwana between 370 and 500 Myr ago. Plate reconstructions provide a possible reason for this: over this time interval, Gondwana was centred on the South Pole, and the bulk of the continent was located between, not over, the two LLSVPs and their marginal PGZs. By the late Devonian period (Fig. 3) Gondwana stretched from the South Pole (by then in South Africa) to the equator, and kimberlites started to erupt along the equatorial and eastern rim of Gondwana (by then in Australia). At this time, kimberlites with economically important diamonds also erupted on the Siberian continent (in Yakutsk; 344–376 Myr ago (refs 2, 23 and 24)). If the stability of LLSVPs and the eruption of large igneous provinces above their margins extend further back than 320 Myr ago, we can constrain Gondwana in longitude at 510 Myr ago and Siberia at 360 Myr ago by placing the Late Cambrian period Kalkarindji large igneous province in Australia and the Yakutsk large igneous province above the LLSVP margins (Fig. 4). In this reconstruction, kimberlites in Siberia between 350 and 360 Myr ago and in Gondwana (the part that became southern Africa) between 500 and 510 Myr ago (Fig. 3) erupted close to vertically above the African PGZ. Going further back, we can show that Cambrian kimberlites with economically important diamonds (535–542 Myr ago (ref. 25)) in Canada fall near the Pacific PGZ, whereas contemporaneous diamond-bearing kimberlites in South Africa erupted above the African PGZ (Fig. 4). Using the principles of plate tectonics and the Yakutsk and Kalkarindji large igneous provinces to

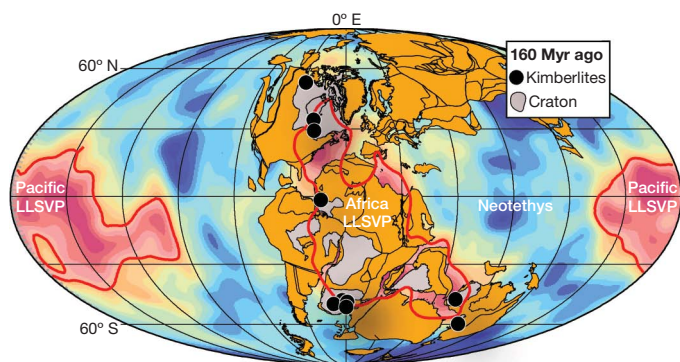


Figure 2 | Late-Jurassic plate reconstruction of continents and kimberlite locations draped on the SMEAN model. We reconstructed kimberlite locations with eruption ages between 155 and 165 Myr ago to the average of 160 Myr ago. Reconstructed kimberlite locations are found near the edges of the African LLSVP (near the 1% slow contour, which is shown as a thick red line) and at the old cratons in North America⁴, northwestern Africa, South Africa (the Kalahari craton¹) and Australia³⁰. The most important cratons for kimberlite eruption since the Carboniferous period are shaded in grey.

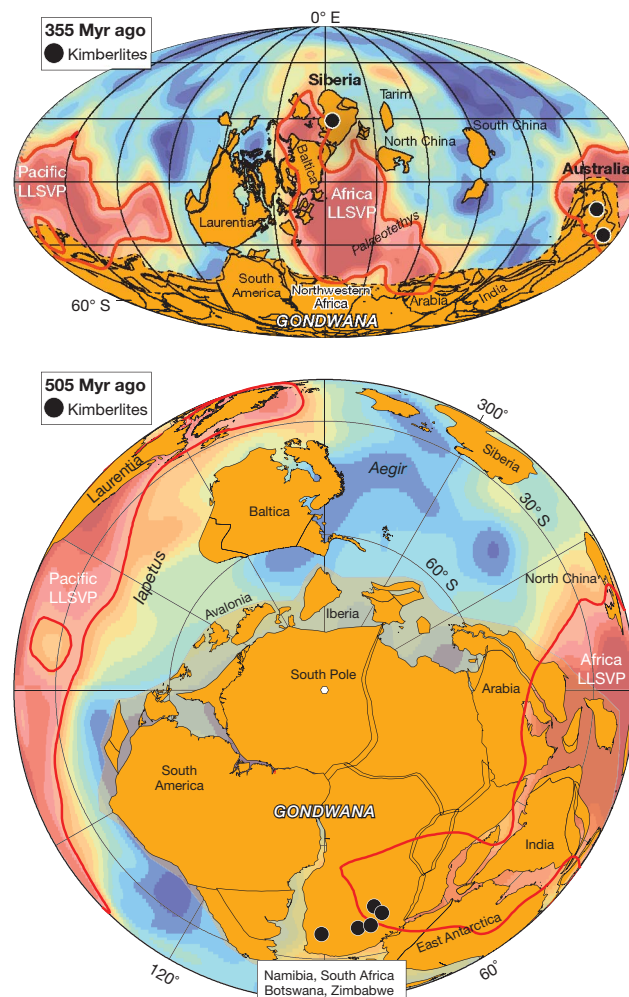


Figure 3 | Devonian and Cambrian period plate reconstructions draped on the SMEAN model. We reconstructed kimberlite locations with eruption ages between 350 and 360 Myr ago to the average of 355 Myr ago, and ages between 500 and 510 Myr ago to the average of 505 Myr ago; they all fall close to vertically above the SMEAN model –1% contours (PGZs), shown as thick red lines.

calibrate our global reconstruction in longitude, we generated semi-absolute reconstructions for the entire Lower and Middle Palaeozoic era, and plotted kimberlite distributions from the major kimberlite-bearing continents (Laurentia, Siberia and Gondwana). These reconstructions show that all kimberlites that erupted between 341 and 542 Myr ago lay, at their times of eruption, above the African (Siberia,

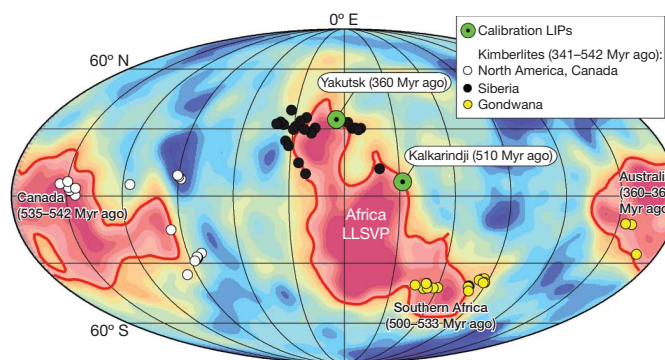


Figure 4 | Reconstructed Palaeozoic kimberlites from Laurentia (North America, Canada), Siberia and core Gondwana draped on the SMEAN model. The SMEAN model –1% contour is shown as a thick red line.

southern Africa) and Pacific (Laurentia, Australia) PGZs (Figs 3 and 4). On average, those kimberlites plot at a distance of $8 \pm 4^\circ$ from the 1% slow contour of the SMEAN model, with 93% lying within a half-width of 10° from that contour (Supplementary Table 1).

Mantle plumes have been argued, using a variety of observations^{9,10,23,26}, to be important in some or all kimberlite eruptions. We have shown that large igneous provinces and hotspot volcanoes develop from the deep-seated mantle plumes that rose from two PGZs^{11,14–16}. Here, for the first time, we show that plumes that have risen from the PGZs at the margins of the sub-African (Figs 1 and 2 and Supplementary Figs 2–6) and the Pacific (Figs 3 and 4) LLSVPs were also involved in kimberlite eruption. This clustering of kimberlites above LLSVP margins is extremely unlikely to be the result of chance, and we estimate that it has a probability of 0.1–1% or less (Supplementary Figs 7 and 8).

Kimberlites are only known within continents, and about 80% of those that have erupted during the past 320 Myr formed within a part of a continent that at the time of the eruption lay close to vertically above a PGZ at the CMB (Fig. 1). The high concentration of economically viable kimberlites in Africa is due to the very old (>2.5 billion years) cratonic parts of the continent lying above a PGZ at various times during the past 320 Myr. This indicates that the search for kimberlites and the diamonds that they contain might be profitably concentrated in areas within the old cratons of continents that once overlay a PGZ (Fig. 2 and Supplementary Figs 2–6). Current limitations in absolute plate reconstructions make it harder to identify such places for times earlier than 320 Myr ago (ref. 16). However, if the relationship of large igneous province and kimberlite eruptions to the PGZs holds true for that time (Figs 3 and 4), we can use the information to position continents close to their original longitude long before the assembly of Pangaea and probably throughout the entire Phanerozoic eon.

We can now show that three distinct kinds of igneous bodies, represented by (1) at least 12 active hotspot volcanoes¹⁸, (2) 23 large igneous provinces of the past 300 Myr (refs 11 and 14–16) and (3) 1,112 kimberlites from the past 320 Myr (this paper) now lie, in the case of the active hotspot volcanoes, or lay at the time of their eruption, in the cases of large igneous provinces and kimberlites, vertically or nearly vertically above a PGZ at the CMB (Fig. 1). The PGZs can be described as narrow loci of an intermittent or continuous upward flux of hot and buoyant material from the CMB. Lateral flow above the CMB may be deflected upwards at the margins of LLSVPs, which are probably chemically distinct^{11,13–17}. This flux appears to be related to the emplacement of large igneous provinces, 'hotspot volcanoes' (of which some, but not all, may lie on tracks that originated in large igneous provinces) and kimberlites.

Large igneous provinces and kimberlites have erupted since Archaean times. Our results show that most of these rocks have been derived from deep plumes originating at the margins of LLSVPs, but whether the African and Pacific LLSVPs have remained in the same places throughout the Earth's history is less certain^{16,27,28}. The stability of LLSVPs in their present locations at the CMB can be demonstrated for large igneous provinces and kimberlites for the past 320 Myr. Most large igneous provinces and kimberlites erupted during roughly the past 200 Myr, so we can be confident about LLSVP stability since then. Explaining those stable LLSVPs and the rising of plumes from their edges requires a new and challenging generation of dynamic mantle models²⁹. We can find a reasonable plate reconstruction with continents placed in a longitude such that the two known large igneous provinces and roughly 200 kimberlites that erupted between 540 and 320 Myr ago fall close to vertically above the present LLSVP margins. This indicates that the near-antipodal locations of the two existing LLSVPs on the equator may have been stationary for as much as 540 Myr, and thus seem to be insensitive to surface plate motions, including those of the formation of Pangaea.

METHODS SUMMARY

We combined reconstructions derived from a hotspot frame for the past 100 Myr with a palaeomagnetic frame dating back to the initial assembly of Pangaea (320 Myr ago). This is known as the global hybrid frame²¹, which we correct here for true polar wander²² between 320 and 100 Myr ago. For data earlier than 320 Myr ago we used the PGZ-reconstruction method to calibrate longitudes¹⁶. This method uses the long-term relationship between large igneous provinces and PGZs to estimate longitudes for large igneous provinces, and we use it here to identify the continents under which the PGZs lay at times of kimberlite eruption. We calibrated longitudes more than 320 Myr ago by placing the Yakutsk large igneous province in Siberia (~360 Myr ago) and the Kalkarindji large igneous province in Australia/Gondwana (~510 Myr ago) over the most likely edges of the African LLSVP¹⁶ (Fig. 4). We know the palaeolatitude for the Yakutsk (~35° N) and Kalkarindji (~9° N) large igneous provinces from palaeomagnetic data from Siberia and Gondwana.

We derived the kimberlite locations from numerous sources (including a recent African compilation³), and included 1,395 'dated' kimberlites for the past 320 Myr. Kimberlite age control varies from excellent (for example, U/Pb dating) to less certain (that is, ages assumed to be similar to those of neighbouring kimberlites). We did not include undated kimberlites or those with vaguely described ages in our analysis. We first rotated each kimberlite site to southern African coordinates using relative rotation parameters²¹, and then rotated them to their correct palaeoposition on the globe (Fig. 1) using the absolute reference frames outlined above. We then draped reconstructed kimberlite eruption sites (symbols in Figs 1–4 may represent multiple sites) on the present-day SMEAN-model anomalies near the CMB, assuming that the African and the Pacific LLSVPs have remained stationary for at least 300 Myr.

We produced diagrams with GMT (gmt.soest.hawaii.edu), GMAP (www.geodynamics.no) and GPlates (www.gplates.org) software, and with SPlates software developed for our industry sponsor (Statoil).

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

Received 21 September 2009; accepted 19 May 2010.

- Jelsma, H. A. *et al.* Preferential distribution along transcontinental corridors of kimberlites and related rocks of southern Africa. *S. Afr. J. Geol.* **107**, 301–324 (2004).
- Kjarsgaard, B. A. in *Mineral Deposits of Canada: A Synthesis of Major Deposit-Types, District Metallogeny, the Evolution of Geological Provinces, and Exploration Methods* (ed. Goodfellow, W. D.) 245–272 (Geol. Assoc. Canada Special Publication 5, 2007).
- Jelsma, H., Barnett, W., Richards, S. & Lister, G. Tectonic setting of kimberlites. *Lithos* **112**, 155–165 (2009).
- Heaman, L. M. & Kjarsgaard, B. A. Timing of eastern North American kimberlite magmatism: continental extension of the Great Meteor hotspot track? *Earth Planet. Sci. Lett.* **178**, 253–268 (2000).
- Jordan, T. H. in *The Mantle Sample: Inclusions in Kimberlites and Other Volcanics* (eds Boyd, F. R. & Meyer, H. O. A.) 1–14 (AGU, 1979).
- Mitchell, R. H. *Kimberlites: Mineralogy, Geochemistry and Petrology* (Plenum, 1986).
- Wyllie, P. J. The origin of kimberlites. *J. Geophys. Res.* **85**, 6902–6910 (1980).
- Ringwood, A. E., Kesson, S. E., Hibberson, W. & Ware, N. Origin of kimberlites and related magmas. *Earth Planet. Sci. Lett.* **113**, 521–538 (1992).
- Haggerty, S. E. A diamond trilogy: superplumes, supercontinents, and supernovae. *Science* **285**, 851–861 (1999).
- Hayman, P. C., Kopylova, M. G. & Kaminsky, F. V. Lower mantle diamonds from Rio Soriso (Juina area, Mato Grosso, Brazil). *Contrib. Mineral. Petrol.* **149**, 430–445 (2005).
- Burke, K., Steinberger, B., Torsvik, T. H. & Smethurst, M. A. Plume generation zones at the margins of large low shear velocity provinces on the core–mantle boundary. *Earth Planet. Sci. Lett.* **265**, 49–60 (2008).
- Bryan, S. & Ernst, R. Revised definition of large igneous provinces (LIPs). *Earth Sci. Rev.* **86**, 175–202 (2008).
- Garnero, E. J., Lay, T. & McNamara, A. K. in *Plates, Plumes, and Planetary Processes* (eds Foulger, G. R. & Jurdy, D. M.) 79–109 (Geol. Soc. Am. Special Paper 430, 2007).
- Torsvik, T. H., Smethurst, M. A., Burke, K. & Steinberger, B. Large igneous provinces generated from the margins of the large low-velocity provinces in the deep mantle. *Geophys. J. Int.* **167**, 1447–1460 (2006).
- Torsvik, T. H., Smethurst, M. A., Burke, K. & Steinberger, B. Long term stability in deep mantle structure: evidence from the ~300 Ma Skagerrak-centered large igneous province (the SCLIP). *Earth Planet. Sci. Lett.* **267**, 444–452 (2008).
- Torsvik, T. H., Steinberger, B., Cocks, L. R. M. & Burke, K. Longitude: linking Earth's ancient surface to its deep interior. *Earth Planet. Sci. Lett.* **276**, 273–283 (2008).
- Thorne, M. S., Garnero, E. J. & Grand, S. Geographic correlation between hot spots and deep mantle lateral shear-wave velocity gradients. *Phys. Earth Planet. Inter.* **146**, 47–63 (2004).

18. Montelli, R., Nolet, G., Dahlen, F. & Masters, G. A catalogue of deep mantle plumes: new results from finite-frequency tomography. *Geochem. Geophys. Geosyst.* **7**, Q11007, doi: 10.1029/2006GC001248 (2006).
19. Davaille, A., Stutzmann, E., Silveira, G., Besse, J. & Courtillot, V. Convective patterns under the Indo-Atlantic. *Earth Planet. Sci. Lett.* **239**, 233–252 (2005).
20. Becker, T. W. & Boschi, L. A comparison of tomographic and geodynamic mantle models. *Geochem. Geophys. Geosyst.* **3**, 1003, doi: 10.1029/2001GC000168 (2002).
21. Torsvik, T. H., Müller, R. D., Van der Voo, R., Steinberger, B. & Gaina, C. Global plate motion frames: toward a unified model. *Rev. Geophys.* **46**, RG3004, doi: 10.1029/2007RG000227 (2008).
22. Steinberger, B. & Torsvik, T. H. Absolute plate motions and true polar wander in the absence of hotspot tracks. *Nature* **452**, 620–623 (2008).
23. Yakubchuk, A. Diamond deposits of the Siberian craton: products of post-1200 Ma plume events affecting the lithospheric keel. *Ore Geol. Rev.* **35**, 155–163 (2009).
24. Kinny, P. D., Griffin, B. J., Heaman, L. M., Brakhfogel, F. F. & Spetsius, Z. V. Shrimp U–Pb ages of perovskite from Yakutian kimberlites. *Russ. Geol. Geophys.* **38**, 97–105 (1997).
25. Heaman, L. M., Kjarsgaard, B. A. & Creaser, R. A. The timing of kimberlite magmatism in North America: implications for global kimberlite genesis and diamond exploration. *Lithos* **71**, 153–184 (2004).
26. Le Roex, A. P., Bell, D. R. & Davis, D. Petrogenesis of group I kimberlites from Kimberley, South Africa: evidence from bulk-rock geochemistry. *J. Petrol.* **44**, 2261–2286 (2003).
27. Zhong, S., Zhang, N., Li, Z. X. & Roberts, J. H. Supercontinent cycles, true polar wander, and very long-wavelength mantle convection. *Earth Planet. Sci. Lett.* **261**, 551–564 (2007).
28. Li, Z. X. & Zhong, S. Supercontinent–superplume coupling, true polar wander and plume mobility: plate dominance in whole-mantle tectonics. *Phys. Earth Planet. Inter.* **176**, 143–156 (2009).
29. Tan, E., Leng, W., Zhong, S. & Gurnis, M. On the fixity of the thermo-chemical piles at the base of mantle. *Eos (Fall Meeting)* **90**, abstr. D112A–08 (2009).
30. Jaques, A. L. Kimberlite and lamproite diamond pipes. *AGSO J. Aust. Geol. Geophys.* **17**, 153–162 (1998).

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

Acknowledgements We thank R. Trønnes, S. Haggerty, M. Gurnis and C. Gaina for comments and discussions, and S. King and D. Evans for reviews. We acknowledge Statoil and the Norwegian Research Council for financial support.

Author Contributions T.H.T. and K.B. developed the conceptual idea for the study, B.S. developed statistical methods and tests and S.J.W. and L.D.A. assembled input data. All authors contributed to discussions and writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.H.T. (trond.torsvik@ngu.no).

METHODS

Our methods depend on several factors, including kimberlite age uncertainties and the choice of both plate and tomography models. In addition, plume sources may have been advected in the mantle^{11,32}, and a kimberlite eruption site may not mark precisely where a plume impinged on the base of the lithosphere, but rather the location of material that may have propagated horizontally within the lithosphere from a plume^{11,33,34,14}. The observation that kimberlites in some cases occur in clusters or lines³ may indicate that their surface distribution is partly structurally controlled; it is therefore complex to estimate the net effect of these individual sources of uncertainty.

We have previously examined nine different shear-wave-velocity models; they all provide broadly similar characteristics near the CMB, so the choice of tomographic model is not critical to our conclusions^{11,16} but may lead to slightly different statistical correlations. As an example, we compare the 1% slow contour of the SMEAN model with the 0.96% slow contour in the Castle *et al.* D" model³⁵ and the 0.77% slow contour in the Kuo *et al.* D" model³⁶ (Supplementary Fig. 6), both of which globally, at the CMB, approximately correspond to the 1% slow contour of the SMEAN model¹¹. Twenty-five reconstructed large igneous provinces plot on average at a distance of $8 \pm 9^\circ$ (mean $\pm$ standard deviation) from the SMEAN model contour, whereas the distances from the CASTLE and KUO contours are reduced to $5 \pm 3^\circ$ and $6 \pm 4^\circ$ (Supplementary Table 1). In the SMEAN model, 80% of all reconstructed large igneous provinces plot within a 10° half-width centred on the 1% slow contour, increasing to 96% for the CASTLE contour (Supplementary Table 1 and Supplementary Fig. 7a). The CASTLE model scores the highest because its contour contains two small subareas at the CMB that plot near the Siberian Traps and the Columbia River Basalt (ST and CB in Supplementary Fig. 6). The CASTLE contour is also continuous further north in the North Atlantic, and thus the Iceland hotspot (Fig. 1) also fits this model best. We consider it likely that the Iceland plume is related to a continuation of the Africa LLSVP, and it is possible that the smaller anomaly now underlying the reconstructed Siberia Trap has also been part of the African LLSVP. Different tomography models therefore do matter in a statistical sense, but all three models (and most other models at the CMB^{11,16}) demonstrate that large igneous provinces correlate with the edges of CMB heterogeneities and never with their centres.

Kimberlite distribution is also sensitive to the specific tomography model, but the roughly 17% of 'anomalous' Late-Cretaceous–Early-Tertiary North American kimberlites in the database of the last 320 Myr ($\sim 12\%$ of the entire Phanerozoic collection) are anomalous in all models. The remaining kimberlites plot at an average distance of $7 \pm 4^\circ$ from the SMEAN model contour, $6 \pm 4^\circ$ from the CASTLE contour and $3 \pm 3^\circ$ from the KUO contour (Supplementary Table 1; 27–314 Myr population). Seventy-three per cent plot within 10° of the SMEAN model contour (Supplementary Fig. 7c). That increases to 85% and 94% for the CASTLE and KUO contours. For comparison, *in-situ* (that is, non-reconstructed) kimberlite locations plot at a distance of $19 \pm 12^\circ$ with only 14% inside the 10° band of the SMEAN model—clearly a much worse fit (Supplementary Fig. 7b). Although we appreciate the better fit for the CASTLE and KUO models, these contours are longer, and hence the area within 10° is larger than for the SMEAN model—1% contour. However, the major reason that the KUO model best fits the kimberlite data (Supplementary Table 1) is that a large population of 80–90-Myr-old kimberlites in South Africa (white arrows marked 2 in Supplementary Fig. 6; see also Supplementary Fig. 4) plot right on top of the 0.77% slow KUO contour, whereas they plot at some distance inside the SMEAN model contour.

An absolute plate motion model must account for the distribution of subducted slab material in the mantle through geological time. Such a reference system, based on information about subducted slabs identified from seismic tomography and plate kinematic models, is still in its infancy, but as a plate model sensitivity test we reconstructed kimberlite eruption sites for the past 300 Myr using the subduction reference frame of van der Meer *et al.*³⁷. Excluding the Late Cretaceous/Early Tertiary North American kimberlites, kimberlites plot at a distance of $9 \pm 4^\circ$, with 65% inside the 10° band for the SMEAN model (Supplementary Fig. 7c). This is slightly worse, but within the error of our hybrid plate model.

We have previously given a statistical argument that the coincidence of reconstructed large igneous provinces with the LLSVP margins is extremely unlikely to have resulted from pure chance¹¹, but how likely is it that a kimberlite distribution near the LLSVP edges occurs by chance? Kimberlites only occur in continents, and the diamond-bearing kimberlites occur in cratons more than 2.5 billion years old. Those old cratons make up roughly 15% of the total area of the continents. In Supplementary Fig. 7d, we plot the fraction of cratons that are within 10° of the PGZ as a function of time, on the basis of three tomography models (SMEAN, CASTLE and KUO). This should be equal to the fraction of kimberlites if they were formed randomly on the cratons. For comparison, 62% (dashed red line) of all kimberlites (980 of 1,588) and 33% of the surface of the sphere are within 10° based on the SMEAN model. The percentage of kimberlites

(62%) is slightly less than the maximum percentage of cratons ($\sim 70\%$) that are within 10° of the SMEAN model slow margin, but at the times when most of the kimberlites were formed a much smaller percentage of cratons were within 10° —about the same as, or even less than, the percentage of the entire surface of the sphere: that is, the result that would be expected if the cratons were placed randomly. This shows that the clustering of kimberlites near the 1% slow margin cannot be due to a clustering of cratons near the 1% slow margin. Numbers for the other tomography models, CASTLE and KUO, are slightly higher but lead to the same conclusion. At the time when most kimberlites formed, the cratons' location relative to the 1% slow margin was more or less as would be expected from a random distribution, but the kimberlites were not formed on the cratons in the way that would be expected in a random distribution. The lighter-coloured dashed lines show that the fraction of kimberlites within 10° of the PGZs decreases if we restrict ourselves to more recent times. This may be partly an effect of less freedom in longitude adjustment for more recent times—we have in fact adjusted longitudes to fit large igneous provinces above PGZs before 320 Myr (Fig. 4). Hence, there may be an increasing bias towards an increased fraction of kimberlites above PGZs further back in time. Nevertheless, even for the most recent time interval, up to 130 Myr ago, in which longitudes can be constrained by hotspot tracks, the fraction of kimberlites within 10° of PGZs is much higher than the fraction of cratons, so the clustering of kimberlites near PGZs cannot be due to freedom in longitude. Although in our reconstruction we switch from a hotspot-based to a palaeomagnetic reference frame at 100 Myr ago, comparison of the two frames shows only a minor difference in longitude between 100 and 130 Myr ago.

The probability that kimberlites were emplaced randomly is further explored in Supplementary Fig. 8. The calculated probabilities are quite variable, depending on which tomography model is used (different colours), whether we consider the fraction of individual kimberlites within a half-width of 10° of the PGZs (lines) or the fraction of kimberlite 'groups' (filled circles), which time interval is considered and how many independent groups there are. Results further depend on the half-width and on which contour is used to define the PGZs (not shown).

Because groups with larger numbers of kimberlites should presumably be given more weight, we expect that the most appropriate estimate for probability in each case lies somewhere between the filled circles and the line of same colour in Supplementary Fig. 8. We estimate that there are about 43 to 55 'independent' groups of kimberlites that have formed in the last 542 Myr. As indicated in Supplementary Fig. 7d and Supplementary Table 1, we expect that with more tight, independent constraints on longitude more than 130 Myr ago, and especially more than 320 Myr ago, the fraction of kimberlites within 10° of the PGZs might be slightly, but not substantially, less; hence, probabilities might be slightly higher than those inferred from the range 43–55 in Supplementary Fig. 8. On the other hand, probability estimates from kimberlites younger than 320 Myr only are rather high, because during that time interval a large fraction of cratons was already within 10° of the PGZs. However, most kimberlites erupted at times when the fraction of cratons within 10° of the PGZs was much less, so these estimates are probably too high. Given all this, we expect that the probability of the distribution of kimberlites relative to PGZs being essentially random is about 0.1–1% or less. We emphasize that this estimate considers all kimberlites, including the 'anomalous' ones from Canada and the large cluster in South Africa, which is reconstructed above the African LLSVP away from its margin, if the SMEAN model 1% slow contour is used. We regard it as highly likely that the distribution of kimberlites is indeed related to the PGZs at the margins of LLSVPs in the lowermost mantle.

It has been suggested that kimberlite eruptions in, for example, North America and Africa occurred during periods of relatively slow continental motion³⁸. To test this idea, we calculated the absolute motion of South Africa and North America for the past 320 Myr (Supplementary Fig. 1b). Our velocity curves differ from those of Speeds and Houseman³⁸, but we notice that South Africa has relatively low speeds ($1\text{--}3.5\text{ cm yr}^{-1}$) during peaks in kimberlite eruption (70–100 Myr ago and 110–120 Myr ago). These lows are also seen for North America, but there are two high-velocity spikes. The spike at 50–60 Myr ago is associated with 'anomalous' kimberlites (Fig. 1, white dots), erupted shortly after the collision of the ribbon continent of the Cordillera with North America, which was a time of tectonic activity in the Canadian Rockies when cracks that fostered decompression melting are likely to have formed in the Slave Province³⁹. Only one lower-mantle mineral assemblage has been reported in Cretaceous–Tertiary kimberlites in Canada, but there is abundant majoritic garnet included in diamond^{40,41}. A transition-zone (410–660 km) activated plume by 'large scale extension' seems to be a reasonable explanation for these 'anomalous' kimberlites.

That the reconstructed positions of at least 23 large igneous provinces, and now the majority of kimberlites (Fig. 1), all fall near the 1% slow contour demonstrates that the majority of both large igneous provinces and kimberlites are derived from the PGZs near the CMB. These observations are undoubtedly incompatible with passive plate-driven models for large igneous province

genesis⁴², because in such alternative models there should be no correlation between surface volcanism and deep-mantle heterogeneities, nor is it very likely that upper mantle and crustal processes could affect the polarity pattern of the geodynamo (Supplementary Fig. 1).

31. Steinberger, B., Sutherland, R. & O'Connell, R. J. Prediction of Emperor–Hawaii seamount locations from a revised model of plate motion and mantle flow. *Nature* **430**, 167–173 (2004).
32. Boschi, L., Becker, T. W. & Steinberger, B. Mantle plumes: dynamic models and seismic images. *Geochem. Geophys. Geosyst.* **8**, Q10006, doi: 10.1029/2007GC001733 (2007).
33. Sleep, N. H. Mantle plumes from top to bottom. *Earth Sci. Rev.* **77**, 231–271 (2006).
34. Courtillot, V., Jaupart, C., Manighetti, I., Tapponnier, P. & Besse, J. On causal links between flood basalts and continental breakup. *Earth Planet. Sci. Lett.* **166**, 177–195 (1999).
35. Castle, J. C., Creager, K. C., Winchester, J. P. & van der Hilst, R. D. Shear wave speeds at the base of the mantle. *J. Geophys. Res.* **105**, 21543–21558 (2000).
36. Kuo, B.-Y., Garnero, E. J. & Lay, T. Tomographic inversion of S–SKS times for shear wave velocity heterogeneity in D'': degree 12 and hybrid models. *J. Geophys. Res.* **105**, 139–157 (2000).
37. van der Meer, D. G., Spakman, W., van Hinsbergen, D. J. J., Amaru, M. L. & Torsvik, T. H. Towards absolute plate motions constrained by lower-mantle slab remnants. *Nature Geosci.* **3**, 36–40 (2010).
38. England, P. & Houseman, G. On the geodynamic setting of kimberlite genesis. *Earth Planet. Sci. Lett.* **67**, 109–122 (1984).
39. Johnston, S. The Cordilleran ribbon continent of North America. *Annu. Rev. Earth Planet. Sci.* **36**, 495–530 (2008).
40. Davies, R., Griffin, W. L., O'Reilly, S. Y. & McCandless, T. E. Inclusions in diamonds from the K14 and K10 kimberlites, Buffalo Hills, Alberta, Canada: diamond growth in a plume? *Lithos* **77**, 99–111 (2004).
41. Stachel, T., Harris, J. W. & Muehlenbachs, K. Sources of carbon in inclusion bearing diamonds. *Lithos* **112**, 625–637 (2009).
42. Foulger, G. R. & Jurdy, D. M. (eds) *Plates, Plumes, and Planetary Processes* (Geol. Soc. Am. Special Paper 430, 2007).

LETTERS

Striations, duration, migration and tidal response in deep tremor

Satoshi Ide¹

Deep tremor in subduction zones is thought to be caused by small repeating shear slip events on the plate interface with significant slow components^{1–4}. It occurs at a depth of about 30 kilometres and provides valuable information on deep plate motion and shallow stress accumulation on the fault plane of megathrust earthquakes. Tremor has been suggested to repeat at a regular interval^{1,2}, migrate at various velocities^{4–7} and be modulated by tidal stress^{6,8,9}. Here I show that some time-invariant interface property controls tremor behaviour, using precise location of tremor sources with event duration in western Shikoku in the Nankai subduction zone, Japan. In areas where tremor duration is short, tremor is more strongly affected by tidal stress and migration is inhibited. Where tremor lasts longer, diffusive migration occurs with a constant diffusivity of $10^4 \text{ m}^2 \text{ s}^{-1}$. The control property may be the ratio of brittle to ductile areas, perhaps determined by the influence of mantle wedge serpentinization on the plate interface. The spatial variation of the controlling property seems to be characterized by striations in tremor source distribution, which follows either the current or previous plate subduction directions. This suggests that the striations and corresponding interface properties are formed through the subduction of inhomogeneous structure, such as seamounts, for periods as long as ten million years.

One of the most important discoveries in Earth science in the last decade is that of deep non-volcanic tremor¹ often accompanied with a slow slip event. The combined phenomena have come to be known as episodic tremor and slip². In western Shikoku, in the Nankai subduction zone, Japan, most seismically observed tremor and geodetically observed slow slip events share a common mechanism—shear slip on the plate interface^{10,11}—and a scaling law in which the event duration is proportional to the seismic moment (that is, the spatial integral of slip multiplied by the rigidity)¹². Similar phenomena have been discovered in other subduction zones^{13,14}, and tremor appears to occur as shear slip on the plate interface at least in Nankai, Cascadia and Costa Rica¹⁵, although there is still ongoing debate on the precise location of tremor relative to the plate interface¹⁴.

Here I focus on tremor in western Shikoku. Tremor in this area has been proposed to be repeating small low-frequency earthquakes⁴, each of which occurs as slip on the plate interface based on both the planar distribution of hypocentres³ and the focal mechanism of low-angle thrust type¹¹. A high ratio of compressional (P)-wave velocity to shear (S)-wave velocity (v_p/v_s) around the tremor sources determined by seismic tomography implies the existence of water at high pore pressure^{3,16}. High-speed tremor migration of up to 10 km in 10 minutes (about 10 m s^{-1}) was also first discovered here⁴. Other categories of slow earthquakes, very-low-frequency earthquakes⁵ and slow slip events¹⁰ have been studied here as well. Here, I reveal the fine internal structure of this well-studied tremor zone by determining tremor source locations and durations.

The data are the velocity seismograms recorded at 23 stations of Hi-net, the high-sensitivity seismograph network maintained by the National Research Institute for Earth Science and Disaster Prevention, Japan, for five years from April 2004 to March 2009 (Fig. 1). I divided continuous data into half-overlapping 300-s time windows and determined the location of the tremor source, if it existed, for each time window. The hypocentre determination method is an envelope correlation method^{1,17}. This method has been widely used to determine the location of tremor phenomena, but has been thought to include significant errors, particularly in depth. However, it can be an effective location method, provided it is accompanied by appropriate processing to reduce the influence of outliers, as detailed in the Methods. In addition to the locations, I estimated the source duration using the stacked envelope aligned with the wave arrival time from the source location. The event duration is defined as the period for which the amplitude exceeds half of the maximum.

Figure 1 shows the locations of tremor sources and stations. For most events, the standard deviation of location determined by the least squares solution is less than 1 km, vertically and horizontally, which is probably underestimated. I can constrain the actual location error to some extent using ordinary earthquakes. I apply the envelope correlation method without eliminating ordinary earthquakes. 95% of ordinary earthquakes larger than magnitude one ($M > 1$) in the catalogue of the Japan Meteorological Agency (JMA) were also detected and located by the envelope correlation method. The estimated depths of these ordinary earthquakes are systematically shallower by about 1.2 km (with a standard deviation of 4.5 km) than those in the JMA catalogue. The epicentres are better constrained, with a standard deviation of about 1.5 km (see the Methods). I take these figures as representative of typical errors in my analysis. These ordinary earthquakes are not shown in Fig. 1, because it includes only events with duration longer than 10 s.

Despite errors of about 5 km in depth, the shape of tremor source distribution clearly shows northwestward-dipping planes, with increasing dip angles from east (A) to west (C) (see Fig. 1). The result is consistent with the previous location of low-frequency earthquakes in this area^{3,15}, but the resolution is insufficient to discuss whether this pattern represents a single plane. At the western end of the tremor zone, the depth may be a little deeper owing to systematic errors in depth, as demonstrated in the Supplementary Information. The relatively high dip angle in this region cannot be explained by this error. Except for the depth of this region, the depth distribution and the spatial pattern of dip angle changes are consistent with the depth of the oceanic Mohorovičić discontinuity (Moho) estimated using receiver functions¹⁸.

Many lineaments within the tremor can be identified, as shown by the arrows in Fig. 1. Their orientation clusters in one of two directions: either N 55° W or N 22° W. The former is the direction of current plate motion¹⁹ and the latter is approximately the direction

¹Department of Earth and Planetary Science, University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo 113-0033, Japan.

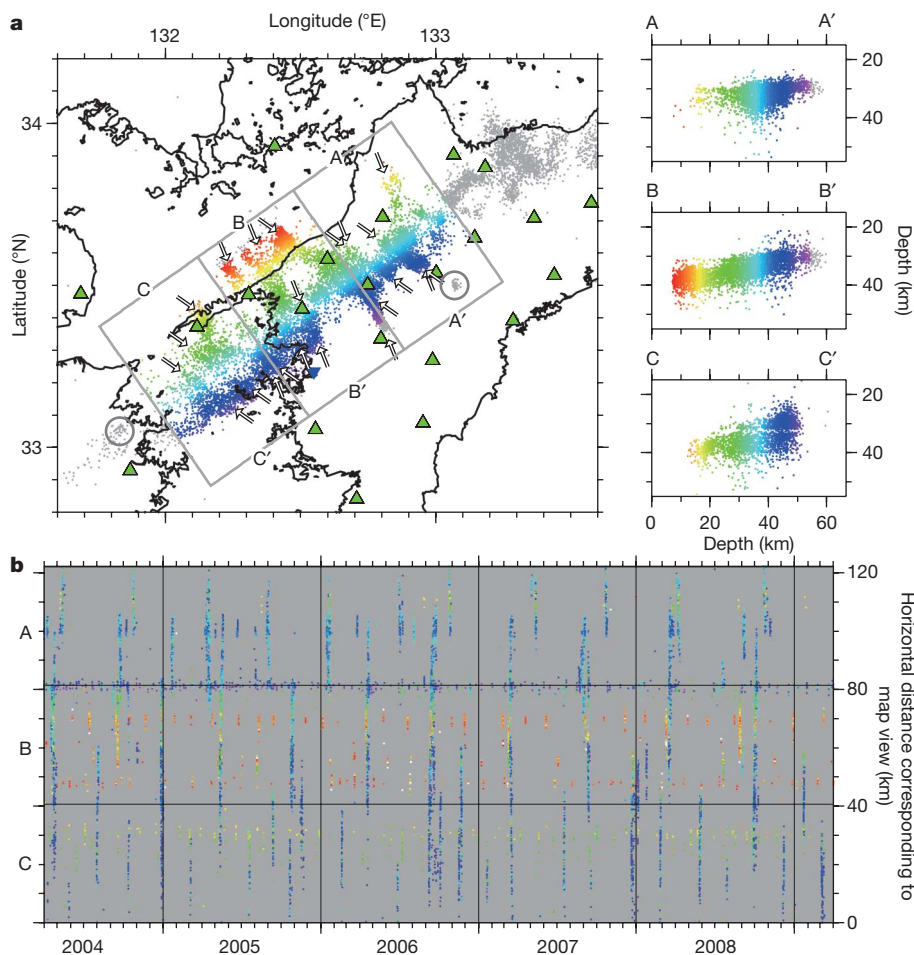


Figure 1 | Distribution of tremor sources having durations longer than 10 s. **a**, Map view (left panel) and three cross-sections (right panels) for the grey rectangular region. Sources in the box are coloured according to the location in the N 40° W direction, and sources outside the box are grey. Green

triangles are Hi-net stations and the blue triangle is the Uwajima tide station. Two grey open circles indicate explosion events in quarries. Arrows are aligned in either N 55° W or N 22° W directions. **b**, Space-time plot of tremor sources for five years. Colours are as in the top panels.

of the past plate motion before 3–5 million years ago, when the direction of subduction of the Philippine Sea plate suddenly changed from NNW to WNW^{20,21}. This suggests that these lineaments are formed by relative plate motion—both previous and current, just as micro-earthquake streaks are observed to follow the slip direction along the creeping section of the San Andreas Fault, California²².

The time-location plot of tremor sources (Fig. 1) shows that large tremor activity repeats at a time interval of 3–6 months, as found previously²³. Some small clusters repeat at much shorter intervals of 1–2 months, or last almost continuously, such as the purple cluster at the boundary between A and B. The intervals seem to be characteristic for each cluster, which suggests they are controlled by some local property of the plate interface.

Tremor distribution in Fig. 1 is substantially different from the distribution of low-frequency earthquakes detected by the JMA²⁴ (Fig. 2a). These low-frequency earthquakes are identified as isolated events in tremor sequences. Therefore, the event duration must be very short. If I select short events with durations between 10–15 s, the distribution (Fig. 2b) is similar to that of low-frequency earthquakes. This distribution is essentially the same as the recently discovered double-peak alignments of tremor activity at the shallow and deep parts of the tremor zone²⁵. On the other hand, the distribution of events with durations between 100–300 s (Fig. 2b) is quite different, and spatially complementary. This means that the event duration is also controlled by some local property of the interface. Figure 2c shows the typical duration measured by the median duration, which has a consistent pattern: the short-duration events are peripheral to, and surround, areas of long duration. There are several circular holes

near the centre of the tremor zone where no tremor was detected. Because the typical duration is long near these holes, it is possible that there are still longer slip events that are undetected by our analysis.

The same local property seems to control the sensitivity to tidal loading. Tremor can be modulated by tidal stress^{6,8,9}, but Fig. 2d shows that does not apply at all locations. I measure sensitivity to tidal stress using the normalized Fourier spectral amplitude, as described in the Methods. Although in some areas tremor occurs quite periodically at the semidiurnal principal lunar (M2) period, there are also wide areas without such correlation. By comparing Fig. 2c and d, in general tremor can be seen to be sensitive to tidal stress where short events occur. In other words, in such areas tremor is easily activated by external load but terminates quickly. Tremor may generate longer slip that is less sensitive to external load in an area where long events occur.

Tremor duration or sensitivity to tidal stress can control overall tremor behaviour. If tremor occurs without tidal modulation in an area where event duration is long, tremor often shows clear diffusive migration, in which the migration time is equal to the square of distance divided by the diffusivity. Figure 3a and b shows such examples. Tremor is slightly controlled by the tide, but migrates even when the tidal stress is in the opposite sense. A parabolic curve assuming the diffusivity of $10^4 \text{ m}^2 \text{ s}^{-1}$ is almost always a good approximation. Diffusion has been considered as a possible mechanism of tremor^{12,26}, but the migration is not always clearly diffusive. Figure 3c and d shows examples in which migration is not clear. In such cases, tremor occurs only when the tide level is high and migration is inhibited.

As shown above, there is some local property controlling event duration, periodicity, sensitivity to tidal stress and diffusive migration.

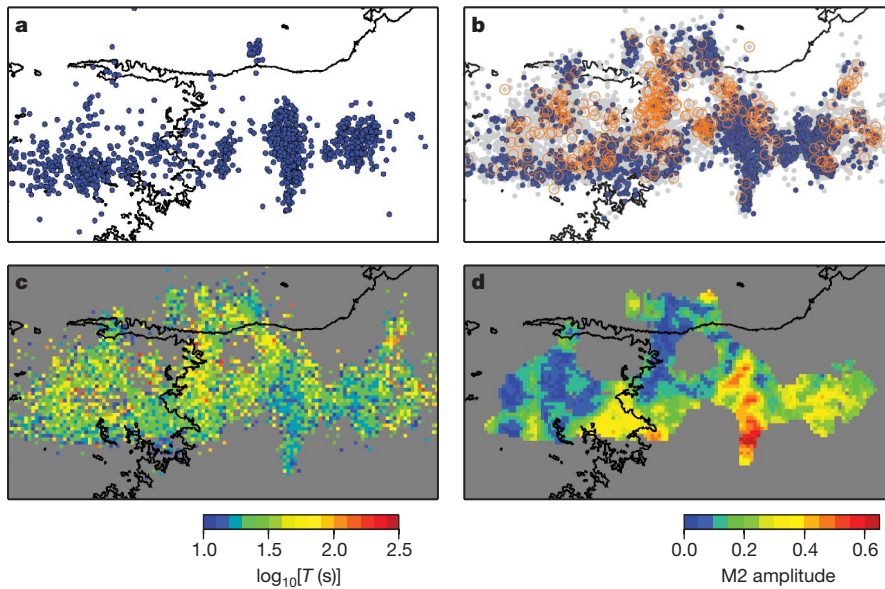


Figure 2 | Distribution of low-frequency earthquakes and tremor sources with duration and sensitivity to the tidal stress. **a**, Distribution of low-frequency earthquakes determined by the JMA. The area in the rectangular box in Fig. 1 is shown. **b**, The distribution of tremor sources is shown by grey circles. Those with duration T between 10 and 15 s and exceeding 100 s are shown by blue circles and orange circles, respectively. **c**, Spatial distribution of median duration estimated in each 1-km² cell. **d**, Sensitivity to the tidal stress estimated using the amplitude of the normalized Fourier spectrum at the semidiurnal principal lunar (M2) period of 12.420601 h.

What is it? Slow earthquakes may occur as a combination of brittle earthquake-like slip and ductile slower slip. In the deeper part of the plate interface, shear motion must be ductile, but in shallower areas a certain part of the interface is brittle. A tremor zone may be an area where a slow slip interface is pinned by numerous small brittle patches. The density of brittle patches could be the local property. Rupture of patches activates slow motion, but the amount of activated slow motion may be spatially variable. For dense patch distribution, only a small slow slip area is activated and the slow slip duration is short. Conversely, if the distribution is sparse, large slow motion can be activated around each patch and grows to some critical size, above which slow slip may break brittle patches without the help of tidal stress. In this case, a feedback system between slow slip and brittle patches produces a diffusive slip event, which is less controlled by tidal stress.

The duration is short along the shallowest edge of the tremor zone, where a clear lineament along strike suggests a drastic change of

material property. This line may correspond to the Moho discontinuity, where dehydration reaction and serpentinization of the stagnant wedge mantle are likely to begin^{27,28}, as suggested by the structural seismic surveys in the Tokai region, Japan²⁹. Serpentinization requires water transport along the subducting plate, which may be enhanced along with subducting irregularities such as seamounts. In the Nankai subduction zone, a subducting seamount near the trough has been identified by a structural survey³⁰. Many similar seamounts must have subducted along the plate interface in the past and some of them may have survived through the shallower locked section where megathrust earthquakes occur. Some of them subducted in the direction of N 22° W before 5 million years ago, whereas others have subducted or are subducting in the direction of N 55° W. I suggest that the current lineaments shown in Fig. 1 are the record of previous seamount subduction, as summarized in Fig. 4. Similar lineaments are also visible in other tremor zones along the Nankai and the Cascadia⁷ subduction zones, suggesting that it is a fundamental

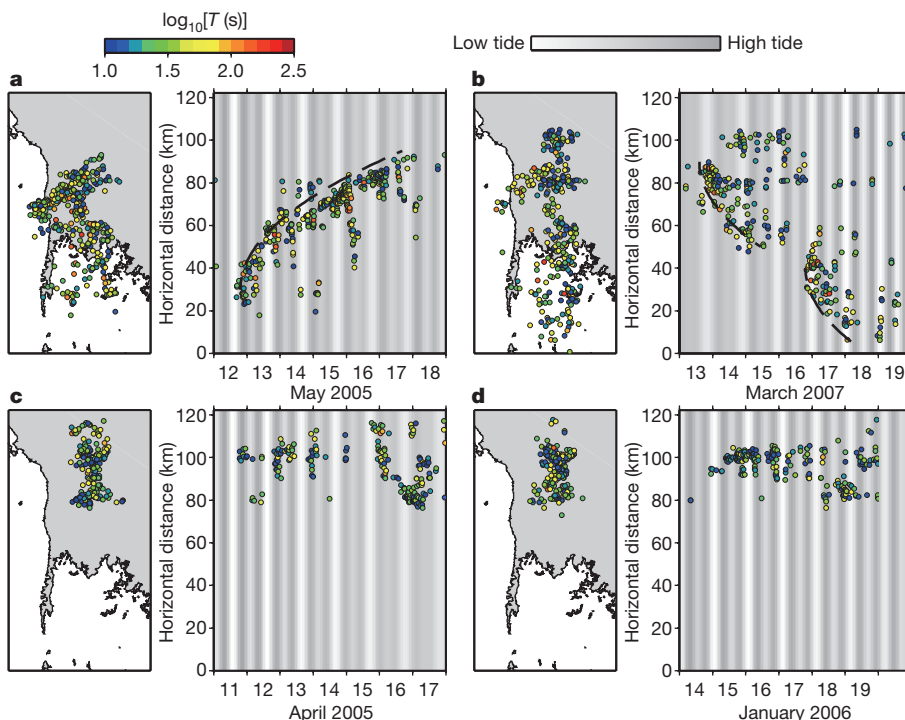


Figure 3 | Map views and space-time plots of tremor sources for active periods of one week. The area in the rectangular box in Fig. 1 is shown in the map view. Each space-time plot shows observed sea level at Uwajima station (Fig. 1) shaded in white to grey (normalized for each period). Coloured circles show the log duration. **a, b**, Tremor migration occurs almost continuously along a dashed parabola calculated using the diffusion coefficient of $10^4 \text{ m}^2 \text{ s}^{-1}$. **c, d**, Migration is not clear. Tremor occurs mainly when sea level is high.

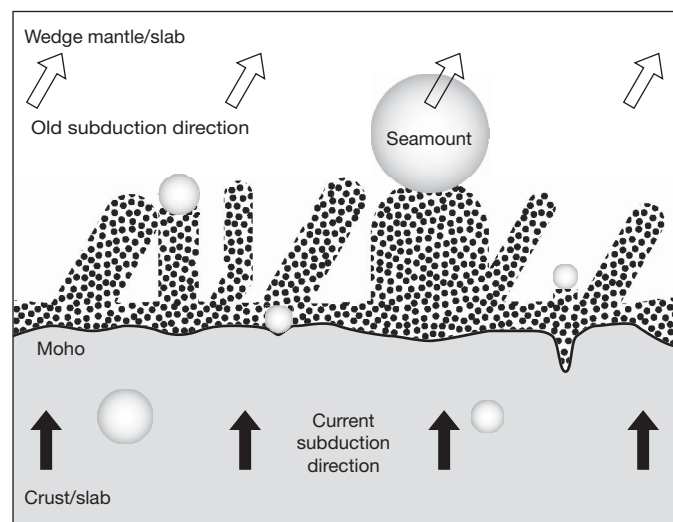


Figure 4 | Schematic illustration of plate subduction and formation of tremor zone. Guided by irregularities such as seamounts, the wedge mantle is serpentinized down to some depth. A change of subduction direction is visible as striations, shown by the dotted areas in the tremor zone.

process for tremor zones, and applicable to many other subduction zones.

METHODS SUMMARY

The hypocentre determination is based on an envelope correlation method: the envelope is ground-velocity bandpass-filtered between 2 and 8 Hz, squared, low-pass filtered below 0.3 Hz, and resampled at one sample per second, in two horizontal directions. Continuous envelope records are divided into 300-s windows. For each window, I seek a pair of envelopes, for which the maximum cross-correlation exceeds 0.6. For these pairs, I find an initial estimate of epicentral location that maximizes the number of station pairs with the same signs of difference in travel time and epicentral distance. The time difference at the maximum cross-correlation for each pair of envelopes is used as data. The data are compared with the difference of calculated S-wave arrival times based on a one-dimensional velocity structure. The best estimation of the hypocentre is given by minimizing the errors between the observed and calculated travel time differences using the Levenberg–Marquardt method. Although I determine a hypocentre for each time window, I discard the result in cases where the number of data points is smaller than 100, or the standard deviation of error is larger than a second. By comparison with the hypocentre catalogue of JMA (Supplementary Fig. 1), I estimate the standard deviation of the location to be about 1.5 km horizontally and 4.5 km vertically.

For the estimation of sensitivity to tidal stress, I prepare a circle centred at one point on the map to include 100 tremor epicentres. If the radius is smaller than 5 km, I consider a sequence of delta functions located at all hypocentral times. I calculate the amplitude of its Fourier spectrum at the semidiurnal principal lunar (M_2) period, which is shown in Fig. 2d.

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

Received 28 March; accepted 1 June 2010.

1. Obara, K. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* **296**, 1679–1681 (2002).
2. Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* **300**, 1942–1943 (2003).
3. Shelly, D. R., Beroza, G. C., Ide, S. & Nakamura, S. Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* **442**, 188–191 (2006).
4. Shelly, D. R., Beroza, G. C. & Ide, S. Non-volcanic tremor and low-frequency earthquake swarms. *Nature* **446**, 305–307 (2007).
5. Ito, Y., Obara, K., Shiomi, K., Sekine, S. & Hirose, H. Slow earthquakes coincident with episodic tremors and slow slip events. *Science* **315**, 503–506 (2007).
6. Shelly, D. R., Beroza, G. C. & Ide, S. Complex evolution of transient slip derived from precise tremor locations in western Shikoku, Japan. *Geochim. Geophys. Geosyst.* **8**, Q10014 (2007).

7. Ghosh, A. *et al.* Tremor bands sweep Cascadia. *Geophys. Res. Lett.* **37**, L08301, doi: 10.1029/2009GL042301 (2010).
8. Nakata, R., Suda, N. & Tsuruoka, H. Non-volcanic tremor resulting from the combined effect of Earth tides and slow slip events. *Nature Geosci.* **1**, 676–678 (2008).
9. Rubinstein, J. L., Rocca, M. L., Vidale, J. E., Creager, K. C. & Wech, A. G. Tidal modulation of nonvolcanic tremor. *Science* **319**, 186–189 (2008).
10. Hirose, H. & Obara, K. Repeating short- and long-term slow slip events with deep tremor activity around the Bungo channel region, southwest Japan. *Earth Planets Space* **57**, 961–972 (2005).
11. Ide, S., Shelly, D. R. & Beroza, G. C. The 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).
12. Ide, S., Beroza, G. C., Shelly, D. R. & Uchide, T. A scaling law for slow earthquakes. *Nature* **447**, 76–79 (2007).
13. Schwartz, S. Y. & Rokosky, J. M. Slow slip events and seismic tremor at circum-pacific subduction zones. *Rev. Geophys.* **45**, RG3004, doi: 10.1029/2006RG000208 (2007).
14. Rubinstein, J. L., Shelly, D. R. & Ellsworth, W. L. in *New Frontiers in Integrated Solid Earth Sciences, International Year of Planet Earth 287–314* (Springer Science and Business Media, 2010).
15. Brown, J. R. *et al.* Deep low-frequency earthquakes in tremor localize to the plate interface in multiple subduction zones. *Geophys. Res. Lett.* **36**, L19306, doi:10.1029/2009GL040027 (2009).
16. Matsubara, M., Obara, K. & Kasahara, K. High-VP/Vs zone accompanying non-volcanic tremors and slow-slip events beneath southwestern Japan. *Tectonophysics* **472**, 6–17 (2009).
17. Maeda, T. & Obara, K. Spatio-temporal distribution of seismic energy radiation from low-frequency tremor in western Shikoku, Japan. *J. Geophys. Res.* **114**, B00A09, doi: 10.1029/2008JB006043 (2009).
18. Shiomi, K., Obara, K. & Sato, H. Moho depth variation beneath southwestern Japan revealed from the velocity structure based on receiver function inversion. *Tectonophysics* **420**, 205–221 (2006).
19. Miyazaki, S. & Heki, K. Crustal velocity field of southwest Japan: Subduction and arc-arc collision. *J. Geophys. Res.* **106**, 4305–4326 (2001).
20. Seno, T. & Maruyama, S. Paleogeographic reconstruction and origin of the Philippine Sea. *Tectonophysics* **102**, 53–84 (1984).
21. Takahashi, M. Tectonic development of Japanese Islands controlled by Philippine Sea plate motion. *J. Geog.* **115**, 116–123 (2006).
22. Rubin, A. M., Gillard, D. & Got, J.-L. Streaks of microearthquakes along creeping faults. *Nature* **400**, 635–641 (1999).
23. Obara, K. & Hirose, H. Non-volcanic deep low-frequency tremors accompanying slow slips in the southwest Japan subduction zone. *Tectonophysics* **417**, 33–51 (2006).
24. 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).
25. Obara, K., Tanaka, S. & Maeda, T. Reevaluation of nonvolcanic tremor activity based on the hybrid method. *Eos* **90** (Fall Meet. Suppl.), abstr. T11C-1835 (2009).
26. Ide, S. A Brownian walk model for slow earthquakes. *Geophys. Res. Lett.* **35**, L17301, doi: 10.1029/2008GL034821 (2008).
27. Hyndman, R. D. & Peacock, S. M. Serpentinization of the forearc mantle. *Earth Planet. Sci. Lett.* **212**, 417–432 (2003).
28. Wada, I., Wang, K., He, J. & Hyndman, R. D. Weakening of the subduction interface and its effects on surface heat flow, slab dehydration, and mantle wedge serpentinization. *J. Geophys. Res.* **113**, B04402, doi: 10.1029/2007JB005190 (2008).
29. Kato, A. *et al.* Variations of fluid pressure within the subducting oceanic crust and slow earthquakes. *Geophys. Res. Lett.* doi: 10.1029/2010GL043723 (in the press).
30. Kodaira, S., Takahashi, N., Nakanishi, A., Miura, S. & Kaneda, Y. Subducted seamount imaged in the rupture zone of the 1946 Nankaido earthquake. *Science* **289**, 104–106 (2000).

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

Acknowledgements I thank G. C. Beroza for many useful comments. Comments from K. Wang and discussions with S. Yoshioka, T. Hori, B. Shibasaki, R. Ando and A. Namiki were helpful. This work is supported by JSPS KAKENHI (20340115) and MEXT KAKENHI (21107007). The figures were prepared using the Generic Mapping Tool (P. Wessel and W. H. F. Smith, 1998).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.I. (ide@eps.s.u-tokyo.ac.jp).

METHODS

Hypocentre determination method. The original records are horizontal-velocity discretized at 100 samples per second. The records are bandpass-filtered between 2 and 8 Hz, squared, low-pass filtered below 0.3 Hz, and resampled at one sample per second. The square root of the resampled data is the ‘envelope’ I use for hypocentre determination. I use envelopes in two horizontal directions separately at 23 stations. Continuous envelope records are divided into 300-s windows, successively overlapped by 150-s.

For each window, I first seek correlated envelopes. If the maximum cross-correlation between envelopes at a station pair exceeds the assumed threshold (0.6 in this study), the two envelopes are judged to be correlated and the corresponding lag time is the travel-time difference between two stations. The difference must not exceed the travel-time difference of a horizontally propagating S-wave at the surface. With a relatively low threshold such as 0.6, even noise often correlates well. Since least-squares methods are susceptible to non-Gaussian observations, that is, outliers, some preprocessing is necessary. I adopt a robust epicentral location method that maximizes the number of station pairs that have the same sign for difference in travel time and epicentral distance. For the estimated epicentre, a datum point is judged to be an outlier when (1) the arrival time at the distant station is earlier by more than 3 s, or (2) the travel-time difference is greater by 3 s than the travel time of the slowest wave between two stations. The threshold 3 s is another control parameter that I chose by trial and error.

After removing outliers, travel-time difference data Δt_{ij} between station i at $\mathbf{x}_i$ and station j at $\mathbf{x}_j$ are used to make an observation equation:

$$\Delta t_{ij} = T(\mathbf{x}_j, \mathbf{x}_0) - T(\mathbf{x}_i, \mathbf{x}_0) + \mathbf{e}$$

where $T(\mathbf{x}, \mathbf{y})$ represents the travel time of an S-wave between $\mathbf{x}$ and $\mathbf{y}$. I calculate travel times assuming a one-dimensional velocity structure based on the tomography result of ref. 3. In fact, the difference of the assumed velocity structure affects the epicentre distribution very little. $\mathbf{x}_0$ is the hypocentre to be estimated. Assuming that $\mathbf{e}$ is the Gaussian error leads to a nonlinear least-squares minimization:

$$\text{Res} = \sum [\Delta t_{ij} - (T(\mathbf{x}_j, \mathbf{x}_0) - T(\mathbf{x}_i, \mathbf{x}_0))]^2$$

I linearize this problem and solve it using the Levenberg–Marquardt method. For this nonlinear problem the selection of the initial parameters may affect the

result. Therefore, I use the estimated epicentre and seven different initial depths—2, 5, 10, 20, 30, 50 and 100 km—and find the solution that minimizes Res. Once the solution is obtained, I calculate the standard deviation of data. If a data residual is more than three times the standard deviation, I remove it and restart the Levenberg–Marquardt method. This process is iterated until there are no removed data points.

Thus I obtain the catalogue of hypocentres estimated for each time window. Some hypocentres are not unique because time windows overlap. I search a pair of hypocentres that occurred within 40 km and 10 s, and removed the one with the larger standard deviation. I also removed poorly estimated hypocentres in cases where the number of data points is not many, smaller than 100, or the standard deviation is larger than 1 s.

Comparison of hypocentres with the JMA catalogue. My method is applicable not only to tremor, but also to ordinary earthquakes. I compare my result with the hypocentre catalogue of JMA. Supplementary Fig. 1 shows the result of comparison. Corresponding events are connected by lines. Generally, the epicentre is similar, with a minor systematic error and a standard deviation of about 1.5 km. The vertical error is a little larger, 4.5 km, and my hypocentral estimate is systematically shallower by about 1 km than that of the JMA. The difference is mostly due to the difference of assumed velocity structure. Moreover, the vertical error has a spatial pattern in that the error is larger at the edge of the network. Therefore, we should be careful not to interpret the hypocentral depth near the edge of the network. Nevertheless, considering the fact that my method uses only S-waves, the consistency between two catalogues is remarkable and the resultant hypocentral catalogue is useful for this analysis.

Estimation of sensitivity to tidal stress. For each point separated at intervals of 2 km in the box area of Fig. 1, I estimated the sensitivity of tremor activity to tidal stress as follows. I prepare a circle centred at one point on the map to include 100 tremor epicentres. If the radius is smaller than 5 km, I consider a sequence of delta functions located at all hypocentral times, t_j . I calculate the amplitude of its Fourier spectrum at the semidiurnal principal lunar (M2) period, $\omega_{M2} = 12.420601$ h, normalized by the number of events $N (= 100)$, which is:

$$\frac{1}{N} \left| \sum_j \exp(i\omega_{M2} t_j) \right|$$

If the sequence is perfectly periodic, the value is 1. Figure 2d shows the distribution of this normalized Fourier spectrum amplitude.

LETTERS

New Oligocene primate from Saudi Arabia and the divergence of apes and Old World monkeys

Iyad S. Zalmout^{1,2}, William J. Sanders^{1,3}, Laura M. MacLatchy^{1,3}, Gregg F. Gunnell¹, Yahya A. Al-Mufarreah⁴, Mohammad A. Ali⁴, Abdul-Azziz H. Nasser⁴, Abdu M. Al-Masari⁴, Salih A. Al-Sobhi⁴, Ayman O. Nadhra⁴, Adel H. Matari⁴, Jeffrey A. Wilson^{1,2} & Philip D. Gingerich^{1,2,3}

It is widely understood that Hominoidea (apes and humans) and Cercopithecoidea (Old World monkeys) have a common ancestry as Catarrhini deeply rooted in Afro-Arabia^{1–4}. The oldest stem Catarrhini in the fossil record are Propliopithecoidea, known from the late Eocene to early Oligocene epochs (roughly 35–30 Myr ago) of Egypt, Oman and possibly Angola^{5–10}. Genome-based estimates for divergence of hominoids and cercopithecoids range into the early Oligocene¹¹; however, the mid-to-late Oligocene interval from 30 to 23 Myr ago has yielded little fossil evidence documenting the morphology of the last common ancestor of hominoids and cercopithecoids, the timing of their divergence, or the relationship of early stem and crown catarrhines. Here we describe the partial cranium of a new medium-sized (about 15–20 kg) fossil catarrhine, *Saadanius hijazensis*, dated to 29–28 Myr ago. Comparative anatomy and cladistic analysis shows that *Saadanius* is an advanced stem catarrhine close to the base of the hominoid–cercopithecoid clade. *Saadanius* is important for assessing competing hypotheses about the ancestral morphotype for crown catarrhines^{12–14}, early catarrhine phylogeny^{12,15} and the age of hominoid–cercopithecoid divergence¹¹. *Saadanius* has a tubular ectotympanic but lacks synapomorphies of either group of crown Catarrhini, and we infer that the hominoid–cercopithecoid split happened later, between 29–28 and 24 Myr ago.

The catarrhine fossil record subsequent to propliopithecoids was unknown except for isolated dentitions of the late-Oligocene *Kamoyapithecus hamiltoni* from Kenya¹⁶ until the Afro-Arabian early Miocene (about 23–16 Myr ago) diversification of apes^{5,16} (considered as basal or ‘eo’¹⁷-hominoids, and/or late stem catarrhines and placed in ‘Proconsuloidea’^{3,12} and ‘Dendropithecoidea’^{3,18}) and cercopithecoids^{2,4} (Supplementary Information). The paucity of fossil evidence from about 30–23 Myr ago has inhibited the testing of competing hypotheses about the attributes of the ancestral morphotype of cercopithecoids and hominoids^{1,12–14}, the age of their divergence¹¹ and the nature of the propliopithecoid–crown-catarrhine relationship^{12,15}.

A new Oligocene catarrhine fossil (SGS-UM 2009-002), preserving substantial facial, palatal and dental morphology, was recently recovered from the top of an oolitic ironstone bone bed of the middle unit of the Shumaysi Formation, at Harrat Al Ujayfa in Al Hijaz Province, western Saudi Arabia (Fig. 1a–c) by a Saudi Geological Survey–University of Michigan expedition. This is the most substantive evidence of catarrhines from the late early Oligocene through to the late Oligocene, further emphasizing Afro-Arabia as the locus of early catarrhine evolution (Fig. 1c).

The Shumaysi Formation was deposited before the early Miocene opening of the Red Sea rift^{19–21}. It rests partly on the late Cretaceous–Eocene Usfan Formation, and is overlain in places by Miocene

sandstones of the Khulays Formation (Fig. 1b)²¹. The middle unit of the formation, from which SGS-UM 2009-002 and its associated fauna derive, represents a non-marine back-mangrove environment²⁰. Previous age estimates for the Shumaysi Formation extend from Eocene to Oligocene–early Miocene^{19,20,22}. Volcanic dykes intruding the upper Shumaysi Formation yield radioisotopic ages of 26–21 Myr (ref. 22), and the oldest lava flows overlying the formation and tuffaceous rocks interbedded in the upper unit produce dates of 25–21 Myr ago²³, indicating that the middle unit is pre-Miocene in age.

Chronostratigraphy of the middle Shumaysi Formation is refined by biochronological analysis of its mammalian assemblage. The fauna includes paenungulates and anthracotheriid artiodactyls with closely related counterparts in the upper part of the Jebel Qatrani Formation sequence of Fayum, Egypt. These taxa are less advanced than comparable animals at the late Oligocene sites of Chilga in Ethiopia and Lothidok in Kenya. They are accompanied by primitive gomphotheres and mammutids reminiscent of proboscideans found at Chilga and Lothidok but never documented in the rich fauna of the Fayum. Together, the assemblage can be temporally interpolated best between the Jebel Qatrani and Chilga Formations, making it 29–28 Myr old (Supplementary Information).

SGS-UM 2009-002 preserves most of the face, anterior neurocranium and palate, with a partial dentition of left and right I2, left C and M1–3, as well as roots of right C and P3 and left P3–4, and alveoli for left and right I1 (Fig. 2a–c). The left canine is broken apically. Some enamel has spalled off the molars, but morphometric details can be determined (Supplementary Information). There is a deep bite mark in the frontal trigon and another, possibly a fatal puncture wound, in the right side of the endocranial cavity. In addition, part of the right temporal bone is preserved, as is the right occipital condyle (Fig. 2d–g).

Salient features of SGS-UM 2009-002 include low hafting of the face on the neurocranium, snout-like projection of the midface, strong anterior convergence of temporal lines, a tubular ectotympanic (Fig. 2d, e), absence of frontal sinuses (Fig. 3a), and large, broad molars. The dental development, length of the canine roots (Fig. 3b), extent of the C–I2 diastemata, molar size, and expression of the temporal lines to form a sagittal crest indicate that the cranium is that of an adult male. Primitive catarrhine features predominate in SGS-UM 2009-002 (Supplementary Information). Together, these features comprise a distinctive taxon that differs sufficiently from early Oligocene–early middle Miocene catarrhines to be placed in a new genus and species that cannot be readily accommodated within any existing catarrhine family or superfamily (Supplementary Information).

¹Museum of Paleontology, University of Michigan, 1109 Geddes Avenue, Ann Arbor, Michigan 48109, USA. ²Department of Geological Sciences, University of Michigan, 2534 C. C. Little Building, 1100 North University Avenue, Ann Arbor, Michigan 48109, USA. ³Department of Anthropology, University of Michigan, 1085 South University Avenue, Ann Arbor, Michigan 48109, USA. ⁴Saudi Geological Survey, Paleontology Unit, PO Box 54141, Jeddah, 21514, Saudi Arabia.

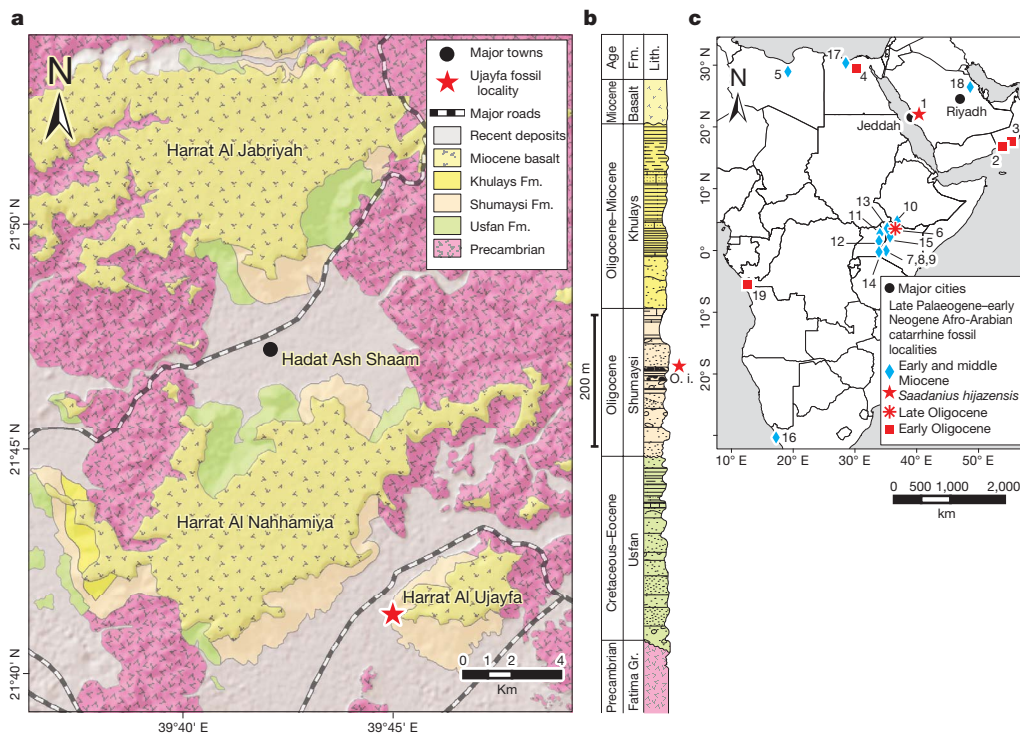


Figure 1 | Geological, geographical and stratigraphic contexts for SGS-UM 2009-002. **a**, Geological map of *Saadanius hijazensis* type site at Harrat Al Ujayfa, Saudi Arabia^{18,20}. **b**, Stratigraphic provenance of SGS-UM 2009-002. O.i., oolitic ironstone horizon. **c**, Major early Oligocene to early middle Miocene Afro-Arabian catarrhine primate sites. Key: 1, Harrat Al Ujayfa, Saudi Arabia; 2, Thaytinit, Oman; 3, Taqah, Oman; 4, Fayum, Egypt; 5,

Gebel Zelten, Libya; 6, Lothidok, Kenya; 7, Meswa Bridge, Kenya; 8, Koru, Kenya; 9, Songhor, Kenya; 10, Buluk, Kenya; 11, Moroto, Uganda; 12, Napak, Uganda; 13, Kalodirr, Kenya; 14, Rusinga, Kenya; 15, Loperot, Kenya; 16, Ryskop, South Africa; 17, Wadi Moghara, Egypt; 18, Ad Dabtiyah, Saudi Arabia; 19, Malembe, Angola.

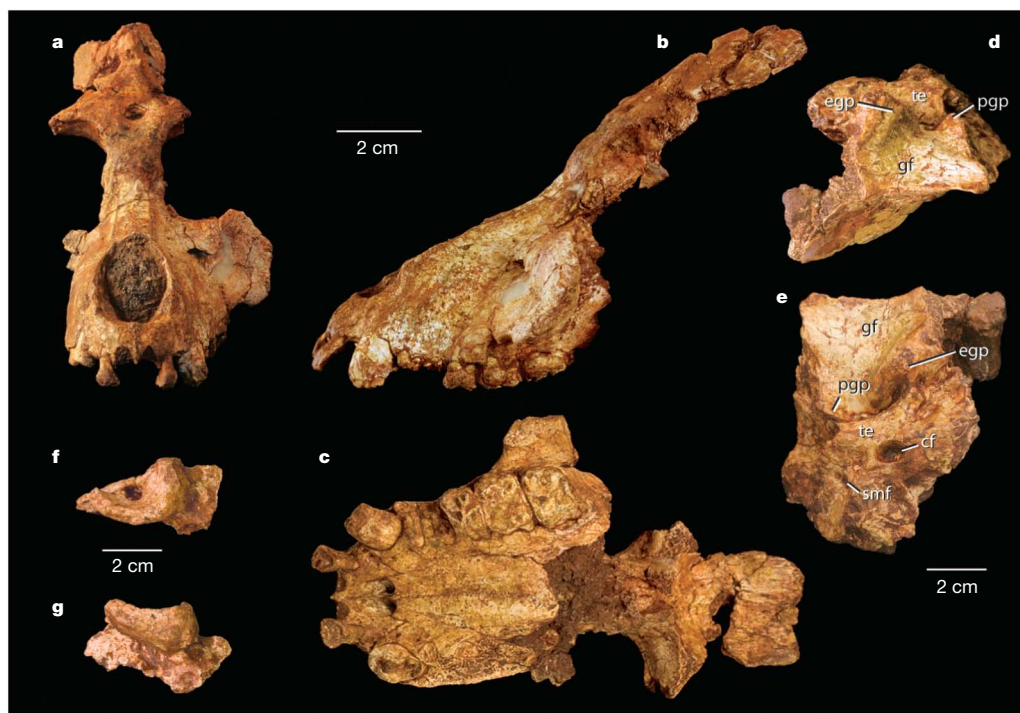


Figure 2 | Photographic images of the partial cranium SGS-UM 2009-002, type specimen of *Saadanius hijazensis* gen. et sp. nov. **a**, Cranium in anterior view. **b**, Cranium in lateral view, anterior to the left. **c**, Ventral view of cranium, anterior to the left. **d**, Anterior view of right temporal bone, ventral at top, lateral to the right. **e**, Ventral view of right temporal bone,

anterior at top, lateral to the left. **f**, Medial view of right occipital condyle, anterior to the left. **g**, Ventral view of right occipital condyle, anterior to the left. cf, carotid foramen; egp, entoglenoid process; gf, glenoid fossa; ppg, postglenoid process; smf, stylomastoid foramen; te, tubular ectotympanic.

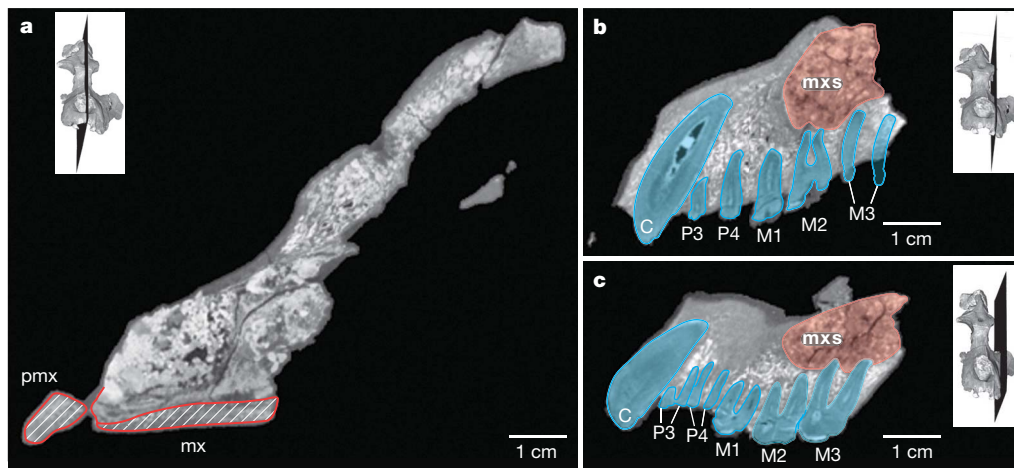


Figure 3 | Micro-CT scan images of the partial cranium SGS-UM 2009-002, type specimen of *Saadanius hijazensis* gen. et sp. nov. **a**, Midline sagittal slice. Offset of the premaxilla (pmx) and palatal process of the maxilla (mx), and thickness of the maxillary palatal process are shown by a red outline with white cross-hatching. Note the absence of frontal sinuses. **b**, Sagittal slice through the left canine. Note the large size and deeply

embedded canine root characteristic of male catarrhines. **c**, Sagittal slice through the greatest anteroposterior extension of the maxillary sinus (mxs). Note the high floor of the sinus and its anterior extension above M2. In **b** and **c** teeth and tooth roots are shaded in blue, and the maxillary sinus is shaded in red.

Order Primates, Linnaeus, 1758

Suborder Anthropoidea, Mivart, 1864

Infraorder Catarrhini, É. Geoffrey Saint-Hilaire, 1812

Superfamily **Saadaniioidea** nov.

Family **Saadaniidae** nov.

Saadanius gen. nov.

Saadanius hijazensis sp. nov.

Etymology. *Saadanius*, after *saadan* (Arabic): collective term for apes and monkeys. The specific term refers to the Al Hijaz region, western province of Saudi Arabia, where the new catarrhine was discovered. **Holotype.** SGS-UM 2009-002 (Fig. 2a–g). The type is housed in the Paleontology Unit of the Saudi Geological Survey in Jeddah, Saudi Arabia.

Locality. Southwest corner of Harrat Al Ujayfa, Al Hijaz Province, western Saudi Arabia (Fig. 1a, c).

Diagnosis. Medium-sized catarrhine (cranium similar in size to *Symphalangus*; about 15–20 kg); upper molars relatively broad, enamel thin, trigon cusps well spaced and set apart from large, distolingually situated hypocones; upper lateral incisors spatulate and probably smaller than I1s; upper canines ovoid in cross-section, with a prominent distal flange and stout, straight, elongate root (Fig. 3b); canine crown area modest relative to molar size; splanchnocranium set below the neurocranium; midfacial projection strong; orbits high; nasal bones long and narrow; nasal aperture large and oval-shaped; anterior facing and vertically oriented, low-rooted malar; strong temporal lines convergent anteriorly, enclosing a small frontal trigon; maxillary sinuses with high floor and anterior limit at M2 (Fig. 3c); no frontal sinus (Fig. 3a); palate shallow and broadest posteriorly; spacious, bifid incisive fossa; premaxillae not overlapping maxillary palatal processes (Fig. 3a); nasolabial clivus low; temporal bone with a short, tubular ectotympanic, broad, flat glenoid process, and massive entoglenoid process; diminutive occipital condyle. *Saadanius* is the type genus of the new family Saadaniidae, which is differentiated from other catarrhine families by features provided in the Supplementary Information.

Comparative and phylogenetic analyses of Catarrhini (Supplementary Information) indicate that *Saadanius hijazensis* is a stem catarrhine, phylogenetically close to crown catarrhines (Fig. 4) and morphologically approximate to the condition of the last common ancestor of crown catarrhines (LCA). A preponderance of features of *Saadanius* are shared with male crania of the early stem catarrhine *Aegyptopithecus zeuxis* or are intermediate between *Aegyptopithecus* and early Miocene apes (Supplementary Information).

A suite of other traits in early Miocene apes, such as frontal sinuses, increased molar trigon crowding, broadened anterior palates, more pronounced lower facial prognathism, relatively enormous male canine crowns and larger body size^{3,13,17,24–26} are absent in *Saadanius*, other stem Catarrhini and basal cercopithecoids (Supplementary Information). Distribution of these traits is consistent with the hypothesis that early Miocene apes are hominoids (Fig. 4) and that initial evolution of the group involved changes to the craniofacial region^{24,26}. The adaptive significance of these traits is probably correlated with shifts in masticatory function, diet (see, for example, ref. 15) and social behaviour, possibly related to body size increase, which itself has numerous advantages²⁷.

Early Miocene apes show few modern hominoid synapomorphies, complicating the identification of basal hominoids²⁷. The ways in which early Miocene apes differ anatomically from *Saadanius* and other stem catarrhines are crucial for providing a method for recognizing basal hominoids. For example, ‘dendropithecoids’ have resisted definitive classification because of their primitive postcranial morphology and a lack of information about cranial features such as the form of the ectotympanic^{3,18}. However, possession of frontal sinuses and relatively enormous male canines differentiate them from propliopithecoids, pliopithecoids, *Saadanius* and basal cercopithecoids, and link them with ‘proconsuloids’ as possible hominoids (see also refs 24, 26, 27). Similarly, a relatively large male canine in the late Oligocene catarrhine *Kamoyapithecus*¹⁶ may signal that it is also a basal hominoid (Supplementary Information).

Unlike propliopithecoids^{8,10} (and pliopithecoids²⁸), however, *Saadanius* shares with crown catarrhines a fully formed tubular ectotympanic, which presumably was present in the LCA and evolved only once in Catarrhini. The presence of a tubular ectotympanic in a cranium dominated by otherwise primitive catarrhine features is compelling evidence of a close phylogenetic relationship between *Saadanius* and crown catarrhines.

Neontology-based character-state analyses reconstruct the LCA cranial morphotype with a short, relatively orthognathic face, short, broad nasals and a globular braincase (see, for example, refs 12, 14). Conversely, studies incorporating observations on early Miocene apes and the basal cercopithecoid *Victoriapithecus* reconstruct an LCA morphotype with a projecting snout and tall face, long, narrow nasal bones and steep frontal with a trigon enclosed by anteriorly convergent temporal lines and supraorbital costae¹. In the second hypothesis, a stem catarrhine similar to the propliopithecoid *Aegyptopithecus* could have had a more direct role in the evolution

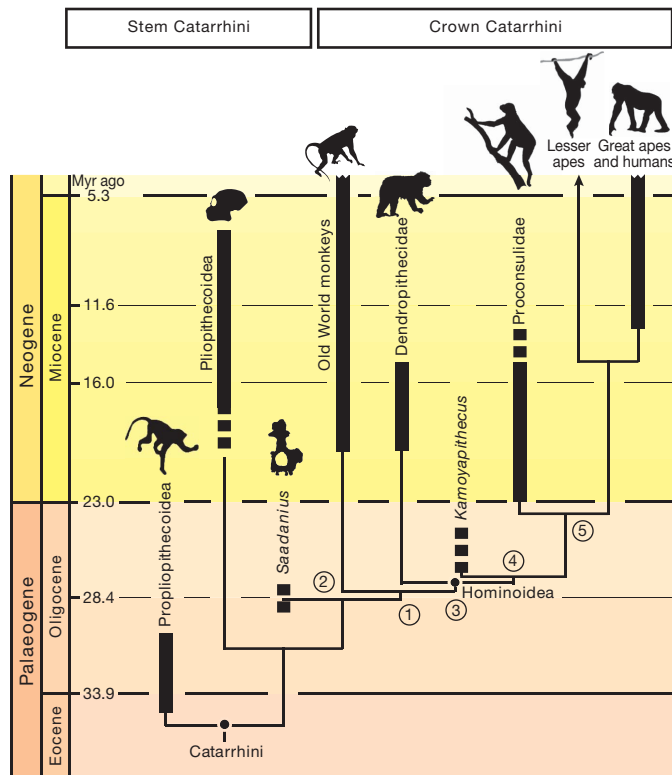


Figure 4 | Proposed relationship of *Saadanius* to other catarrhine primates.

The diagram is based on Supplementary Information and refs 1, 2, 13, 16, 17, 24, 26, 27. Thick solid vertical lines indicate known temporal ranges of taxa; thick dotted vertical lines show intervals of occurrence or possible extension for temporal ranges of taxa. The following characters define these clades: (1) *Saadanius* + crown catarrhines, presence of complete tubular ectotympanic; (2) Cercopithecoidea, bilophodont molars, absence of maxillary sinuses, absence of atrioturbinal/vomeronasal functional complex, premaxillary-maxillary suture extending to top of nasals, upper canine sulcus extending onto root; (3) Hominoidea, presence of frontal sinus, male upper canines relatively very large in cross-section; (4) *Kamoyapithecus* + proconsulids + crown hominoids (including lesser apes, great apes and humans), body size increase; (5) proconsulids + crown hominoids, palate broadest anteriorly, pronounced lower facial prognathism, loss of tail, deep, narrow humeral zona conoidea, deep olecranon fossa, absence of entepicondylar foramen and epitrochlear fossa.

of later crown catarrhines, because its cranial morphology closely approximates this latter ancestral morphotype.

The second, paleontology-based, hypothesis¹ is more strongly corroborated by the presence in *Saadanius* of nearly all craniodental features predicted by it for the LCA and is further supported by the observation that many of the traits present in *Aegyptopithecus* and *Saadanius* (for example midfacial projection and long nasals^{6,8}) are retained in early Miocene apes, most notably *Afropithecus*^{15,25}, *Morotopithecus* and *Proconsul*^{17,29}. Although primitive, the consistent expression of these features suggests a conservation of the stem catarrhine male cranial morphotype well into the hominoid lineage. These features are less well developed in female propiopithecoid and ape crania, as a result of the strong degree of sexual dimorphism characteristic of early catarrhines (see, for example, refs 6, 10, 25). The correspondence of the intermediate temporal occurrence and intermediate morphological development of *Saadanius* between propiopithecoids and Miocene apes (Fig. 4 and Supplementary Information) is further suggestive of phylogenetic continuity among these taxa.

The age of the hominoid–cercopithecoidean divergence has frequently been estimated as 25–23 Myr. However, recent genomic investigation calculated an earlier incidence of this event between 34.5 and 29.2 Myr ago¹¹, nearly coincident with the entire temporal range of propiopithecoids⁹. It is possible that wider sampling of Afro-Arabia might

document this event within the older time interval, but no fossils of crown catarrhines have been recovered from this time span, and the oldest documented proconsulids, dendropithecids and cercopithecoideans only date to 23–20 Myr ago^{2–4}. *Saadanius* provides new evidence consistent with a divergence date after 29–28 Myr ago, and its comparative study offers a basis for recognition of the first hominoids subsequent to this event. If the late-Oligocene *Kamoyapithecus*¹⁶ is a hominoid, then the split occurred between 29–28 Myr and 24 Myr ago.

METHODS SUMMARY

Micro-computed tomography (micro-CT) scan imaging. SGS-UM 2009-002 was scanned with the OMNI-X Industrial Scanner at the Center for Quantitative Imaging, Pennsylvania State University. This was scanned in volume mode with a Feldkamp reconstruction algorithm.

Phylogenetic analysis. Phylogenetic analysis was carried out with TNT (Tree Analysis Using New Technology) made freely available by the Willi Hennig Society³⁰. A total of 36 characters were analysed for 19 taxa (Supplementary Information) including *Pan* and *Hylobates* as extant representatives of Hominoidea and using the basal fossil catarrhine *Aegyptopithecus* to root all trees. The Implicit Enumeration Algorithm was used to ensure that all possible trees were explored, with and without the collapse-tree option. Results in each case were identical. No characters were weighted and two characters were ordered. Consistency and Retention Indices were calculated on the basis of minimum, maximum and most-parsimonious steps for each resulting tree. Analyses were performed for all taxa and characters; all taxa, excluding postcranial characters; appropriate taxa collapsed into Dendropithecidae and Proconsulidae and all characters; and collapsed taxa, excluding postcranial characters.

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

Received 15 February; accepted 16 April 2010.

- Benefit, B. R. & McCrossin, M. L. Facial anatomy of *Victoriapithecus* and its relevance to the ancestral cranial morphology of Old World monkeys and apes. *Am. J. Phys. Anthropol.* **92**, 329–370 (1993).
- Benefit, B. R. & McCrossin, M. L. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 241–253 (Cambridge Univ. Press, 2002).
- Harrison, T. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 311–338 (Cambridge Univ. Press, 2002).
- Miller, E. R. et al. Systematics of early and middle Miocene Old World monkeys. *J. Hum. Evol.* **57**, 195–211 (2009).
- Pickford, M. Première découverte d'une faune mammalienne terrestre paléogène d'Afrique sub-saharienne. *C. R. Acad. Sci. Paris (II)* **302**, 1205–1210 (1986).
- Simons, E. L. New faces of *Aegyptopithecus* from the Oligocene of Egypt. *J. Hum. Evol.* **16**, 273–289 (1987).
- Thomas, H. et al. in *Fossil Vertebrates of Arabia. With Emphasis on the Late Miocene Faunas, Geology, and Palaeoenvironments of the Emirate of Abu Dhabi, United Arab Emirates* (eds Whybrow, P. J. & Hill, A.) 430–442 (Yale Univ. Press, 1999).
- Rasmussen, D. T. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 203–220 (Cambridge Univ. Press, 2002).
- Seiffert, E. R. Evolution and extinction of Afro-Arabian primates near the Eocene–Oligocene boundary. *Folia Primatol. (Basel)* **78**, 314–327 (2007).
- Simons, E. L., Seiffert, E. R., Ryan, T. M. & Attia, Y. A remarkable female cranium of the early Oligocene *Aegyptopithecus zeuxis* (Catarrhini, Propiopithecidae). *Proc. Natl Acad. Sci. USA* **104**, 8731–8736 (2007).
- Steiper, M. E., Young, N. M. & Sukarna, T. Y. Genomic data support the hominoid slowdown and an early Oligocene estimate for the hominoid–cercopithecoidean divergence. *Proc. Natl Acad. Sci. USA* **101**, 17021–17026 (2004).
- Harrison, T. The phylogenetic relationships of the early catarrhine primates: a review of the current evidence. *J. Hum. Evol.* **16**, 41–80 (1987).
- Rae, T. C. in *Function, Phylogeny, and Fossils. Miocene Hominoid Evolution and Adaptations* (eds Begun, D. R., Ward, C. V. & Rose, M. D.) 59–77 (Plenum, 1997).
- Delson, E. & Andrews, P. in *Phylogeny of the Primates. A Multidisciplinary Approach* (eds Luckett, W. P. & Szalay, F. S.) 405–446 (Plenum, 1975).
- Leakey, M. G. et al. Similarities in *Aegyptopithecus* and *Afropithecus* facial morphology. *Folia Primatol. (Basel)* **56**, 65–85 (1991).
- Leakey, M. G., Ungar, P. S. & Walker, A. A new genus of large primate from the late Oligocene of Lothidok, Turkana District, Kenya. *J. Hum. Evol.* **28**, 519–531 (1995).
- Begun, D. R. in *Phylogeny of the Neogene Hominoid Primates of Eurasia* (eds de Bonis, L., Koufos, G. D. & Andrews, P.) 231–253 (Cambridge Univ. Press, 2001).
- Harrison, T. A taxonomic revision of the small catarrhine primates from the early Miocene of East Africa. *Folia Primatol. (Basel)* **50**, 59–108 (1988).
- Al-Shanti, A. M. S. Oolitic iron ore deposits in Wadi Fatima between Jeddah and Mecca, Saudi Arabia. *Saudi Arabian Dir. Gen. Min. Resources Bull.* **2**, 1–51 (1966).
- Srivastava, S. K. & Binda, P. L. Depositional history of the early Eocene Shumaysi Formation, Saudi Arabia. *Palynology* **15**, 47–61 (1991).

21. Abou Ouf, M. A. & Gheith, A. M. in *Sedimentation and Tectonics of Rift Basins: Red Sea and Gulf of Aden* (eds Purser, B. H. & Bosence, D. W. J.) 135–145 (Chapman & Hall, 1998).
22. Moltzer, J. G. & Binda, P. L. Micropaleontology and palynology of the middle and upper members of the Shumaysi Formation, Saudi Arabia. *Saudi Arabian Dir. Gen. Min. Resources Bull.* **4**, 47–76 (1981).
23. Féraud, G., Zumbo, V. & Sebai, A. $^{40}\text{Ar}/^{39}\text{Ar}$ age and duration of tholeiitic magmatism related to the early opening of the Red Sea rift. *Geophys. Res. Lett.* **18**, 195–198 (1991).
24. Rae, T. C. Mosaic evolution in the origin of the Hominoidea. *Folia Primatol. (Basel)* **70**, 125–135 (1999).
25. Leakey, R. E., Leakey, M. G. & Walker, A. C. Morphology of *Afropithecus turkanensis* from Kenya. *Am. J. Phys. Anthropol.* **76**, 289–307 (1988).
26. Rae, T. C. Miocene hominoid craniofacial morphology and the emergence of great apes. *Ann. Anat.* **186**, 417–421 (2004).
27. Kelley, J. in *Function, Phylogeny, and Fossils. Miocene Hominoid Evolution and Adaptations* (eds Begun, D.R., Ward, C.V. & Rose, M.D.) 173–208 (Plenum, 1997).
28. Begun, D. R. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 221–240 (Cambridge Univ. Press, 2002).
29. Teaford, M. F., Beard, K. C., Leakey, R. E. & Walker, A. New hominoid facial skeleton from the Early Miocene of Rusinga Island, Kenya, and its bearing on the relationship between *Proconsul nyanzae* and *Proconsul africanus*. *J. Hum. Evol.* **17**, 461–477 (1988).
30. Goloboff, P., Farris, J. & Nixon, K. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).

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

Acknowledgements We thank Saudi Geological Survey (Jeddah) officials Z. A. Nawab, A. M. Al Attas and A. F. Al Khattabi for the invitation to undertake fieldwork in western Saudi Arabia; E. Mbua (National Museums of Kenya) for permission to study of catarrhine fossils in her care; A. Walker for access to micro-CT facilities at Pennsylvania State University; T. Ryan for scanning; A. Rountrey, D. Erickson and B. Miljour for assistance with interpretation of micro-CT scans, photography, and figures; T. Harrison for phylogenetic discussion and advice; E. Seiffert for CT images and photographs of propliopithecoid fossils; B. Benefit for information on the ectotympanic of *Victoriapithecus*; D. T. Rasmussen and M. Gutiérrez for information about fossil mammals from Lothidok, Kenya; M. Halawani, A. Memesh, S. Dini, C. Spencer and G. W. Hughes for information about the geology and lithostratigraphy of the research area; and E. Seiffert for improving the manuscript. The research was supported by a US National Science Foundation grant (EAR-0517773) to P.D.G.

Author Contributions I.S.Z., Y.A.A., M.A.A., A.H.N., A.M.A., S.A.A., A.O.N., A.H.M. and J.A.W. undertook the palaeontological survey, fossil collection and field geology. W.J.S. prepared the fossils. G.F.G., L.M.M. and W.J.S. performed the comparative study and phylogenetic analysis of catarrhine primates. L.M.M. and P.D.G. coordinated the micro-CT scanning. G.F.G. and W.J.S. made comparative studies of the non-primate mammalian fauna and identified the taxa. All authors contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to I.S.Z. (zalmouti@umich.edu) or W.J.S. (wsanders@umich.edu), and requests for materials to P.D.G. (gingeric@umich.edu).

METHODS

Micro-CT scan imaging. Specimen SGS-UM 2009-002 was scanned on the OMNI-X HD600 High Resolution Computed Tomography scanner in the Center for Quantitative Imaging at Pennsylvania State University. Scans were collected with the X-TEK X-ray tube with the following settings: energy settings 180 kV, 0.2 mA; source to object distance 215.03 mm; scanned with 1,440 views and five samples averaged per view; field of view 65.9 mm; $x = y = 0.0644$ mm (pixel size); $z = 0.069$ mm (slice thickness and spacing); matrix size $1,024 \times 1,024$ 16-bit TIFF images. There were 1,864 slices taken; the specimen was scanned in volume mode with a Feldkamp reconstruction algorithm.

Measurements. Standard craniodental measurements were taken of the holotype specimen of *Saadanius hijazensis* (Supplementary Table 3). All measurements were recorded to the nearest 0.01 mm using either digital callipers or a Leica S6E binocular dissecting microscope fitted with a measuring reticule.

Phylogenetic analyses. The morphological data set on which our phylogenetic analyses were based included 36 characters and 19 taxa. Besides *Saadanius*, other taxa scored included the propliopithecoid *Aegyptopithecus*, the cercopithecoid *Victoriapithecus*, Pliopithecoida (a composite taxon including *Dionysopithecus*, *Platodontopithecus*, *Pliopithecus*, *Epipliopithecus*, *Egarapithecus*, *Plesiopliopithecus*, *Anapithecus*, *Laccopithecus* and *Paidopithecus*)²⁸, the enigmatic *Kamoyapithecus*, the dendropithecids *Dendropithecus*, *Micropithecus* and *Simiolus*, the proconsulids *Afropithecus*, *Heliopithecus*, *Morotopithecus*, *Turkanapithecus*, *Nyanzapithecus*, *Rangwapithecus*, *Nacholapithecus*, *Equatorius* and *Proconsul* (combining two species *P. nyanzae* and *P. heseloni*), and the extant apes *Pan* and *Hylobates*. We also scored dendropithecids and proconsulids as single entities, which caused several polymorphic characters to be recognized. Character states were scored on original specimens housed at the Kenya National Museums, high-quality casts housed at the University of Michigan Department of Anthropology and Museum of Paleontology, or from figures and descriptions provided in the literature.

A data matrix was built with TNT (Tree Analysis Using New Technology)³⁰. Taxa and characters were defined under the Data option in TNT; character states

were then assigned to each taxon on the basis of our scoring. Using the Character Settings option in TNT we were able to define characters as additive (ordered) or non-additive (unordered) and to include or exclude characters to explore different analysis options. Our matrix was then saved as a *.tnt file using the Save Data option.

This data matrix was then analysed with the Implicit Enumeration Algorithm under the Analyse menu in TNT to ensure that all possible trees were explored, both with and without the collapse-tree option (results in each case were identical). Phylogenetic analysis was performed using the basal fossil catarrhine (propliopithecoid) *Aegyptopithecus* to root all trees. All characters were treated as being of equal weight, and two characters were ordered (16, upper canine size and 30, presence or absence of an entepicondylar foramen on the distal humerus; see Supplementary Table 5).

Consistency Index (CI) and Retention Index (RI) were calculated on the basis of minimum (MIN) and maximum (MAX) and most-parsimonious (MP) steps for each resulting tree (using the show max-min option in TNT). $CI = MIN/MP$; $RI = (MAX - MP)/(MAX - MIN)$.

Analyses were performed in four ways: for all taxa using all characters (Supplementary Fig. 3c); for all taxa, excluding postcranial characters (Supplementary Fig. 3d); for appropriate taxa collapsed into Dendropithecidae and Proconsulidae using all characters (Supplementary Fig. 3a); and for collapsed taxa, excluding postcranial characters (Supplementary Fig. 3b). Trees were examined with the View Trees option in TNT, and consensus trees were viewed with the TNT Consensus option in the Trees menu. The TNT Optimize menu was used to view tree statistics (lengths, branch lengths and min-max) and to explore the distribution of different character states by mapping characters and synapomorphies onto trees.

As a final test of the trees produced by TNT, we saved our *.tnt file as a NEXUS file and analysed that matrix using PAUP with the Branch-and-Bound option to ensure finding the shortest trees. Results were identical to those produced in TNT.

Replacing underperforming protected areas achieves better conservation outcomes

Richard A. Fuller^{1,2}, Eve McDonald-Madden^{1,2}, Kerrie A. Wilson¹, Josie Carwardine^{1,2}, Hedley S. Grantham¹, James E. M. Watson¹, Carissa J. Klein¹, David C. Green³ & Hugh P. Possingham¹

Protected areas vary enormously in their contribution to conserving biodiversity, and the inefficiency of protected area systems is widely acknowledged^{1–3}. However, conservation plans focus overwhelmingly on adding new sites to current protected area estates⁴. Here we show that the conservation performance of a protected area system can be radically improved, without extra expenditure, by replacing a small number of protected areas with new ones that achieve more for conservation. Replacing the least cost-effective 1% of Australia's 6,990 strictly protected areas could increase the number of vegetation types that have 15% or more of their original extent protected from 18 to 54, of a maximum possible of 58. Moreover, it increases markedly the area that can be protected, with no increase in overall spending. This new paradigm for protected area system expansion could yield huge improvements to global conservation at a time when competition for land is increasingly intense.

Protected areas are one of the most important tools in modern conservation, with over 100,000 sites covering about 12% of the area of countries and their territorial waters worldwide⁵. Historically, the placement of protected areas has been driven more by a lack of potential for economic development in an area than its contribution to conservation goals⁶. Consequently, many species and habitats remain inadequately protected and vulnerable to threatening processes. For example, only 0.01% of the global extent of coral reefs occurs within effective protected areas⁷, the distributions of 20% of the world's threatened bird species do not overlap at all with protected areas¹, and 83% of threatened plants in New Caledonia are found only outside protected areas⁸.

Schemes for improving the performance of protected area systems typically begin by determining whether biodiversity features (for example, ecosystems, species) are adequately represented in the existing system, and then go on to identify further areas to fill the gaps^{1,3,9–11}. This approach means that resources will continue to be expended on managing poor quality sites within the existing protected area system, and large amounts of financial capital will remain tied up. Clearly, achieving a given level of biodiversity protection by adding to an inefficient system is ultimately much more expensive than one efficiently planned from the outset by comparing the costs and benefits of candidate protected areas^{12,13}.

A more radical approach to expanding protected area systems would be to reverse the protection status of the least cost-effective sites and use the resulting capital to establish and manage new protected areas. Whether this is worth doing depends on the magnitude of the potential gains for conservation. Here we measure the improvement in protected area system performance that could be delivered by replacing a small number of sites that contribute the least for conservation given the capital they absorb. We illustrate this

approach for Australia, a country that spans ecosystems from tropical lowland rainforest to subalpine meadows, arid scrub and desert. It has among the world's most systematically designed protected area systems¹⁴, and therefore one would expect the benefits of protected area replacement to be small relative to many other regions around the world.

The protected area estate in Australia comprised 6,990 sites covering 629,352 km² (~8% of Australia's landmass) managed under the International Union for Conservation of Nature (IUCN) management categories I–IV in 2006¹⁵. The distributions of 60 vegetation types in 1750 (before widespread clearance by Europeans) and in 2000 are mapped by the National Vegetation Information System (Version 3.0; available at <http://www.environment.gov.au>). We used a conservation target of 15% of the pre-clearance extent of each vegetation type to measure the performance of the system, although this could be substituted by, or augmented with, any desired performance metric including those reflecting cultural values, human recreation or other needs.

We superimposed the 6,990 protected areas onto 62,630 subcatchments (mean area = 123 km²) mapped across Australia¹⁶, resulting in 64,991 distinct planning units for which we had cost and vegetation mapping information. We estimated the value of all land covered by native vegetation within each planning unit using data on average unimproved land values at the scale of local government areas. These values represented the purchase price or the funds realized when degazetting (removing protection and selling) the land occupied by all areas of native vegetation in each planning unit. We also applied a fixed area-based transaction cost levied when either acquiring or degazetting land (Supplementary Information). Our cost estimates assume that land has not had value added by virtue of being in or near a protected area. This has a conservative effect on our analysis, as there are examples of protected areas positively affecting the value of adjacent land¹⁷.

The cost effectiveness of each planning unit, E_j , is derived from its contribution to conserving vegetation types within Australia (the benefit), B_j , relative to the cost of protecting that planning unit, C_j : $E_j = \frac{B_j}{C_j}$. Here, $B_j = \sum_{i=1}^m A_{ij}/r_i$, where m is the total number of vegetation types, A_{ij} is the area covered by vegetation type i in planning unit j and r_i is the proportion of each vegetation type i remaining in Australia from pre-clearance levels. The cost, C_j , is the estimated land value for each planning unit. We calculated E_j for each planning unit, and sequentially removed the least cost-effective protected areas, replacing them progressively with the most cost-effective new sites until the funds yielded by degazettement were exhausted. Our measure of cost effectiveness prioritizes planning units containing vegetation types that have been heavily cleared.

¹The Ecology Centre, University of Queensland, St Lucia, Queensland 4072, Australia. ²CSIRO Climate Adaptation Flagship and CSIRO Sustainable Ecosystems, St Lucia, Queensland 4072, Australia. ³Information Technology Services, University of Queensland, St Lucia, Queensland 4072, Australia.

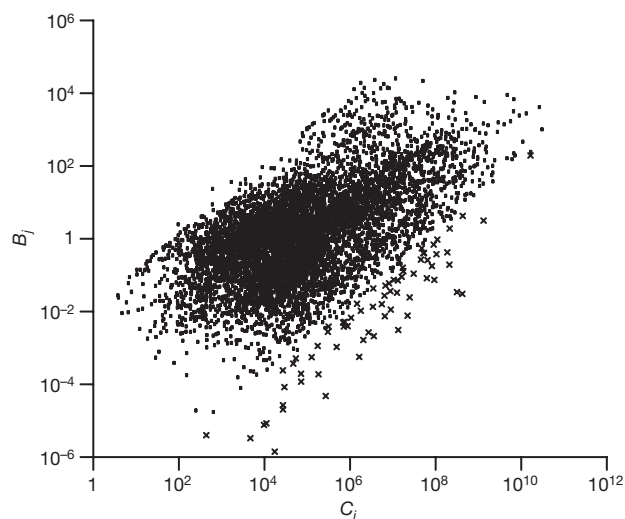


Figure 1 | Cost effectiveness in Australia's protected areas. The contribution of Australian protected areas to conserving vegetation types relative to their rarity (B_j) is positively related to the estimated cost of acquisition and management of the sites (C_j). However, there is a great deal of scatter in the cost effectiveness among the 6,990 protected areas; here the least cost-effective 1% of sites (70 protected areas) are denoted by crosses.

Being a static measure that is calculated once for each planning unit at the start of the analysis, it continues to prioritize threatened vegetation types without specifying a threshold for adequate protection, and it does not incorporate complementarity¹⁸. It is thus conservative for this application, because a dynamic prioritization algorithm would be even more efficient¹⁹. Furthermore, our static cost-effectiveness metric is transparent, simple to interpret and easy to implement.

The relationship between the estimated value of the land occupied by a protected area (C_j) and its contribution to protecting heavily cleared vegetation types (B_j) was strongly positive (standardized major axis regression: $\beta = 0.85$; $r^2 = 0.34$; $P < 0.001$), driven by the fact that both cost and benefit increase with the size of the protected area (Fig. 1). However, there is substantial variation in cost effectiveness among current protected areas. The least cost-effective protected areas span a wide range of land values and benefit values, and it is worth noting that many of the most expensive sites are still cost effective (Fig. 1).

Replacing the least cost-effective protected areas with the most cost-effective new sites produces a very rapid increase in system performance (Fig. 2). By replacing only the lowest ranked 1% of protected areas we can increase the number of vegetation types for which 15% of their pre-clearance extent is conserved in Australia from 18 to 54, of a maximum possible of 58. Two of the original 60 vegetation types (brigalow *Acacia harpophylla* forests and woodlands; mallee with a

tussock grass understorey) have less than 15% of their original extent remaining and thus the reservation target cannot be met. Moreover, by reconfiguring the protected area system we can increase the total area under protection by a factor of ten. We repeated the analysis assuming that only 80% of the estimated value of a land parcel is realized when sold, and 120% of its value is required for purchase. The results were nearly as marked, with vegetation types protected to 15% again rising from 18 to 54, and the total area under protection increasing by a factor of 9.8 (Supplementary Information). A dynamic version of this analysis that aims to equalize protection levels across vegetation types produces slightly stronger results, with vegetation types protected to 15% rising from 18 to 55, and the total area under protection increasing by a factor of ten (Supplementary Information). We emphasize that all these performance gains are budget neutral, achieved using only the funds realized from degazettement.

Enormous efficiency gains could be achieved by modest and careful adjustments to a protected area system. Although we do not advocate wholesale reassembly of protected area systems that have been built up over decades, our results show that a more flexible approach to the expansion of protected area systems could ultimately protect much more biodiversity. With increasing pressure on environmental budgets, demonstrating value for money is more urgent than ever, and protected area replacement produces demonstrable gains in representing biodiversity.

There are many possible cost-effectiveness metrics that could be used to construct this kind of analysis, reflecting the full range of reasons why protected areas have been designated and are valued by society. Whatever metrics are used, wide variation among protected areas in their cost and contribution to system performance will inevitably result in large gains to performance by some judicious changes. For example, a cost-effectiveness metric based on human amenity value using the number of people living within 50 km of protected areas yields equally strong efficiency gains (Supplementary Information). However, as more objectives are added, the performance of the protected area system in relation to any individual objective will inevitably decline.

The idea of degazettement usually has negative connotations, and it is highly controversial among conservation scientists and practitioners^{20–22}. Yet designing processes to deal with protected area replacement is essentially no different to solving other complex environmental management questions where costs and benefits are spread unequally across society and among generations²³. Formalizing the process using biological criteria might even improve accountability, particularly if replacement with a higher quality site is formally mandated. For example, instead of such decisions being framed as giving up on protection, the focus could be on allowing fair economic use of downlisted sites while insisting on protection for new, more cost-effective sites. The reality is that excision of commercially valuable parts of protected areas, and degazettement or downgrading of entire sites, are becoming increasingly common in many

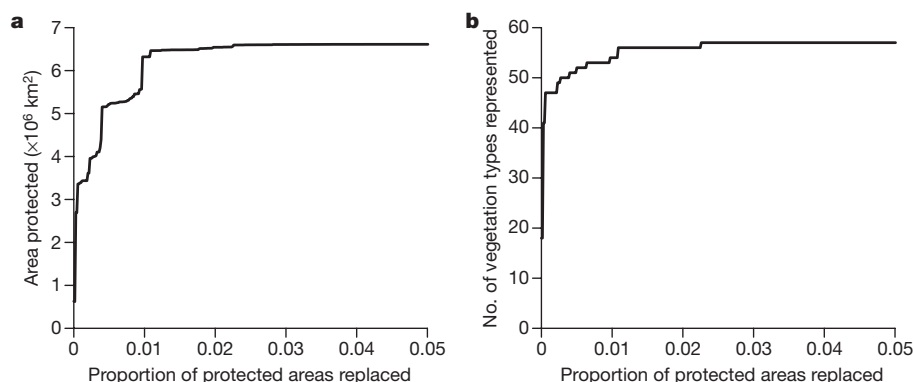


Figure 2 | Conservation outcomes delivered by protected area replacement. **a**, Change in total area under protection. **b**, The number of vegetation types for which at least 15% of their pre-clearance extent are represented, as existing protected areas are progressively replaced with more efficient sites.

countries, often with no form of mitigation^{24–26}. This is especially likely where the opportunity costs of protected areas are greater than the revenue they generate, and is currently a particularly acute problem in east Africa²⁷.

Relaxing the protection status of any existing protected area will probably prove politically difficult. As an alternative to protected area replacement, investment in the ongoing management of a protected area could be scaled up or down according to its cost effectiveness. This does not necessarily mean abandoning sites with low conservation value. Providing management costs are low it might be worth continuing to manage such a site. In some cases, and particularly where a site is near an urban area, alternative commercial sources of funds for management could be sought, for example, revenue from tourism or recreational users. This would enable biodiversity-conservation-specific funds to be directed towards areas with the greatest biological value.

Growth in global spending on protected areas has stagnated in recent decades²⁸, and the rate of gazetting new protected areas is slowing as countries begin to meet their obligations under the Convention on Biological Diversity. If the rate of new investment in protected areas continues to decline, ensuring that the best places are retained in the face of rampant land clearance will be more important than ever. Protected areas form a large part of our conservation legacy to future generations, and handing down a healthy and well-reasoned set of sites is, in our view, the best way of justifying this investment.

Received 17 February; accepted 18 May 2010.

Published online 30 June 2010.

- Rodrigues, A. S. L. *et al.* Effectiveness of the global protected area network in representing species diversity. *Nature* **428**, 640–643 (2004).
- Wiersma, Y. F. & Nudds, T. D. Efficiency and effectiveness in representative reserve design in Canada: the contribution of existing protected areas. *Biol. Conserv.* **142**, 1639–1646 (2009).
- Ceballos, G. Conservation priorities for mammals in megadiverse Mexico: the efficiency of reserve networks. *Ecol. Appl.* **17**, 569–578 (2007).
- Scott, J. M. *et al.* Gap analysis: a geographic approach to protection of biological diversity. *Wildl. Monogr.* **123**, 1–41 (1993).
- World Database on Protected Areas. (<http://www.wdpa.org/>) (2009).
- Ando, A., Camm, J., Polasky, S. & Solow, A. Species distributions, land values, and efficient conservation. *Science* **279**, 2126–2128 (1998).
- Mora, C. *et al.* Coral reefs and the global network of marine protected areas. *Science* **312**, 1750–1751 (2006).
- Jaffre, T., Bouchet, P. & Veillon, J.-M. Threatened plants of New Caledonia: is the system of protected areas adequate? *Biodivers. Conserv.* **7**, 109–135 (1998).
- Rodrigues, A. S. L. *et al.* Global gap analysis: priority regions for expanding the global protected-area network. *Bioscience* **54**, 1092–1100 (2004).
- Carwardine, J. *et al.* Cost-effective priorities for global mammal conservation. *Proc. Natl Acad. Sci. USA* **105**, 11446–11450 (2008).
- Underwood, E. C. *et al.* Expanding the global network of protected areas to save the imperiled Mediterranean biome. *Conserv. Biol.* **23**, 43–52 (2009).
- Naidoo, R. *et al.* Integrating economic costs into conservation planning. *Trends Ecol. Evol.* **21**, 681–687 (2006).
- Strange, N., Thorsen, B. J. & Bladt, J. Optimal reserve selection in a dynamic world. *Biol. Conserv.* **131**, 33–41 (2006).
- Department of the Environment, Water, Heritage and the Arts; Australian Government. *Australia's Strategy for the National Reserve System 2009–2030* (<http://www.environment.gov.au/parks/publications/nrs/nrsstrat.html>) (2009).
- Department of the Environment, Water, Heritage and the Arts; Australian Government. *Collaborative Australia Protected Area Database* (<http://www.environment.gov.au/parks/nrs/science/capad/2006/index.html>) (2009).
- Stein, J. in *Linking Rivers to Landscapes* (eds Rutherford, I., Wiszniewski, I., Askey-Doran, M. & Glazik, R.) 448–552 (Department of Primary Industries, Water and Environment, 2005).
- Roberts, C. M., Bohnsack, J. A., Gell, F., Hawkins, J. P. & Goodridge, R. Effects of marine reserves on adjacent fisheries. *Science* **294**, 1920–1923 (2001).
- Vane-Wright, R. I., Humphries, C. J. & Williams, P. H. What to protect?—Systematics and the agony of choice. *Biol. Conserv.* **55**, 235–254 (1991).
- Meir, E., Andelman, S. & Possingham, H. P. Does conservation planning matter in a dynamic and uncertain world? *Ecol. Lett.* **7**, 615–622 (2004).
- Vandergeest, P. Property rights in protected areas: obstacles to community involvement as a solution in Thailand. *Environ. Conserv.* **23**, 259–268 (1996).
- Dearden, P., Chettamart, S. & Emphandu, D. Protected areas and property rights in Thailand. *Environ. Conserv.* **25**, 195–197 (1998).
- Vandergeest, P. Reply: Protected areas and property rights in Thailand. *Environ. Conserv.* **26**, 7–9 (1999).
- Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
- Veit, P. G., Nshala, R., Ochieng' Odhiambo, M. & Manyindo, J. *Protected Areas and Property Rights: Democratizing Eminent Domain in East Africa* (World Resources Institute, 2008).
- Child, K. Civil society in Uganda: the struggle to save the Mabira Forest Reserve. *J. East. Afr. Stud.* **3**, 240–258 (2009).
- Sussman, R. W., Green, G. M. & Sussman, L. K. Satellite imagery, human ecology, anthropology, and deforestation in Madagascar. *Hum. Ecol.* **22**, 333–354 (1994).
- Norton-Griffiths, M. & Southey, C. The opportunity costs of biodiversity conservation in Kenya. *Ecol. Econ.* **12**, 125–139 (1995).
- Emerton, L., Bishop, J. & Thomas, L. *Sustainable Financing of Protected Areas: A Global Review of Challenges and Options* (IUCN, 2006).

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

Acknowledgements We thank J. Stein for providing subcatchment data, and L. Barr, C. Fuller, B. Kendall, T. Martin and H. Wilson for discussion. This work was funded by the Centre for Applied Environmental Decision Analysis, an Australian Commonwealth Environment Research Facility.

Author Contributions All authors designed the research. E.M.-M., D.C.G. and R.A.F. performed the analysis, and R.A.F. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.A.F. (r.a.fuller@dunelm.org.uk).

LETTERS

Functional impact of global rare copy number variation in autism spectrum disorders

A list of authors and their affiliations appears at the end of the paper.

The autism spectrum disorders (ASDs) are a group of conditions characterized by impairments in reciprocal social interaction and communication, and the presence of restricted and repetitive behaviours¹. Individuals with an ASD vary greatly in cognitive development, which can range from above average to intellectual disability². Although ASDs are known to be highly heritable (~90%)³, the underlying genetic determinants are still largely unknown. Here we analysed the genome-wide characteristics of rare (<1% frequency) copy number variation in ASD using dense genotyping arrays. When comparing 996 ASD individuals of European ancestry to 1,287 matched controls, cases were found to carry a higher global burden of rare, genic copy number variants (CNVs) (1.19 fold, $P = 0.012$), especially so for loci previously implicated in either ASD and/or intellectual disability (1.69 fold, $P = 3.4 \times 10^{-4}$). Among the CNVs there were numerous *de novo* and inherited events, sometimes in combination in a given family, implicating many novel ASD genes such as *SHANK2*, *SYNGAPI*, *DLGAP2* and the X-linked *DDX53-PTCHD1* locus. We also discovered an enrichment of CNVs disrupting functional gene sets involved in cellular proliferation, projection and motility, and GTPase/Ras signalling. Our results reveal many new genetic and functional targets in ASD that may lead to final connected pathways.

Twin and family studies indicate a predominantly genetic basis for ASD susceptibility and provide support for considering these disorders as a clinical spectrum. Some 5–15% of individuals with an ASD have an identifiable genetic aetiology corresponding to known rare single-gene disorders (for example, fragile X syndrome) and chromosomal rearrangements (for example, maternal duplication of 15q11-q13). Rare mutations have been identified in synaptic genes, including *NLGN3*, *NLGN4X* (ref. 4) and *SHANK3* (ref. 5), and micro-array studies have revealed copy number variation (CNV) as risk factors⁶. CNV examples include *de novo* events observed in 5–10% of ASD cases^{7–9}, *de novo* or inherited hemizygous deletions and duplications of 16p11.2 (refs 9–11) and *NRXN1* (ref. 7), and exceptionally rare homozygous deletions in consanguineous families¹². Genome-wide association studies using single nucleotide polymorphisms (SNPs) have highlighted two potential ASD risk loci at 5p14.1 (ref. 13) and 5p15.2 (ref. 14), but these data indicate that common variation will account for only a small proportion of the heritability in ASD.

To delineate further the contribution of rare genomic variants to autism we genotyped 1,275 ASD cases and their parents using the Illumina Infinium 1M single SNP microarray (Fig. 1). A set of 1,981 controls used for comparison studies was genotyped on the same platform¹⁵ and both data sets were subjected to the same quality control procedures. Ultimately, we analysed 996 ASD cases (876 trios) and 1,287 controls of European ancestry to minimize confounds due to population differences (Supplementary Figs 1 and 2 and Supplementary Table 1)¹⁶.

Two CNV prediction algorithms (QuantiSNP¹⁷ and iPattern (unpublished data)) and additional extensive quality control procedures were used to establish a stringent data set of non-redundant CNVs

called by both algorithms in an individual (Fig. 1, Supplementary Tables 1–3 and Supplementary Fig. 3). This stringent data set of 5,478 rare CNVs in 996 cases and 1,287 controls of European ancestry (Supplementary Table 4) had the following characteristics: (1) CNV present at <1% frequency in the total sample (cases and controls); (2) CNV ≥ 30 kb in size (because >95% of these could be confirmed); and (3) all CNVs further verified using combined evidence from the PennCNV algorithm¹⁸ and child–parent intensity fold changes, genotype proportions (to verify deletions) and visual inspection (for chromosome X).

We assessed the impact of rare CNV in cases compared to controls using three primary measures of CNV burden: the number of CNVs per individual, the estimated CNV size, and the number of genes affected by CNVs (Table 1). No significant difference was found in the former two

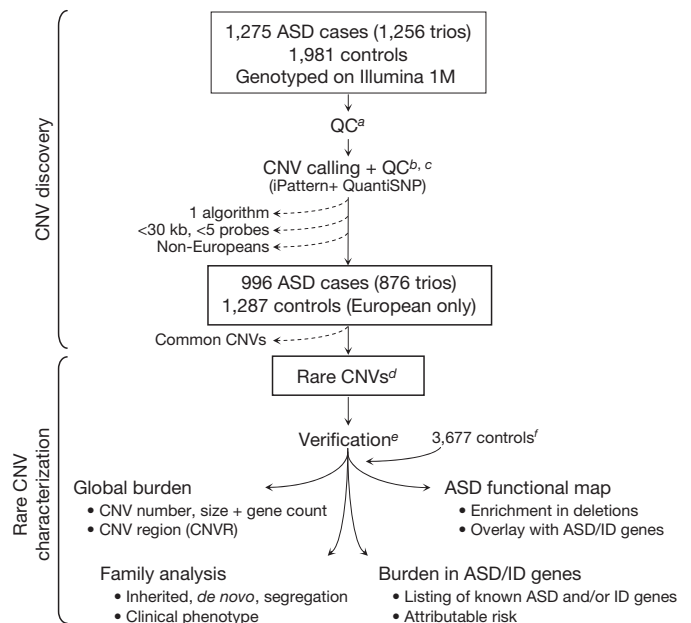


Figure 1 | CNV discovery and characterization. Comprehensive procedures were used to identify the rare CNV data set (boxed). Dashed arrows indicate CNVs not included in downstream analyses. Labels a–f are as follows: a, SNP and intensity quality control (QC) with ancestry estimation; b, QC for CNV calls; c, pilot validation experiments using quantitative PCR were used to evaluate the false discovery rate; d, rare CNVs in samples of European ancestry were defined as ≥ 30 kb in size and present in the total sample set at a frequency <1%. A total of 70 out of 996 (17%) of ASD cases were analysed on different lower-resolution arrays in previous studies^{9,10,28}. Label e indicates that all CNVs were computationally verified and at least 40% of case CNVs were also experimentally validated by qPCR and/or independent Agilent or other SNP microarrays; f, 3,677 additional European ancestry controls were used to test specific loci from the primary burden analyses. Additional details are in the Methods and Supplementary Information. ID, intellectual disability.

Table 1 | Global burden of genic rare CNVs in cases versus controls

Type	Classification	Total CNVs (n)	P	Case/control ratio	Baseline rate (ctrl)	P _{corr}
All	None					
	All	5,478	0.012*	1.19	3.59	0.003*
	Deletions only	2,757	0.008*	1.26	1.08	0.006*
Duplications only	All	2,721	0.072	1.16	2.51	0.015*
All	CNV frequency					
	2–6×	1,831	0.058	1.26	1.03	0.062
	1×	1,419	0.375	1.03	0.83	0.180
Deletions only	2–6×	1,094	0.004*	1.57	0.43	0.011*
	1×	880	0.036*	1.26	0.30	0.036*
Duplications only	2–6×	977	0.203	1.16	1.05	0.118
	1×	896	0.749	0.92	0.72	0.442
All	CNV size					
	30–500 kb	5,086	0.313	1.03	2.72	0.181
	≥500 kb	392	0.005*	1.69	0.88	0.005*
Deletions only	30–500 kb	2,645	0.004*	1.24	0.85	0.004*
	≥500 kb	112	0.209	1.32	0.23	0.143
Duplications only	30–500 kb	2,441	0.801	0.93	1.86	0.518
	≥500 kb	280	0.007*	1.82	0.65	0.012*

Table 1 shows gene count in cases versus controls. Total burden for genes intersected by CNVs in ASD cases ($n = 996$) compared to controls ($n = 1,287$). Gene coordinates were defined by the RefSeq boundaries plus a 10-kb region on either side. All genomic analyses used NCBI Build 36. Genome-wide P -values were estimated in 100,000 permutations (one-sided) and additionally corrected (P_{corr}) for global case–control differences in CNV rate and size. Analyses were further stratified according to CNV type ('deletions only' and 'duplications only') and frequency (single occurrences (1×) or CNVs observed two to six times in the total sample (2–6×)). Note that for the two lower-frequency groups, the 'deletions only' and 'duplications only' counts are not expected to sum to the 'all' count (see ref. 26 and Supplementary Information).

* Significant differences ($P \leq 0.05$) are indicated.

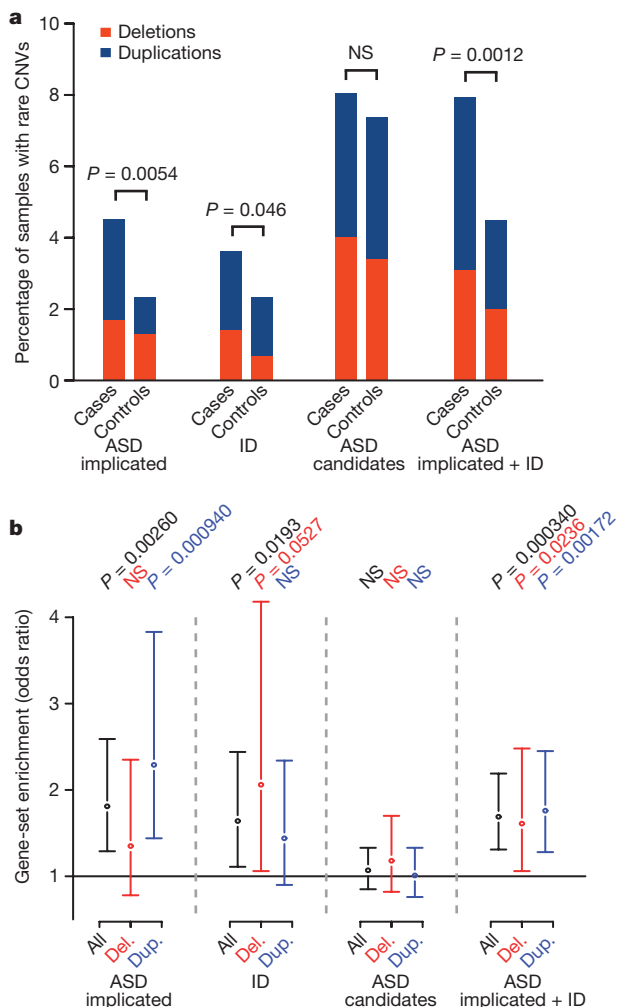
measures (Supplementary Tables 4a and 5), even after controlling for fine-level ancestry differences by pair-matching cases and controls (Supplementary Information)¹⁶. In contrast, we discovered a significant increase in the number of genes intersected by rare CNV in cases when focusing on gene-containing segments (1.19-fold increase, empirical

$P = 0.012$). This ASD association with genic CNV was stronger for deletions (1.26-fold increase, empirical $P = 8.0 \times 10^{-3}$). These differences remained after we further controlled for potential case–control differences that could be present due to biological differences or technical biases. Restricting our analysis to autosomal CNVs (that is, after removing CNVs located on chromosome X) also resulted in a consistent enriched gene count in ASD cases compared to controls. Single-occurrence CNV deletions had increased rates in ASD cases over controls, indicating that some could be pathogenic.

We then examined parent–child transmission and confirmed that 5.7% (50 out of 876) of ASD cases had at least one *de novo* CNV with $>0.6\%$ carrying two or more *de novo* events (Supplementary Tables 4a, 6 and 7). The *de novo* CNV rate in our simplex and multiplex families was 5.6% (22 out of 393) and 5.5% (19 out of 348), respectively, in contrast with previous studies showing a higher rate in simplex families^{8,9}. A total of 226 validated *de novo* (7) and inherited (219) CNVs not observed in controls and affecting single genes were found (Supplementary Table 8).

Numerous novel candidate ASD loci such as *SHANK2*, *SYNGAP1* and *DLGAP2* were identified on the basis of the observation that *de novo* CNV affects these genes in cases but not controls (Supplementary Table 6). The relatedness of *SHANK2* to the causal ASD gene *SHANK3* (ref. 5), involvement of *SYNGAP1* in intellectual disability¹⁹, and interaction of *DLGAP* family proteins with *SHANK* proteins²⁰ further support their role in ASDs. Maternally inherited X-linked deletions at *DDX53–PTCHD1* (7 cases) implicate this locus in ASD. We tested an additional 3,677 European ancestry controls (Fig. 1) and again found no CNV at these genes, and *DDX53–PTCHD1* emerged as a significant ASD risk factor ($P = 3.1 \times 10^{-3}$ with the initial 1,287 controls; $P = 3.6 \times 10^{-6}$ with combined controls; Supplementary Fig. 4).

Figure 2 | CNV burden in known ASD and/or intellectual disability genes. **a**, Proportion of samples with CNVs overlapping genes and loci known to be associated in ASD with or without intellectual disability (ID) or intellectual disability only, as well as published candidate genes and loci for ASD (Supplementary Table 9). To select for CNVs with maximal impact, they needed to intersect genes and overlap the target loci by $\geq 50\%$ of their length. Fisher's exact test P -values for significant differences ($P \leq 0.05$, one tailed) are shown. NS, not significant. **b**, Enrichment analysis for genes overlapped by rare CNVs in cases compared to controls for the three gene sets in **a**, relative to the whole genome. Odds ratio and 95% confidence intervals are given for each gene set. Empirical P -values for gene-set enrichment are indicated above each odds ratio. All P -values < 0.1 are listed.



Association studies of individual rare CNV often have insufficient power to discriminate benign from disease-causing variants. Here, we assessed whether genes and CNVs previously associated with ASD and/or intellectual disability were enriched in cases compared with controls, in order to help identify pathogenic events. We defined three gene lists based on evidence from previous studies of their involvement in ASDs (Supplementary Table 9): (1) 'ASD implicated' list consisting of 36 disease genes and 10 loci strongly implicated in ASD and identified in subjects with ASD or ASD and intellectual disability; (2) 'intellectual disability' consisting of 110 disease genes and 17 loci implicated in intellectual disability but not yet in ASD; and (3) 'ASD candidates' including 103 genes from previous studies of common and rare variants.

We observed a higher proportion of cases with rare CNVs overlapping 'ASD implicated' disease genes compared to controls (4.3% versus 2.3%, Fisher exact test $P = 5.4 \times 10^{-3}$; Fig. 2a), corresponding to a significant enrichment for genes in this set (odds ratio (OR) = 1.8; 95% confidence interval (CI) 1.3–2.6, empirical $P = 2.6 \times 10^{-3}$; Fig. 2b, see also Supplementary Information). This effect was stronger for duplications, which may also disrupt genes (OR = 2.3; 95% CI 1.4–3.8, empirical $P = 9.4 \times 10^{-4}$). Enrichment was also found for rare CNVs overlapping intellectual disability genes, more notably for deletions (OR = 2.1; 95% CI 1.1–4.2, empirical $P = 0.053$). In contrast, there was no evidence of enrichment among case CNVs compared to control CNVs for genes in the ASD candidates set (empirical $P > 0.3$). When the two disease gene sets 'ASD implicated' and 'intellectual disability' were combined, we observed 7.6% of cases with rare CNVs preferentially affecting ASD/intellectual disability genes compared to 4.5% in controls (Fisher exact test $P = 1.2 \times 10^{-3}$; Fig. 2a). The observed enrichments did not change when potential case–control genome-wide differences for CNV rate and size were considered.

Our global analyses of these putative pathogenic loci use subjective boundaries for CNV overlap. Manual inspection of the data yields more accurate results. After eliminating CNVs that are less likely to have an aetiological role (heterozygous CNVs that disrupt autosomal recessive loci, events outside the critical region of overlap of genomic

disorders, X-linked genes in females inherited from non-ASD fathers, duplications inherited from non-ASD parents, and intronic CNVs in *NRXN1*), 25 CNVs remained in the ASD group, compared to only four in the controls ($P = 3.6 \times 10^{-6}$; Supplementary Table 10). Moreover, the latter four CNVs were all duplications at 1q21.1, 16p11.2 or 22q11.2, loci known to exhibit incomplete penetrance and variable expressivity⁶. The population attributable risk provided by the combination of all ASD CNVs that overlap ASDs and/or intellectual disability genes is estimated to be 3.3% (Supplementary Table 11). We also identified rare *de novo* chromosomal abnormalities and large CNVs likely to be aetiological (Supplementary Table 10).

We then tested for functional enrichment of gene sets among those genes affected by CNVs to identify biological processes involved in ASD (Fig. 3). Here, the term gene set refers to groups of genes that share a common function or operate in the same pathway. Such a functional enrichment mapping approach can combine single-gene effects into biologically meaningful groups²¹.

We compiled comprehensive collections of gene sets (Supplementary Table 12) and used the Fisher's exact test to assess which gene sets were more frequently affected by rare CNV events in ASD cases compared to controls. An estimate of the false-discovery rate (FDR) at each gene set was obtained by random permutation of case and control labels (Supplementary Information). To visualize enriched gene sets, overlap scores were used to organize these sets graphically into a functional enrichment map (or network) using Cytoscape²². We identified the 'seed' gene sets for the network at an FDR q -value of 5% and further relaxed the thresholds to 12.5% to better capture the network topology²³.

Using these criteria only deletions were found to be significantly enriched in gene sets in cases over controls (Supplementary Fig. 5), consistent with the global burden results (Table 1). Specifically, 76 gene sets affected by deletions (2.18% of sets tested) were found to be enriched and used to construct a functional map (Fig. 3a and Supplementary Figs 6 and 7). We tested for possible bias, including measures of CNV size and number for cases versus controls per gene set, as well as genome proximity, but no differences were found that

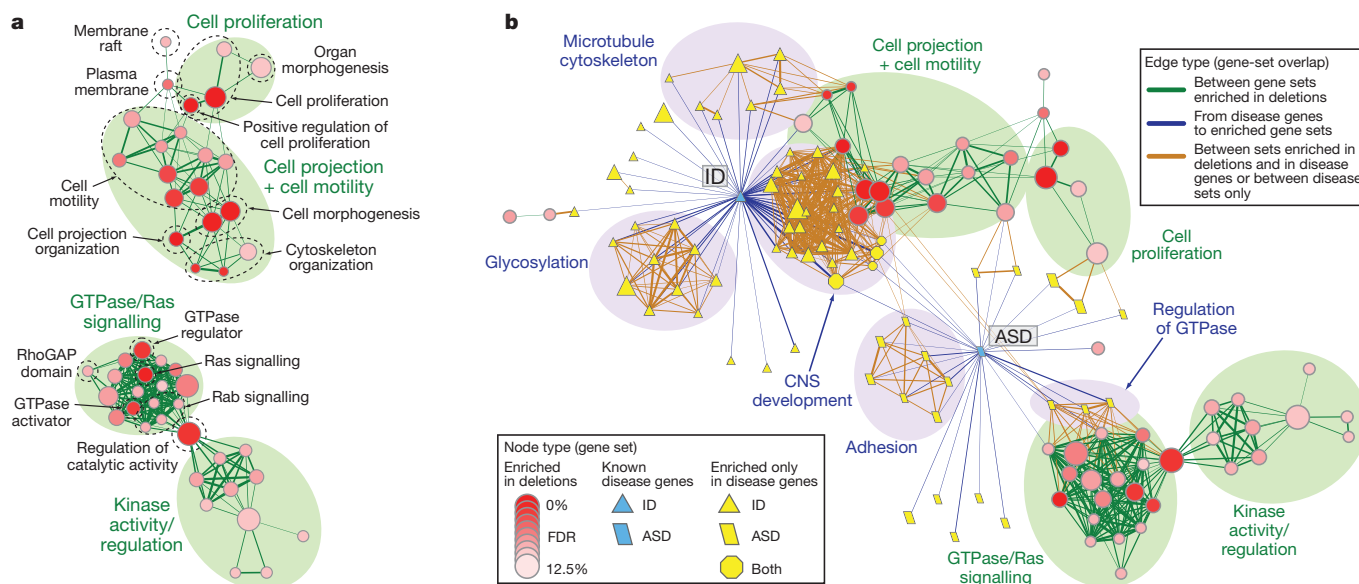


Figure 3 | A functional map of ASD. Enrichment results were mapped as a network of gene sets (nodes) related by mutual overlap (edges), where the colour (red, blue or yellow) indicates the class of gene set. Node size is proportional to the total number of genes in each set and edge thickness represents the number of overlapping genes between sets. **a**, Gene sets enriched for deletions are shown (red) with enrichment significance (FDR q -value) represented as a node colour gradient. Groups of functionally related gene sets are circled and labelled (groups, filled green circles; subgroups, dashed line). **b**, An expanded enrichment map shows the

relationship between gene sets enriched in deletions (**a**) and sets of known ASD/intellectual disability genes. Node colour hue represents the class of gene set (that is, enriched in deletions, red; known disease genes (ASD and/or intellectual disability (ID) genes), blue; enriched only in disease genes, yellow). Edge colour represents the overlap between gene sets enriched in deletions (green), from disease genes to enriched sets (blue), and between sets enriched in deletions and in disease genes or between disease gene-sets only (orange). The major functional groups are highlighted by filled circles (enriched in deletions, green; enriched in ASD/intellectual disability, blue).

might explain the observed enrichments (Supplementary Figs 8 and 9).

We identified enrichments in gene sets known to be involved in ASDs and also discovered new candidate ASD pathways (Fig. 3a and Supplementary Table 13). For example, gene sets involved in cell and neuronal development and function (including projection, motility and proliferation) previously reported in ASD-associated phenotypes were identified²⁴. Novel observations included gene sets involved in GTPase/Ras signalling, with component Rho GTPases known to be involved in regulating dendrite and spine plasticity and associated with intellectual disability. We also found a tentative link to sets in the kinase activity/regulation functional group where only minorities of these sets meet a stringent 5% FDR q -value threshold (Supplementary Fig. 10).

We further assessed the relationship of our functional enrichment map with known ASD/intellectual disability genes (Fig. 3b and Supplementary Fig. 11) and found genes enriched in sets linked to microtubule cytoskeleton, glycosylation and CNS development/adhesion²⁵. The two groups of genes found to be enriched in deletions (Fig. 3a) also displayed connectivity to the ASD/intellectual disability disease gene sets, either directly or through intermediates (Fig. 3b and Supplementary Fig. 12). Although ASD genes seem to be enriched in different subsets of genes compared to intellectual-disability-only genes, we cannot discount the possibility that this is the result of selection bias, and we expect that more intellectual disability genes may yet be linked to ASD.

Our findings provide strong support for the involvement of multiple rare genic CNVs, both genome-wide and at specific loci, in ASD. These findings, similar to those recently described in schizophrenia²⁶, suggest that at least some of these ASD CNVs (and the genes that they affect) are under purifying selection²⁷. Genes previously implicated in ASD by rare variant findings have pointed to functional themes in ASD pathophysiology^{6,28}. Molecules such as NRXN1, NLGN3/4X and SHANK3, localized presynaptically or at the post-synaptic density (PSD), highlight maturation and function of glutamatergic synapses. Our data reveal that *SHANK2*, *SYNGAP1* and *DLGAP2* are new ASD loci that also encode proteins in the PSD. We also found intellectual disability genes to be important in ASD²⁹. Furthermore, our functional enrichment map identifies new groups such as GTPase/Ras, effectively expanding both the number and connectivity of modules that may be involved in ASD. The next step will be to relate defects or patterns of alterations in these groups to ASD endophenotypes. The combined identification of higher-penetrance rare variants and new biological pathways, including those identified in this study, may broaden the targets amenable to genetic testing and therapeutic intervention.

METHODS SUMMARY

Raw data from ASD family (accession phs000267.v1.p1) and SAGE control (Accession: phs000092.v1.p1) genotyping are at NCBI dbGAP. CNVs were analysed using PLINK v1.07³⁰, R stats and custom scripts. See Supplementary Information for details. A list of all CNVs passing quality control is available in Supplementary Table 8.

Received 3 December 2009; accepted 7 May 2010.

Published online 9 June 2010.

1. Veenstra-Vanderweele, J., Christian, S. L. & Cook, E. H. Jr. Autism as a paradigmatic complex genetic disorder. *Annu. Rev. Genomics Hum. Genet.* **5**, 379–405 (2004).
2. Chakrabarti, S. & Fombonne, E. Pervasive developmental disorders in preschool children: confirmation of high prevalence. *Am. J. Psychiatry* **162**, 1133–1141 (2005).
3. Bailey, A. et al. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol. Med.* **25**, 63–77 (1995).
4. Jamain, S. et al. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature Genet.* **34**, 27–29 (2003).
5. Durand, C. M. et al. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature Genet.* **39**, 25–27 (2007).
6. Cook, E. H. Jr & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923 (2008).

7. Szatmari, P. et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genet.* **39**, 319–328 (2007).
8. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
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. Kumar, R. A. et al. Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
12. Morrow, E. M. et al. Identifying autism loci and genes by tracing recent shared ancestry. *Science* **321**, 218–223 (2008).
13. Wang, K. et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
14. Weiss, L. A., Arking, D. E., Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802–808 (2009).
15. Bierut, L. J. et al. A genome-wide association study of alcohol dependence. *Proc. Natl Acad. Sci. USA* **107**, 5082–5087 (2010).
16. Lee, A. B., Luca, D., Klei, L., Devlin, B. & Roeder, K. Discovering genetic ancestry using spectral graph theory. *Genet. Epidemiol.* **34**, 51–59 (2010).
17. Colella, S. et al. 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).
18. 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).
19. Hamdan, F. F. et al. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* **360**, 599–605 (2009).
20. Romorini, S. et al. A functional role of postsynaptic density-95-guanylate kinase-associated protein complex in regulating Shank assembly and stability to synapses. *J. Neurosci.* **24**, 9391–9404 (2004).
21. O'Dushlaine, C. et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol. Psychiatry* doi:10.1038/mp.2010.7 (16 February 2010).
22. Cline, M. S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**, 2366–2382 (2007).
23. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
24. Meechan, D. W., Tucker, E. S., Maynard, T. M. & LaMantia, A. S. Diminished dosage of 22q11 genes disrupts neurogenesis and cortical development in a mouse model of 22q11 deletion/DiGeorge syndrome. *Proc. Natl Acad. Sci. USA* **106**, 16434–16445 (2009).
25. Wegiel, J. et al. The neuropathology of autism: defects of neurogenesis and neuronal migration, and dysplastic changes. *Acta Neuropathol.* doi:10.1007/s00401-010-0655-4 (3 March 2010).
26. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
27. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
28. Glessner, J. T. et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
29. Skuse, D. H. Rethinking the nature of genetic vulnerability to autistic spectrum disorders. *Trends Genet.* **23**, 387–395 (2007).
30. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

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

Acknowledgements The authors acknowledge the families participating in the study and the main funders of the Autism Genome Project Consortium (AGP): Autism Speaks (USA), the Health Research Board (HRB; Ireland), The Medical Research Council (MRC; UK), Genome Canada/Ontario Genomics Institute, and the Hilibrand Foundation (USA). Additional support for individual groups was provided by the US National Institutes of Health (NIH grants HD055751, HD055782, HD055784, HD35465, MH52708, MH55284, MH57881, MH061009, MH06359, MH066673, MH080647, MH081754, MH66766, NS026630, NS042165, NS049261), the Canadian Institute for Advanced Research (CIFAR), the Canadian Institutes for Health Research (CIHR), Assistance Publique-Hôpitaux de Paris (France), Autistica, Canada Foundation for Innovation/Ontario Innovation Trust, Deutsche Forschungsgemeinschaft (grant Po 255/17-4) (Germany), EC Sixth FP AUTISM MOLGEN, Fundação Calouste Gulbenkian (Portugal), Fondation de France, Fondation FondaMental (France), Fondation Orange (France), Fondation pour la Recherche Médicale (France), Fundação para a Ciência e Tecnologia (Portugal), the Hospital for Sick Children Foundation and University of Toronto (Canada), INSERM (France), Institut Pasteur (France), the Italian Ministry of Health (convention 181 of 19.10.2001), the John P. Hussman Foundation (USA), McLaughlin Centre (Canada), Ontario Ministry of Research and Innovation (Canada), the Seaver Foundation (USA), the Swedish Science Council, The Centre for Applied Genomics (Canada), the Utah Autism Foundation (USA) and the Wellcome Trust core award 075491/Z/04 (UK). D.P. is supported by fellowships from the Royal Netherlands Academy of Arts and

Sciences (TMF/DA/5801) and the Netherlands Organization for Scientific Research (Rubicon 825.06.031). S.W.S. holds the GlaxoSmithKline-CIHR Pathfinder Chair in Genetics and Genomics at the University of Toronto and the Hospital for Sick Children (Canada).

Author Contributions D.P., J.D.B., R.M.C., E.H.C., H.C., M.C., B.D., S.E., L.G., D.H.G., M.G., J.L.H., J.H., J.M., A.P.M., J.I.N., A.D.P., M.A.P.-V., G.D.S., P.S., A.M.V., V.J.V., E.M.W., J.S.S., C.B. and S.W.S. were leading contributors in the design, analysis and writing of this study. A.J.B., A.B., G.D., C.M.F., H.H., S.M.K., E.M., S.F.N., G.O., J.P., T.H.W., J.D.B., R.M.C., E.H.C., H.C., B.D., S.E., L.G., D.H.G., M.G., J.L.H., J.H., A.P.M., J.I.N., A.D.P., M.A.P.-V., G.D.S., P.S., A.M.V., V.J.V., E.M.W., S.W.S., J.S.S. and C.B. are Lead Autism Genome Project Consortium (AGP) investigators who contributed equally to this project. All other authors were either involved in phenotype and clinical assessments or have participated in experiments and analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.W.S. (stephen.scherer@sickkids.ca).

Dalila Pinto¹, Alistair T. Pagnamenta², Lambertus Klei³, Richard Anney⁴, Daniele Merico⁵, Regina Regan⁶, Judith Conroy⁶, Tiago R. Magalhaes^{7,8}, Catarina Correia^{7,8}, Brett S. Abrahams⁹, Joana Almeida¹⁰, Elena Bacchelli¹¹, Gary D. Bader^{5,72}, Anthony J. Bailey¹², Gillian Baird¹³, Agatino Battaglia¹⁴, Tom Berney^{15,56}, Nadia Bolshakova⁴, Sven Bölte¹⁶, Patrick F. Bolton¹⁷, Thomas Bourgeron¹⁸, Sean Brennan⁴, Jessica Brian¹⁹, Susan E. Bryson²⁰, Andrew R. Carson¹, Guillermo Casallo¹, Jillian Casey⁶, Brian H.Y. Chung¹, Lynne Cochrane⁴, Christina Corsello²¹, Emily L. Crawford²², Andrew Crossett²³, Cheryl Cytrynbaum¹, Geraldine Dawson^{24,25}, Maretha de Jonge²⁶, Richard Delorme²⁷, Irene Dmric¹⁹, Eftichia Duketis¹⁶, Frederico Duque¹⁰, Annette Estes²⁸, Penny Farrar², Bridget A. Fernandez²⁹, Susan E. Folstein³⁰, Eric Fombonne³¹, Christine M. Freitag¹⁶, John Gilbert³⁰, Christopher Gillberg³², Joseph T. Glessner³³, Jeremy Goldberg³⁴, Andrew Green⁶, Jonathan Green³⁵, Stephen J. Guter³⁶, Hakon Hakonarson^{33,37}, Elizabeth A. Heron⁴, Matthew Hill⁴, Richard Holt², Jennifer L. Howe¹, Gillian Hughes⁴, Vanessa Hus²¹, Roberta Igliozzi¹⁴, Cecilia Kim³³, Sabine M. Klauck³⁸, Alexander Kolevzon³⁹, Olena Korvatska⁴⁰, Vlad Kustanovich⁴¹, Clara M. Lajonchere⁴¹, Janine A. Lamb⁴², Magdalena Laskawiec¹², Marion Leboyer⁴³, Ann Le Couteur^{15,56}, Bennett L. Leventhal^{44,45}, Anath C. Lionel¹, Xiao-Qing Liu¹, Catherine Lord²¹, Linda Lotspeich⁴⁶, Sabata C. Lund²², Elena Maestrini¹¹, William Mahoney⁴⁷, Carine Mantoulan⁴⁸, Christian R. Marshall¹, Helen McConachie^{15,56}, Christopher J. McDougle⁴⁹, Jane McGrath⁴, William M. McMahon⁵⁰, Alison Merikangas⁴, Ohsuke Migita¹, Nancy J. Minshew⁵¹, Ghazala K. Mirza², Jeff Munson⁵², Stanley F. Nelson⁵³, Carolyn Noakes¹⁹, Abdul Noor⁵⁴, Gudrun Nygren⁵², Guiomar Oliveira¹⁰, Katerina Papanikolaou⁵⁵, Jeremy R. Parr⁵⁶, Barbara Parrini¹⁴, Tara Paton¹, Andrew Pickles⁵⁷, Marion Pilorge⁵⁸, Joseph Piven⁵⁹, Chris P. Ponting⁶⁰, David J. Posey⁴⁹, Annemarie Poustka^{38,†}, Fritz Poustka¹⁶, Aparna Prasad¹, Jiannis Ragoussis², Katy Renshaw¹², Jessica Rickaby¹, Wendy Roberts¹⁹, Kathryn Roeder²³, Bernadette Roge⁴⁸, Michael L. Rutter⁶¹, Laura J. Bierut⁶², John P. Rice⁶², Jeff Salt³⁶, Katherine Sansom¹, Daisuke Sato¹, Ricardo Segurado⁴, Ana F. Sequeira^{7,8}, Lili Senman¹⁹, Naisha Shah⁶, Val C. Sheffield⁶³, Latha Soorya³⁹, Inês Sousa², Olaf Stein⁶⁴, Nuala Sykes², Vera Stoppioni⁶⁵, Christina Strawbridge³⁴, Raffaella Tancredi¹⁴, Katherine Tansey⁴, Bhooma Thiruvahindrapuram¹, Ann P. Thompson³⁴, Susanne Thomson²², Ana Tryfou³⁹, John Tsiatis⁵⁵, Herman Van Engeland²⁶, John B. Vincent⁵⁴, Fred Volkmar⁶⁶, Simon Wallace¹², Kai Wang³³, Zhouzhi Wang¹, Thomas H. Wassink⁶⁷, Caleb Webber⁶⁰, Rosanna Weksberg¹, Kirsty Wing², Kerstin Wittmeyer⁴⁸, Shawn Wood³³, Jing Wu²³, Brian L. Yaspan²², Danielle Zurawiecki³⁹, Lonnie Zwaigenbaum⁶⁸, Joseph D. Buxbaum³⁹, Rita M. Cantor⁵³, Edwin H. Cook³⁶, Hilary Coon⁵⁰, Michael L. Cuccaro³⁰, Bernie Devlin³, Sean Ennis⁶, Louise Gallagher⁴, Daniel H. Geschwind⁹, Michael Gill⁴, Jonathan L. Haines⁶⁹, Joachim Hallmayer⁴⁶, Judith Miller⁵⁰, Anthony P. Monaco⁷, John I. Nurnberger Jr.⁴⁹, Andrew D. Paterson¹, Margaret A. Pericak-Vance³⁰, Gerard D. Schellenberg⁷⁰, Peter Szatmari³⁴, Astrid M. Vicente^{7,8}, Veronica J. Vieland⁶⁴, Ellen M. Wijsman⁷¹, Stephen W. Scherer^{1,72}, James S. Sutcliffe²² & Catalina Betancur⁵⁸

¹The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Ontario M5G 1L7, Canada. ²Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. ³Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. ⁴Autism Genetics Group, Department of Psychiatry, School of Medicine, Trinity College, Dublin 8, Ireland. ⁵Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. ⁶School of Medicine and Medical Science University College, Dublin 4, Ireland. ⁷Instituto Nacional de Saude Dr Ricardo Jorge 1649-016 Lisbon and Instituto Gulbenkian de Ciencia, 2780-156 Oeiras, Portugal. ⁸BioFiG—Center for Biodiversity, Functional and Integrative Genomics, Campus da FCUL, C-2.2.12, Campo Grande, 1749-016 Lisboa, Portugal. ⁹Program in Neurogenetics, Department of Neurology and Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine at UCLA, Los Angeles, California 90095, USA. ¹⁰Hospital Pediátrico de Coimbra, 3000 – 076 Coimbra, Portugal. ¹¹Department of Biology, University of Bologna, 40126 Bologna, Italy. ¹²Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford OX3 7JX, UK. ¹³Newcomen Centre, Guy's

Hospital, London SE1 9RT, UK. ¹⁴Stella Maris Institute for Child and Adolescent Neuropsychiatry, 56128 Calambrone (Pisa), Italy. ¹⁵Child and Adolescent Mental Health, University of Newcastle, Sir James Spence Institute, Newcastle upon Tyne NE1 4LP, UK. ¹⁶Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, J.W. Goethe University Frankfurt, 60528 Frankfurt, Germany. ¹⁷Department of Child and Adolescent Psychiatry, Institute of Psychiatry, London SE5 8AF, UK. ¹⁸Human Genetics and Cognitive Functions, Institut Pasteur; University Paris Diderot-Paris 7, CNRS URA 2182, Fondation FondaMental, 75015 Paris, France. ¹⁹Autism Research Unit, The Hospital for Sick Children and Bloorview Kids Rehab, University of Toronto, Toronto, Ontario M5G 1X8, Canada. ²⁰Department of Pediatrics and Psychology, Dalhousie University, Halifax, Nova Scotia B3K 6R8, Canada. ²¹Autism and Communicative Disorders Centre, University of Michigan, Ann Arbor, Michigan 48109-2054, USA. ²²Department of Molecular Physiology and Biophysics, Vanderbilt Kennedy Center, and Centers for Human Genetics Research and Molecular Neuroscience, Vanderbilt University, Nashville, Tennessee 37232, USA. ²³Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. ²⁴Autism Speaks, New York 10016, USA. ²⁵Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-3366, USA. ²⁶Department of Child Psychiatry, University Medical Center, Utrecht 3508 GA, The Netherlands. ²⁷INSERM U 955, Fondation FondaMental, APHP, Hôpital Robert Debré, Child and Adolescent Psychiatry, 75019 Paris, France. ²⁸Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington 98195, USA. ²⁹Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St John's Newfoundland A1B 3V6, Canada. ³⁰The John P. Hussman Institute for Human Genomics, University of Miami, Miami, Florida 33101, USA. ³¹Division of Psychiatry, McGill University, Montreal, Quebec H3A 1A1, Canada. ³²Department of Child and Adolescent Psychiatry, Göteborg University, Göteborg S41345, Sweden. ³³The Center for Applied Genomics, Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. ³⁴Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario L8N 3Z5, Canada. ³⁵Academic Department of Child Psychiatry, Booth Hall of Children's Hospital, Blackley, Manchester M9 7AA, UK. ³⁶Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60612, USA. ³⁷Department of Pediatrics, Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. ³⁸Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. ³⁹The Seaver Autism Center for Research and Treatment, Department of Psychiatry, Mount Sinai School of Medicine, New York 10029, USA. ⁴⁰Department of Medicine, University of Washington, Seattle, Washington 98195, USA. ⁴¹Autism Genetic Resource Exchange, Autism Speaks, Los Angeles, California 90036-4234, USA. ⁴²Centre for Integrated Genomic Medical Research, University of Manchester, Manchester M13 9PT, UK. ⁴³INSERM U995, Department of Psychiatry, Groupe Hospitalier Henri Mondor-Albert Chenevier, AP-HP; University Paris 12, Fondation FondaMental, Créteil 94000, France. ⁴⁴Nathan Kline Institute for Psychiatric Research (NKI), 140 Old Orangeburg Road, Orangeburg, New York 10962, USA. ⁴⁵Department of Child and Adolescent Psychiatry, New York University and NYU Child Study Center, 550 First Avenue, New York, New York 10016, USA. ⁴⁶Department of Psychiatry, Division of Child and Adolescent Psychiatry and Child Development, Stanford University School of Medicine, Stanford, California 94304, USA. ⁴⁷Department of Pediatrics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada. ⁴⁸Centre d'Etudes et de Recherches en Psychopathologie, University de Toulouse Le Mirail, Toulouse 31200, France. ⁴⁹Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. ⁵⁰Psychiatry Department, University of Utah Medical School, Salt Lake City, Utah 84108, USA. ⁵¹Departments of Psychiatry and Neurology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. ⁵²Department of Psychiatry and Behavioural Sciences, University of Washington, Seattle, Washington 98195, USA. ⁵³Department of Human Genetics, University of California—Los Angeles School of Medicine, Los Angeles, California 90095, USA. ⁵⁴Centre for Addiction and Mental Health, Clarke Institute and Department of Psychiatry, University of Toronto, Toronto, Ontario M5G 1X8, Canada. ⁵⁵University Department of Child Psychiatry, Athens University, Medical School, Agia Sophia Children's Hospital, 115 27 Athens, Greece. ⁵⁶Institutes of Neuroscience and Health and Society, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK. ⁵⁷Department of Medicine, School of Epidemiology and Health Science, University of Manchester, Manchester M13 9PT, UK. ⁵⁸INSERM U952 and CNRS UMR 7224 and UPMC Univ Paris 06, Paris 75005, France. ⁵⁹Carolina Institute for Developmental Disabilities, University of North Carolina at Chapel Hill, North Carolina 27599-3366, USA. ⁶⁰MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. ⁶¹Social, Genetic and Developmental Psychiatry Centre, Institute Of Psychiatry, London SE5 8AF, UK. ⁶²Department of Psychiatry, Washington University in St Louis, School of Medicine, St Louis, Missouri 63130, USA. ⁶³Department of Pediatrics and Howard Hughes Medical Institute Carver College of Medicine, University of Iowa, Iowa City, Iowa 52242, USA. ⁶⁴Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital and The Ohio State University, Columbus, Ohio 43205, USA. ⁶⁵Neuropsychiatria Infantile, Ospedale Santa Croce, 61032 Fano, Italy. ⁶⁶Child Study Centre, Yale University, New Haven, Connecticut 06520, USA. ⁶⁷Department of Psychiatry, Carver College of Medicine, Iowa City, Iowa 52242, USA. ⁶⁸Department of Pediatrics, University of Alberta, Edmonton, Alberta T6G 2J3, Canada. ⁶⁹Center for Human Genetics Research, Vanderbilt University Medical Centre, Nashville, Tennessee 37232, USA. ⁷⁰Pathology and Laboratory Medicine, University of Pennsylvania, Pennsylvania 19104, USA. ⁷¹Departments of Biostatistics and Medicine, University of Washington, Seattle, Washington 98195, USA. ⁷²Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A1, Canada.

†Deceased.

Blindsight depends on the lateral geniculate nucleus

Michael C. Schmid¹, Sylwia W. Mrowka¹, Janita Turchi¹, Richard C. Saunders¹, Melanie Wilke¹, Andrew J. Peters¹, Frank Q. Ye² & David A. Leopold^{1,2}

Injury to the primary visual cortex (V1) leads to the loss of visual experience. Nonetheless, careful testing shows that certain visually guided behaviours can persist even in the absence of visual awareness^{1–4}. The neural circuits supporting this phenomenon, which is often termed blindsight, remain uncertain⁴. Here we demonstrate that the thalamic lateral geniculate nucleus (LGN) has a causal role in V1-independent processing of visual information. By comparing functional magnetic resonance imaging (fMRI) and behavioural measures with and without temporary LGN inactivation, we assessed the contribution of the LGN to visual functions of macaque monkeys (*Macaca mulatta*) with chronic V1 lesions. Before LGN inactivation, high-contrast stimuli presented to the lesion-affected visual field (scotoma) produced significant V1-independent fMRI activation in the extrastriate cortical areas V2, V3, V4, V5/middle temporal (MT), fundus of the superior temporal sulcus (FST) and lateral intraparietal area (LIP) and the animals correctly located the stimuli in a detection task. However, following reversible inactivation of the LGN in the V1-lesioned hemisphere, fMRI responses and behavioural detection were abolished. These results demonstrate that direct LGN projections to the extrastriate cortex have a critical functional contribution to blindsight. They suggest a viable pathway to mediate fast detection during normal vision.

We acclimated two adult macaque monkeys with chronic V1 aspiration-lesions (Methods Summary) to the fMRI-testing environment. During the experiments, the animals sat upright in a custom-made

chair placed in a vertical 4.7 T magnetic resonance scanner, and fixated on a small point in the centre of the screen while we recorded the eye position and presented visual stimuli (Supplementary Methods). We established the boundaries of the retinotopically organized visual areas using standard functional mapping methods⁵ (Supplementary Figs 1, 2). The centre of the lesion was located at the representation of the horizontal meridian, and extended several millimetres both dorsally and ventrally into V1 (black area in Fig. 1a, area between red bars in Fig. 1b, black area in Supplementary Fig. 2), corresponding to a visual eccentricity of $\sim 2^\circ$ to $\sim 7^\circ$ of visual angle. Previous work has shown that this type of lesion in V1 does not alter the retinotopic organization assessed with fMRI in several extrastriate areas⁶.

To assess whether the extrastriate cortex could be activated in the absence of V1 input, we presented a small (2° diameter) rotating chequerboard pattern, known to effectively drive responses in the early visual cortex, to the visual field region affected by the lesion (scotoma, Fig. 1c, top). In independent experimental runs, we presented the stimulus to the corresponding region in the healthy hemisphere as a control (Fig. 1c, bottom). As expected, the stimulus shown to the unaffected hemifield elicited strong, circumscribed, contralateral responses in V1, neighbouring extrastriate areas V2, V3, V4, V5/MT and FST (Fig. 2a, d), and parietal area LIP (Supplementary Fig. 3a, c). When the stimulus was presented inside the scotoma region, there was no V1 response because the cortex had been removed; nonetheless, there were stimulus-driven responses in

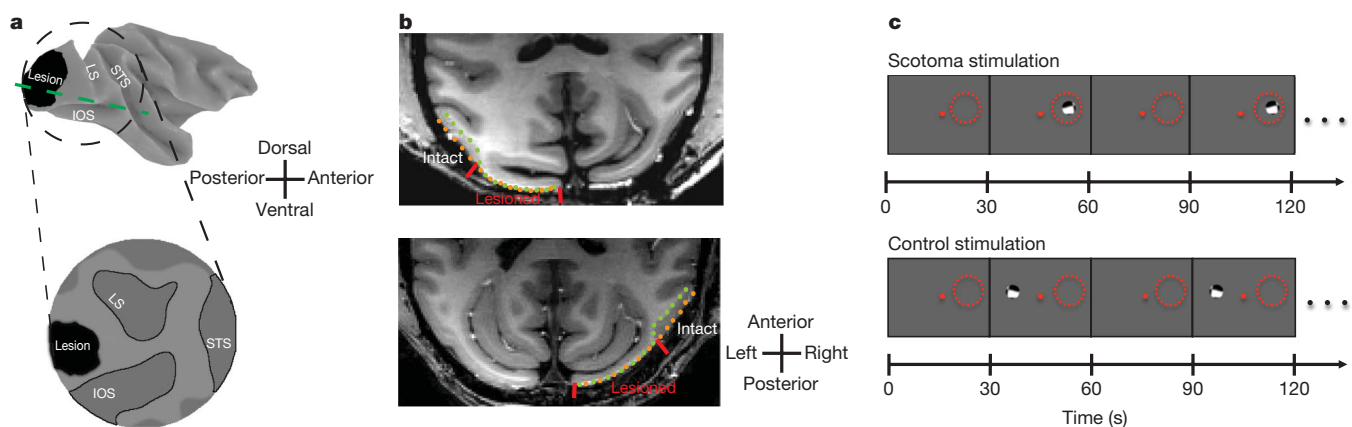


Figure 1 | Experimental set-up. **a**, A side view on the right hemisphere of an inflated macaque brain. An area of $\sim 400 \text{ mm}^2$ of grey matter in the opercular part of V1, representing the visual field between $\sim 2^\circ$ and 7° , has been surgically aspirated and is shown in black. Extrastriate areas, the subject of analysis in this study, are hidden in the sulci surrounding V1, including the lunate (LS), inferior occipital (IOS) and superior temporal (STS) sulci. To facilitate the visual examination of extrastriate cortex, the occipital lobe (dashed circled area) was cut and flattened (inset). **b**, Axial sections of monkey 1's (top) and monkey 2's (bottom) occipital lobes at the

position indicated by the green dashed line in **a**. The outward borders of white and grey matter are highlighted by green and orange dotted lines, respectively. The lesions are evident by the absence of grey matter (area between red line markers). **c**, To compare visually elicited responses in the extrastriate cortex in the presence of V1 input to those in its absence, we spatially restricted (2° diameter) rotating chequerboard stimuli and presented them either inside (top) or outside (bottom) the scotoma (part of the visual field affected by the V1 lesion, indicated here by red dotted circles; the single red dot indicates the fixation point for the animal).

¹Laboratory of Neuropsychology, National Institute of Mental Health (NIMH), 49 Convent Drive, Bethesda, Maryland 20892, USA. ²Neurophysiology Imaging Facility, NIMH, National Institute of Neurological Disorders and Stroke (NINDS), National Eye Institute (NEI), 49 Convent Drive, Bethesda, Maryland 20892, USA.

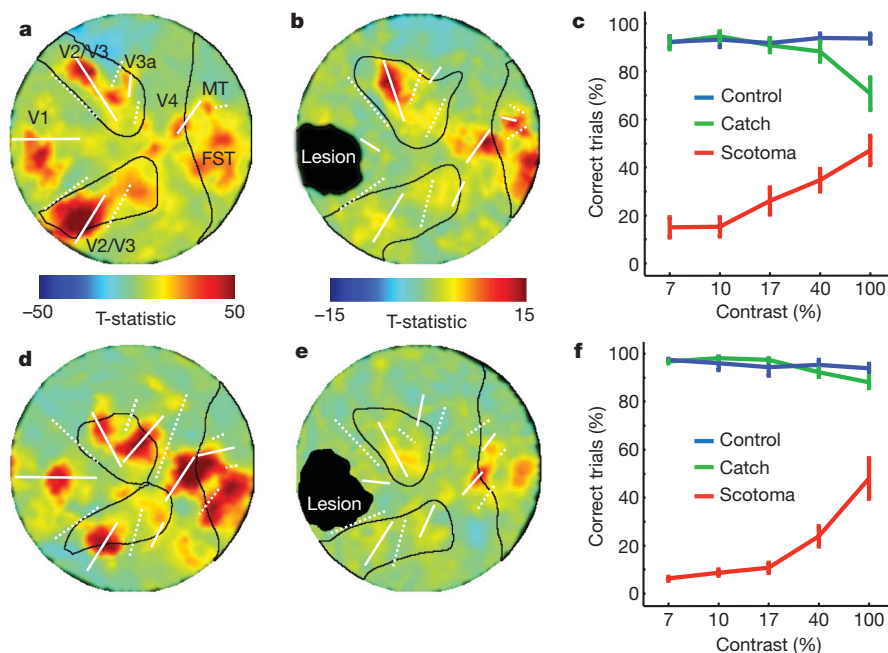


Figure 2 | Visual processing in V1-lesioned monkeys. **a**, Functional activation map (t -statistic) of monkey 1's non-lesioned visual cortex to 85 cycles of visual stimulation outside the scotoma (see Fig. 1c, bottom). The map has been horizontally flipped for easier comparison with the lesioned hemisphere. White dotted and solid lines show the position of the vertical and horizontal meridian representations, respectively, derived from independent retinotopic mapping experiments (Supplementary Figs 1, 2) to reveal the functional boundaries of extrastriate areas^{5,6}. **b**, Activation map of monkey 1's lesioned hemisphere to 85 cycles of visual stimulation inside the scotoma (see Fig. 1c, bottom). The position of the stimulus inside the scotoma was effective in that the lesion surrounding the V1 cortex with intact grey matter was not activated. In the absence of V1 input, areas V2, V3, V4 and V5/MT continue to be visually responsive. **c**, Behavioural performance of monkey 1 in detecting visual stimuli (0.2° diameter)

extrastriate areas V2, V3, V4, V5/MT, FST and LIP (Fig. 2b, e and Supplementary Fig. 3b, d), indicating that stimulus information reached these areas in the absence of V1 input. Moreover, comparing the activation patterns of the lesioned and control hemispheres revealed that the responses within each area were localized to their normal retinotopic positions. One prominent difference between the lesioned and control conditions was the emergence of dorsoventral asymmetry in areas V2 and V3, with only dorsal, but not ventral, portions of these areas exhibiting V1-independent responses. This effect, which has also been observed in human blindsight⁷, cannot be attributed to the position of either the lesion or the stimulus, as the retinotopically matched stimulus in the opposite visual field evoked roughly equivalent responses in the dorsal and ventral parts.

On the behavioural level, both monkeys retained the ability to detect and make a saccadic eye movement to small (0.2° diameter) high-contrast, but not low-contrast, visual targets presented inside the scotoma, albeit with diminished performance. Target contrasts were adjusted based on performance in the control hemifield, such that there was reliable detection even at the lowest contrast level (7%) (Fig. 2c, f and Supplementary Methods). When low-contrast stimuli were presented inside the scotoma, both monkeys consistently maintained central fixation, indicating that they were unaware that any stimulus was being presented². By comparison, when we showed high-contrast (100%) stimuli, the monkeys were able to detect roughly half of the presentations, consistent with previous reports of blindsight in humans and monkeys^{1-4,8}. Finally, a direct comparison between detection performance and fMRI responses to the same set of stimuli (2° diameter) presented in the scotoma at varying luminance contrast levels confirmed the tight relationship between the two measures of blindsight (Supplementary Fig. 4).

presented inside (red line) or outside (blue line) the scotoma at different luminance contrast levels compared to a constant grey background. On one third of the trials, no stimulus was presented and the monkey was rewarded for maintaining central fixation (green line). Data represent mean $\pm$ s.e.m. from five experiments. **d**, Functional activation map of monkey 2's non-lesioned hemisphere to 95 cycles of visual stimulation outside the scotoma. **e**, Activation map of monkey 2's lesioned hemisphere to 95 cycles of visual stimulation inside the scotoma. **f**, Behavioural performance of monkey 2 for detecting visual stimuli inside (red line) and outside (blue line) the scotoma or during catch trials (green line). Data represent mean $\pm$ s.e.m. from five experiments. Although both monkeys display a large visual deficit, they continue to process visual information to some extent as performance improves with stimulus contrast.

Several controls ruled out the possibility that scattered light was contributing to the observed behavioural and fMRI responses to stimuli in the scotoma effects^{9,10}: first, behavioural responses were spatially accurate, as performance was calculated by considering only saccades that were within 1° of the target. Second, when the same high-contrast stimuli were presented monocularly in the monkeys' blindspots (the area covered by the optic disk in the retina), their performance fell to zero (Supplementary Fig. 5). Third, no systematic fMRI modulation was measured in the intact V1 of either monkey adjacent to the lesion (Fig. 2b, e). Fourth, strong extrastriate activation was found in a third monkey, in which we made a much larger lesion, encompassing opercular V1 and a portion of the adjacent V2 (Supplementary Fig. 6). As well as ruling out non-specific effects of scattering light^{9,10} as responsible for the observed effects, these experiments demonstrate a viable visual pathway that can operate in the absence of V1, in accordance with studies in human patients and lesion studies in monkeys^{1-8,11,12}.

We next investigated the components of the pathway that provide V1-bypassing input to the extrastriate cortex and support visual performance. Previous work has shown that direct anatomical projections exist from the LGN, the main thalamic relay between the retina and the primary visual cortex, to several extrastriate visual areas^{13,14}. We therefore proposed that residual activity in the extrastriate cortex, and corresponding behavioural performance, is the result of sensory signals transmitted directly from the LGN. To examine this possibility, we temporarily disrupted neural activity within the LGN of V1-lesioned monkeys by locally injecting the GABA_A-receptor agonist THIP¹⁵ (Methods Summary), and measured the effects on cortical fMRI responses and behavioural performance. Specifically, we reversibly inactivated the posterior part of the LGN, which represents the parafoveal

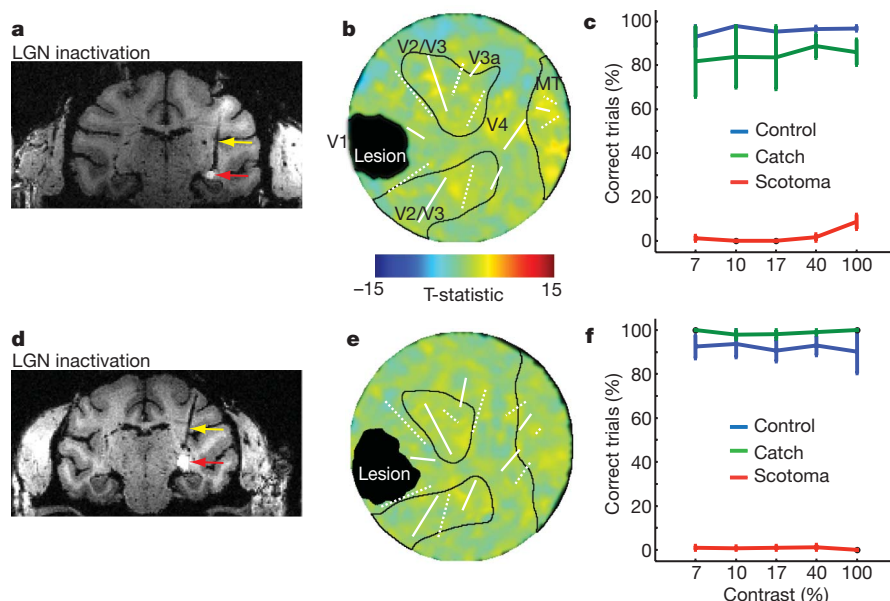


Figure 3 | Role of the LGN in driving V1-independent visual processing. **a**, We inactivated the LGN by injecting the GABA_A agonist THIP (Methods Summary). We co-injected the drug with the magnetic resonance contrast-agent gadolinium (total volume 2 μ l, Methods Summary), to visualize the site in magnetic resonance images. Here, a coronal section through the posterior part of monkey 1's LGN (+7 mm on the anterior-posterior axis in stereotactic coordinates) is shown. Injection of gadolinium resulted in a localized increase in the intensity of the T1-weighted magnetic resonance signal, with a diameter of $\sim$ 3 mm (red arrow). We ensured that injections were reproducible across experiments by permanently implanting a magnetic-resonance-compatible cannula (yellow arrow). **b**, Functional activation map of macaque 1's left, lesioned, hemisphere to visual

stimulation inside the scotoma (35 stimulation cycles) during inactivation of the LGN. LGN inactivation results in the elimination of V1-independent visual responses (Fig. 2b). **c**, Monkey 1's performance in detecting visual targets at different luminance contrasts. Data represent mean $\pm$ s.e.m. performance from three experiments with THIP injections into the LGN. The injections eliminated the monkey's ability to detect a target inside the scotoma. **d**, Inactivation of macaque 2's posterior LGN. **e**, Activation map of monkey 2's right, lesioned, hemisphere to visual stimulation inside the scotoma (60 stimulation cycles) during LGN inactivation. **f**, Monkey 2's performance for correctly detecting targets during LGN inactivation. Data represent the mean $\pm$ s.e.m. performance in three experiments with THIP injections.

visual hemifield and covers the scotoma-affected region (Methods Summary), on multiple occasions via chronically implanted guide tubes compatible with magnetic resonance. The magnetic resonance contrast-agent gadolinium was co-injected along with THIP, allowing us to visualize the spread of the injection in the tissue. In both monkeys (Fig. 3a, d), the 2 μ l injection diffused to an effective diameter of approximately 3 mm in the caudal LGN, as visualized by the gadolinium (Fig. 3a, d).

After we had inactivated the LGN, virtually all extrastriate responses in the V1-lesioned hemispheres disappeared (Fig. 3b, e), indicating that the residual activation to stimuli presented in the scotoma had indeed reached the extrastriate cortex by means of direct projections from the LGN. Moreover, inactivation of the LGN abolished the animals' residual capacity to detect high-contrast stimuli presented to the scotoma region of the visual field (red line, Fig. 3c, f), demonstrating that the LGN is the critical thalamic link that supports behavioural performance in blindsight. As a control, detection of visual stimuli in the opposite hemifield (outside the scotoma) was unaffected.

To obtain a more quantitative assessment of the amount of information transmitted directly via the LGN, we compared the strength of fMRI responses under normal visual stimulation conditions to those obtained inside the scotoma, with and without additional LGN inactivation, across all experimental sessions (Fig. 4). On average, across all monkeys and extrastriate areas, scotoma stimulation with the LGN intact had fMRI activation of $\sim$ 20% of that under normal conditions. This finding is in good qualitative agreement with previously reported fMRI activation patterns of human and monkey blindsight subjects^{6,7,11} and with single-unit recordings in area V5/MT of macaque monkeys with chronic V1 lesions¹². In addition, inactivating the LGN reduced activation levels to less than 5% of normal.

Apart from demonstrating the role of the LGN in supporting blindsight, these data reveal several interesting features of V1-independent

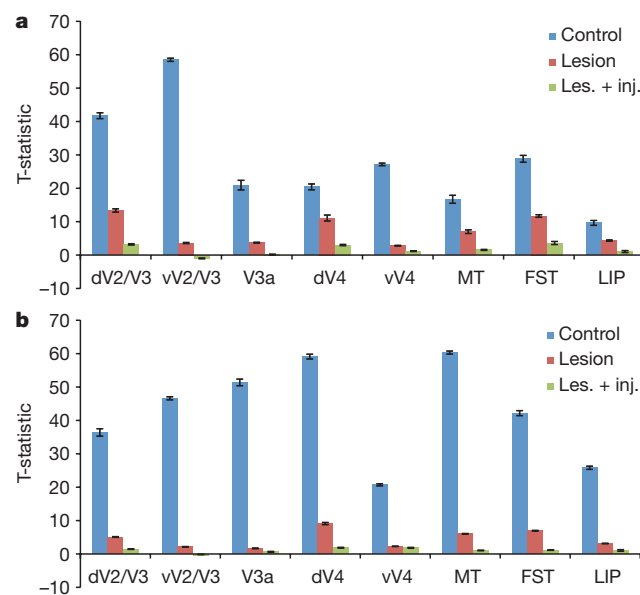


Figure 4 | Quantitative summary of mean fMRI activation levels. In extrastriate areas under normal conditions (V1 and LGN intact, blue bars); in the absence of V1 input (lesion, red bars); and in the absence of input from V1 and the LGN (Les. + inj., green bars). **a**, Data collected from monkey 1 and corresponding to the mean $\pm$ s.e.m. *t*-statistic obtained during 85 visual stimulation cycles without LGN inactivation and 35 stimulation cycles with LGN inactivation. **b**, Data from experiments with monkey 2, in which the mean $\pm$ s.e.m. *t*-statistic has been computed over 95 stimulation cycles without LGN inactivation and 60 stimulation cycles with LGN inactivation. On average across all areas and monkeys, fMRI activation in extrastriate areas is reduced by $\sim$ 80% when V1 input is missing. Additional LGN inactivation reduces activity by more than 95% compared to normal levels.

vision. For example, the asymmetry of visual responses in areas V2 and V3 to stimuli in the scotoma, which has previously been observed in blindsight patient GY⁷, may reflect the differential contributions of parallel visual streams to upper- and lower-field vision: namely, that a larger proportion of neurons in the retina and the LGN respond to lower-visual-field stimulation than to upper-visual-field stimulation¹⁶. Psychophysical evidence suggests that this bias is primarily related to the magnocellular system¹⁷, which, at the level of the retina, is also the system that is less affected by retrograde degeneration following V1 injury¹⁸. Thus, one possibility for the observed asymmetry following striate cortex lesions is that the normal visual-field asymmetry of the magnocellular system is unmasked by post-lesional neurodegeneration.

The near-complete drop in both extrastriate activation and behavioural performance not only implicates the LGN as a critical hub for blindsight, but also argues against the existence of pathways that do not involve the LGN. Blindsight functions have been observed for decades in a number of different species, and have traditionally been attributed to a second visual pathway in which retinal information is relayed via the superior colliculus and a secondary thalamic nucleus, the pulvinar, to the extrastriate cortex¹⁹. Although the role of the superior colliculus in mediating blindsight functions and V1-independent responses in area V5/MT has been demonstrated^{20,21}, there is presently no direct evidence that the pulvinar is involved. On the contrary, the pulvinar may have minimal contribution for the following reasons: its anatomical basis as a first-order visual-relay is in question²² (however, see refs 23 and 24); its neural responses appear to be driven more by cortical than by collicular inputs²⁵; there is no residual vision following LGN lesions²⁶; and neural activity in area V5/MT is eliminated during LGN inactivation²⁷.

Conversely, contribution of the LGN to blindsight has sometimes been left as an open possibility^{4,21}, particularly since the neural degeneration following V1 ablation selectively spares LGN cells that project directly to the extrastriate cortex²⁸. These V1-bypassing projection neurons appear to belong chiefly to the koniocellular system, whose neurons reside primarily in the intercalated LGN layers^{13,14,22}. Interestingly, it is these koniocellular-rich layers that receive input from the superior colliculus^{22,29}. This may explain the previously described effects of superior colliculus ablation on V1-independent visual processing^{20,21}. Finally, recent observations using diffusion-tensor magnetic resonance imaging in a human blindsight patient many years after a V1 lesion show significant projections between the LGN and area V5/MT³⁰, although the direction of the projection could not be established.

Our data demonstrate that the LGN has a causal role in V1-independent visual function. We propose that this residual function is mediated by neurons in the intercalated layers of the LGN, whose direct projections to the extrastriate cortex may not only support residual vision following V1 lesions, but may also serve as a shortcut to the high-level cortex during normal vision. In providing short-latency, eye-specific signals to high-level visual cortex, such a shortcut could serve to facilitate some forms of rapid behavioural responses to visual stimuli.

METHODS SUMMARY

The main experiments were carried out in two healthy adult monkeys (*Macaca mulatta*): one male (8 kg weight; monkey 1) and one female (5 kg weight; monkey 2). Additional control experiments were conducted in a third healthy adult monkey (female, 5 kg weight). All procedures followed the Institute of Laboratory Animal Research (part of the National Research Council of the National Academy of Sciences) guidelines and were approved by the National Institute of Mental Health (NIMH) Animal Care and Use Committee. To immobilize the head during experiments and to record eye movements during behavioural testing, we implanted headposts and eye coils following standard procedures. We caused V1-area lesions by coagulating pial vessels over the intended lesion area on the V1 operculum, and by aspirating grey matter within this area. In both monkeys the lesions were located ~2–7° visual eccentricities away from the fovea (monkeys had no problems maintaining fixation), covering

between one-third and one-half of opercular V1. We used a lesion in a third monkey, covering all of opercular V1 and large parts of the adjacent V2 (Supplementary Fig. 6), to evaluate the possibility of scattered-light effects on intact grey-matter tissue in the first two, smaller lesioned, monkeys. To reach the LGN for inactivation, we chronically implanted magnetic-resonance-compatible fused-silica guide tubes (Plastics One) in monkeys 1 and 2, using a frameless stereotaxy procedure (Brainsight, Rogue Research). For the experiments, we injected the GABA_A agonist THIP (Tocris; concentration: 6.67 µg µl⁻¹, dissolved in sterile saline, pH 7.4; volume: 2 µl; rate: 0.5–1 µl min⁻¹) for LGN inactivation together with the magnetic resonance contrast-agent gadolinium (Berlex Imaging, 5 mM) to visualize the site and extent of the injection in a magnetic resonance image.

Received 1 December 2009; accepted 18 May 2010.

Published online 23 June; corrected 15 July 2010 (see full-text HTML version for details).

- Weiskrantz, L., Warrington, E. K. & Sanders, M. D. Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* **97**, 709–728 (1974).
- Cowey, A. & Stoerig, P. Blindsight in monkeys. *Nature* **373**, 247–249 (1995).
- Keating, E. G. Residual spatial vision in the monkey after removal of striate and precipital cortex. *Brain Res.* **187**, 271–290 (1980).
- Cowey, A. The blindsight saga. *Exp. Brain Res.* **200**, 3–24 (2010).
- Brewer, A. A., Press, W. A., Logothetis, N. K. & Wandell, B. A. Visual areas in macaque cortex measured using functional magnetic resonance imaging. *J. Neurosci.* **22**, 10416–10426 (2002).
- Schmid, M. C., Panagiotaropoulos, T., Augath, M. A., Logothetis, N. K. & Smirnakis, S. M. Visually driven activation in macaque areas V2 and V3 without input from the primary visual cortex. *PLoS ONE* **4**, e5527, doi:10.1371/journal.pone.0005527 (2009).
- Baseler, H. A., Morland, A. B. & Wandell, B. A. Topographic organization of human visual areas in the absence of input from primary cortex. *J. Neurosci.* **19**, 2619–2627 (1999).
- Cowey, A. & Stoerig, P. Visual detection in monkeys with blindsight. *Neuropsychologia* **35**, 929–939 (1997).
- Collins, C. E., Lyon, D. C. & Kaas, J. H. Responses of neurons in the middle temporal visual area after long-standing lesions of the primary visual cortex in adult new world monkeys. *J. Neurosci.* **23**, 2251–2264 (2003).
- Campion, J., Latto, R. & Smith, Y. M. Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behav. Brain Sci.* **6**, 423–447 (1983).
- Goebel, R., Muckli, L., Zanella, F. E., Singer, W. & Stoerig, P. Sustained extrastriate cortical activation without visual awareness revealed by fMRI studies of hemianopic patients. *Vision Res.* **41**, 1459–1474 (2001).
- Rodman, H. R., Gross, C. G. & Albright, T. D. Afferent basis of visual response properties in area MT of the macaque. I. Effects of striate cortex removal. *J. Neurosci.* **9**, 2033–2050 (1989).
- Sincich, L. C., Park, K. F., Wohlgenuth, M. J. & Horton, J. C. Bypassing V1: a direct geniculate input to area MT. *Nature Neurosci.* **7**, 1123–1128 (2004).
- Bullier, J. & Kennedy, H. Projection of the lateral geniculate nucleus onto cortical area V2 in the macaque monkey. *Exp. Brain Res.* **53**, 168–172 (1983).
- Cope, D. W., Hughes, S. W. & Crunelli, V. GABA_A receptor-mediated tonic inhibition in thalamic neurons. *J. Neurosci.* **25**, 11553–11563 (2005).
- Curcio, C. A. & Allen, K. A. Topography of ganglion cells in human retina. *J. Comp. Neurol.* **300**, 5–25 (1990).
- McAnany, J. J. & Levine, M. W. Magnocellular and parvocellular visual pathway contributions to visual field anisotropies. *Vision Res.* **47**, 2327–2336 (2007).
- Cowey, A., Stoerig, P. & Perry, V. H. Transneuronal retrograde degeneration of retinal ganglion cells after damage to striate cortex in macaque monkeys: selective loss of Pβ cells. *Neuroscience* **29**, 65–80 (1989).
- Diamond, I. T. & Hall, W. C. Evolution of neocortex. *Science* **164**, 251–262 (1969).
- Mohler, C. W. & Wurtz, R. H. Role of striate cortex and superior colliculus in visual guidance of saccadic eye movements in monkeys. *J. Neurophysiol.* **40**, 74–94 (1977).
- Rodman, H. R., Gross, C. G. & Albright, T. D. Afferent basis of visual response properties in area MT of the macaque. II. Effects of superior colliculus removal. *J. Neurosci.* **10**, 1154–1164 (1990).
- Stepniewska, I., Qi, H. X. & Kaas, J. H. Do superior colliculus projection zones in the inferior pulvinar project to MT in primates? *Eur. J. Neurosci.* **11**, 469–480 (1999).
- Berman, R. A. & Wurtz, R. H. Functional identification of a pulvinar path from superior colliculus to cortical area MT. *J. Neurosci.* **30**, 6342–6354 (2010).
- Lyon, D. C., Nassi, J. J. & Callaway, E. M. A dysynaptic relay from superior colliculus to dorsal stream visual cortex in macaque monkey. *Neuron* **65**, 270–279 (2010).
- Bender, D. B. Visual activation of neurons in the primate pulvinar depends on cortex but not colliculus. *Brain Res.* **279**, 258–261 (1983).
- Schiller, P. H., Logothetis, N. K. & Charles, E. R. Functions of the colour-opponent and broad-band channels of the visual system. *Nature* **343**, 68–70 (1990).
- Maunsell, J. H., Nealey, T. A. & DePriest, D. D. Magnocellular and parvocellular contributions to responses in the middle temporal visual area (MT) of the macaque monkey. *J. Neurosci.* **10**, 3323–3334 (1990).
- Cowey, A. & Stoerig, P. Projection patterns of surviving neurons in the dorsal lateral geniculate nucleus following discrete lesions of striate cortex: implications for residual vision. *Exp. Brain Res.* **75**, 631–638 (1989).

29. Harting, J. K., Huerta, M. F., Hashikawa, T. & van Lieshout, D. P. Projection of the mammalian superior colliculus upon the dorsal lateral geniculate nucleus: organization of tectogeniculate pathways in nineteen species. *J. Comp. Neurol.* **304**, 275–306 (1991).
30. Bridge, H., Thomas, O., Jbabdi, S. & Cowey, A. Changes in connectivity after visual cortical brain damage underlie altered visual function. *Brain* **131**, 1433–1444 (2008).

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

Acknowledgements We thank A. Maier and D. McMahon for comments on the manuscript; S. Smirnakis, R. Berman, R. Wurtz, B. Richmond, S. Guderian and M. Fukushima for discussions; C. Zhu and H. Merkle for magnetic resonance coil construction; K. Smith, N. Phipps, J. Yu, G. Dold, D. Ide and T. Talbot for technical assistance; D. Sheinberg for developing visual stimulation software; and members of the Brian Wandell laboratory for developing and sharing mrVista software. This

work was supported by the Intramural Research Programme of the NIMH, the NINDS, and the NEI.

Author Contributions M.C.S. took the primary lead for all aspects of this work and wrote the paper; S.W.M. helped with experiments and analysis; J.T. helped with the experiments and developed the inactivation method; R.C.S. created the lesions; M.W. developed the inactivation method; A.J.P. helped with experiments and analysis; F.Q.Y. developed pre-processing software and optimized magnetic resonance sequences; and D.A.L. provided resources, acted in a supervisory role on all aspects of this work and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.C.S. (schmidmicha@gmail.com)

LETTERS

Planar cell polarity breaks bilateral symmetry by controlling ciliary positioning

Hai Song¹, Jianxin Hu¹, Wen Chen¹, Gene Elliott², Philipp Andre¹, Bo Gao¹ & Yingzi Yang¹

Defining the three body axes is a central event of vertebrate morphogenesis. Establishment of left–right (L–R) asymmetry in development follows the determination of dorsal–ventral and anterior–posterior (A–P) body axes^{1,2}, although the molecular mechanism underlying precise L–R symmetry breaking in reference to the other two axes is still poorly understood. Here, by removing both *Vangl1* and *Vangl2*, the two mouse homologues of a *Drosophila* core planar cell polarity (PCP) gene *Van Gogh* (*Vang*), we reveal a previously unrecognized function of PCP in the initial breaking of lateral symmetry. The leftward nodal flow across the posterior notochord (PNC) has been identified as the earliest event in the *de novo* formation of L–R asymmetry^{3,4}. We show that PCP is essential in interpreting the A–P patterning information and linking it to L–R asymmetry. In the absence of *Vangl1* and *Vangl2*, cilia are positioned randomly around the centre of the PNC cells and nodal flow is turbulent, which results in disrupted L–R asymmetry. PCP in mouse, unlike what has been implicated in other vertebrate species, is not required for ciliogenesis, cilium motility, Sonic hedgehog (*Shh*) signalling or apical docking of basal bodies in ciliated tracheal epithelial cells. Our data suggest that PCP acts earlier than the unidirectional nodal flow during bilateral symmetry breaking in vertebrates and provide insight into the functional mechanism of PCP in organizing the vertebrate tissues in development.

Identifying the mechanism underlying the initial bilateral symmetry breaking has been an overriding goal in understanding how nature is able to produce organisms with consistent and distinct left and right anatomies. In mice, bilateral symmetry is broken by a leftward flow across a pit-like, teardrop-shaped PNC (also referred to as ‘the node’) generated by posteriorly localized motile cilia^{3–6}. However, an earlier asymmetric event that positions the cilia to the posterior sides of PNC cells in response to the A–P patterning signals remains to be identified. PCP, which originally referred to an asymmetry of epithelial cells along an axis orthogonal to the apical–basal axis, is an evolutionarily conserved mechanism by which positional information is transmitted and maintained between and within cells⁷. A group of core PCP proteins including the transmembrane protein *Vang* have been identified in *Drosophila*⁸. Although mutants of core PCP genes have so far failed to show defects in L–R asymmetry, this is probably due to functional redundancy among members of a core PCP component. To overcome this and uncover important regulatory functions of PCP, we inactivated both *Vangl1* and *Vangl2*, the two mouse homologues of *Drosophila* *Vang*.

Vangl1 and *Vangl2* were broadly expressed during L–R patterning (Supplementary Fig. 1). Before the appearance of ciliated PNC cells, levels of *Vangl1* and *Vangl2* protein were too low to be detected. However, after embryonic day (E)7.75, the protein levels of both *Vangl1* and *Vangl2* were selectively and progressively upregulated

in the PNC cells, and asymmetrical localization of *Vangl1* and *Vangl2* protein was observed along the A–P axis (Supplementary Fig. 2a–c), indicating that the PNC cells are indeed polarized and PCP may be required by the PNC cells to regulate early embryonic development. The *Vangl1* and *Vangl2* localization patterns are also consistent with the role of PCP in regulating convergent extension of neurulating vertebrate embryos⁹. In the sensory epithelia of the cochlea, *Vangl1* and *Vangl2* demonstrated similar asymmetrical localization patterns in both hair cells and supporting cells (Supplementary Fig. 2d, e). These data suggest that *Vangl1* and *Vangl2* are redundant in PCP, which should be severely disrupted by removing both *Vangl* genes.

A null allele of *Vangl1* (*Vangl1^{gt}*) was generated with a gene-trap embryonic stem-cell line (Supplementary Fig. 3). We also generated a floxed *Vangl2* allele, *Vangl2^f*, from which the null *Vangl2^Δ* allele was derived (Supplementary Fig. 3). *Vangl1* and *Vangl2* regulated PCP in a gene dose-dependent manner (Fig. 1 and Supplementary Fig. 4). Asymmetrical *Vangl1* localization was still detected in the cochlea of the *Vangl2^{Δ/Δ}* mutant (Fig. 1a). The *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos had the most severe developmental defects in comparison with embryos with other genotypes, particularly when sensory hair cell polarity and convergent extension were examined to assess PCP. Thus, *Vangl1* and *Vangl2* function of regulating PCP was removed in the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos.

The *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos had multiple laterality defects, which were not detected in embryos with other genotypes (Fig. 2a). The *Vangl2^{Δ/+}*; *Vangl1^{gt/+}* mice were fertile, morphologically normal and served as controls in our study. The *Vangl2^{Δ/+}*; *Vangl1^{gt/+}* and *Vangl2^{Δ/Δ}*; *Vangl1^{gt/+}* embryos always turned to the right side and their hearts correctly looped to the right ($n > 50$), but in all of the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos assayed ($n > 10$), both embryonic turning and heart looping failed. In addition, the normal mouse lung has five lobes, four on the right and one on the left. In the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos, however, mild left lung isomerism was observed with three lobes on the right side (4/9). Because PCP in visceral organs may also control morphogenesis in later stages, to assess the early L–R asymmetry defects in the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos further, we examined the components in the Nodal signal cascade that are asymmetrically expressed and required for L–R patterning¹⁰. In control embryos, *Nodal* was first expressed symmetrically around the PNC (somite 0–1). Then, between somite 2 and somite 7 stages, *Nodal* expression on the left side of the PNC became progressively stronger and its expression in the left lateral plate mesoderm (LPM) was also detected^{11,12}. In the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos, however, *Nodal* expression around the PNC was decreased (6/11) (Fig. 2b, c). Strikingly, more robust *Nodal* expression on the right side of the PNC was also observed in the *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}* embryos (5/11) (Fig. 2b, c). In the LPM of *Vangl2^{Δ/Δ}*; *Vangl1^{gt/gt}*

¹Developmental Genetics Section, Genetic Disease Research Branch, National Human Genome Research Institute, Bethesda, Maryland 20892, USA. ²Transgenic Mouse Core, Genetic Disease Research Branch, National Human Genome Research Institute, Bethesda, Maryland 20892, USA.

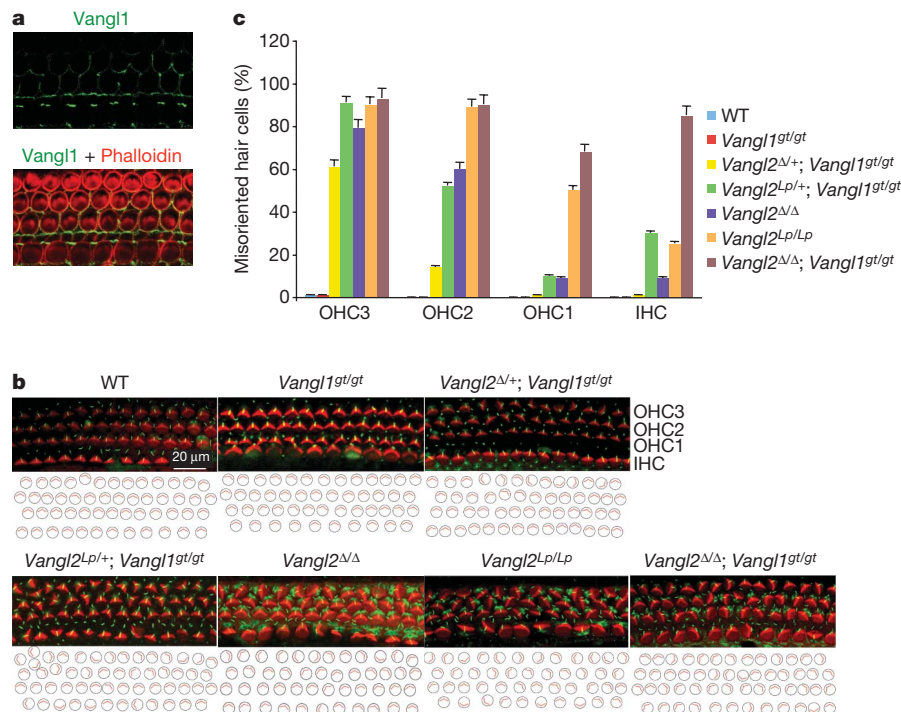


Figure 1 | The *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos have the most severe polarity defects. **a**, Asymmetrical localization of Vangl1 protein (green) in the sensory hair cells of the *Vangl2*^{Δ/Δ} cochlea at E18.5. **b**, Sensory hair cell polarity at E18.5 is shown by the orientation of actin-based stereocilium bundles (red) and the microtubule-based kinocilia (green). WT, wild type. **c**, Quantification of misoriented hair cells in embryos with different

genotypes. Results are means and s.d. for three samples. The orientation defects in the cochlea ranged from few misoriented cells (*Vangl2*^{Δ/+}; *Vangl1*^{gt/gt}) to several affected cells (*Vangl2*^{Δ/Δ}, *Vangl2*^{Lp/Lp}; *Vangl1*^{gt/gt} and *Vangl2*^{Lp/Lp}) to almost complete loss of polarity (*Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt}). OHC, outer hair cells; IHC, inner hair cells.

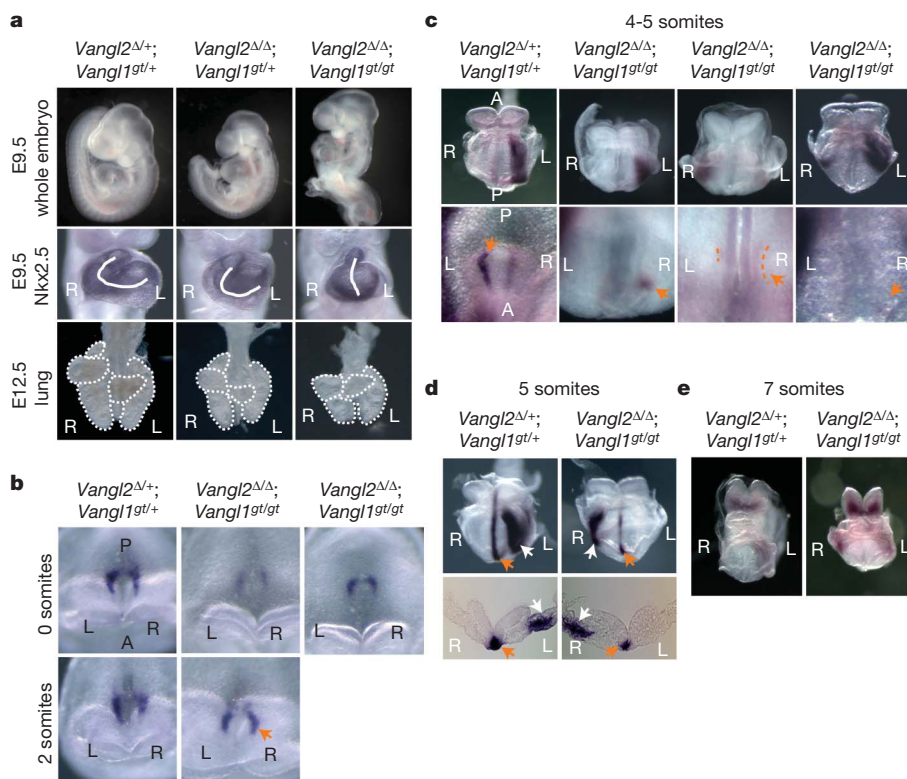


Figure 2 | Laterality defects in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. **a**, Failure of embryonic turning and heart (marked by *Nkx2.5* expression) looping in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. The *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} lung has three lobes (outlined) instead of four on the right side. **b**, Perinodal *Nodal* expression is enhanced on the right side of the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryo (arrow). **c**, Ventral views of *Nodal* expression in LPM (top) and

around the node (bottom, arrows). Weak perinodal *Nodal* expression is outlined. **d**, *Lefty1* expression (orange arrow) on the left side of prospective floor plate and *Lefty2* expression (white arrow) in the LPM, shown in ventral views (top) and transverse sections (bottom). **e**, Ventral views of *Pitx2* expression.

embryos, reversed (1/8) or bilateral (3/8) *Nodal* expression sites were observed (Fig. 2c). Controlled by *Nodal* signalling, expression of *Lefty1*, *Lefty2* and *Pitx2* was observed only on the left side of the control embryos^{10,13–15}. In the double mutant embryos, *Lefty1* and *Lefty2* expression sites (1/2) were reversed (Fig. 2d) and bilateral *Pitx2* expression (2/7) was found (Fig. 2e). These results demonstrate that PCP is required upstream of *Nodal* to pattern the L–R axis. The incomplete penetrance of the laterality defects in the *Vangl1/2* mutant embryo is possibly due to one or more other inputs to the PCP pathway¹⁶.

Because PCP has been implicated in ciliogenesis¹⁷, we examined the morphology and movement of cilia in PNC cells. Cilia are long, thin microtubule-based projections from the basal bodies of most mammalian cells¹⁸. The PNC cell has a single motile cilium³, which is essential for establishing leftward fluid flow across the ventral surface of the PNC. When examined by scanning electron microscopy and immunofluorescent labelling, cilia appeared normal in the PNC of the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos (*n* = 4 in each case) at E8.0 (Fig. 3a, b). Thus, the L–R asymmetry defects in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos are not secondary to defects in ciliogenesis. The leftward nodal flow is enabled by the chirality of the cilium, as it rotates clockwise³, and the posterior ciliary localization in the PNC cell, which results in a posterior tilt of the cilium and its rotation axis^{5,6}. Because it has been postulated that the posterior positioning of the basal body might be controlled by PCP⁶, we examined basal body localization in the PNC cells of the *Vangl* mutant embryos (Fig. 3c, d). The basal bodies were mostly localized posteriorly in both *Vangl2*^{Δ/+}; *Vangl1*^{gt/+} and *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/+} embryos. However, in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos the PNC cell density was slightly decreased (Supplementary Fig. 5a) and the number of posteriorly localized basal bodies was significantly decreased. Most basal bodies remained around the centre and anteriorly localized

basal bodies were increased, resulting in an even distribution of basal bodies on the anterior and posterior sides of the PNC cells in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos (Fig. 3d and Supplementary Fig. 5b, c). This loss of posterior localization of basal bodies in the *Vangl* mutant was not a result of developmental delay (Supplementary Fig. 5c). Thus, posterior positioning of basal bodies in the PNC cells is indeed controlled by PCP, in the absence of which the basal bodies are mostly localized around the centre. Cilia in the PNC emerge from the centre of each cell at late streak stage (about E7.25). By early head fold stage (E7.75), most cilia are located in the posterior side of the PNC cells^{6,19}. Because positioning of the kinocilium in sensory hair cells is also an event downstream of PCP²⁰, ciliary positioning may be a common readout of PCP.

To test whether altered ciliary positioning in the PNC cells affects function, we examined nodal flow (Fig. 4A and Supplementary Movies 1 and 2). In control embryos, fast and smooth leftward bead movement across the PNC was observed. In the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos, bead movement was still biased leftwards, although beads were often trapped in swirling vortices and required more time to cross the PNC. Whereas the nodal flow of control embryos generated straight and parallel bead paths, in the mutant, bead paths contained ‘knots’ and abrupt turns and were often crossed with each other (2/2). When observed directly, the cilia in the PNC of the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryo were beating at a speed and direction similar to those of the control (Fig. 4B and Supplementary Movies 3 and 4). The abnormal nodal flow in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos also disrupted the left-biased perinodal rise of intracellular calcium in control embryos²¹ (Fig. 4C, D). The incidence of bilateral increase or decrease in intracellular calcium concentration was increased in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. The ectopic calcium signal on the right side was weaker and the domain was narrower and shorter. These results, taken together, show that unidirectionality of nodal flow was significantly

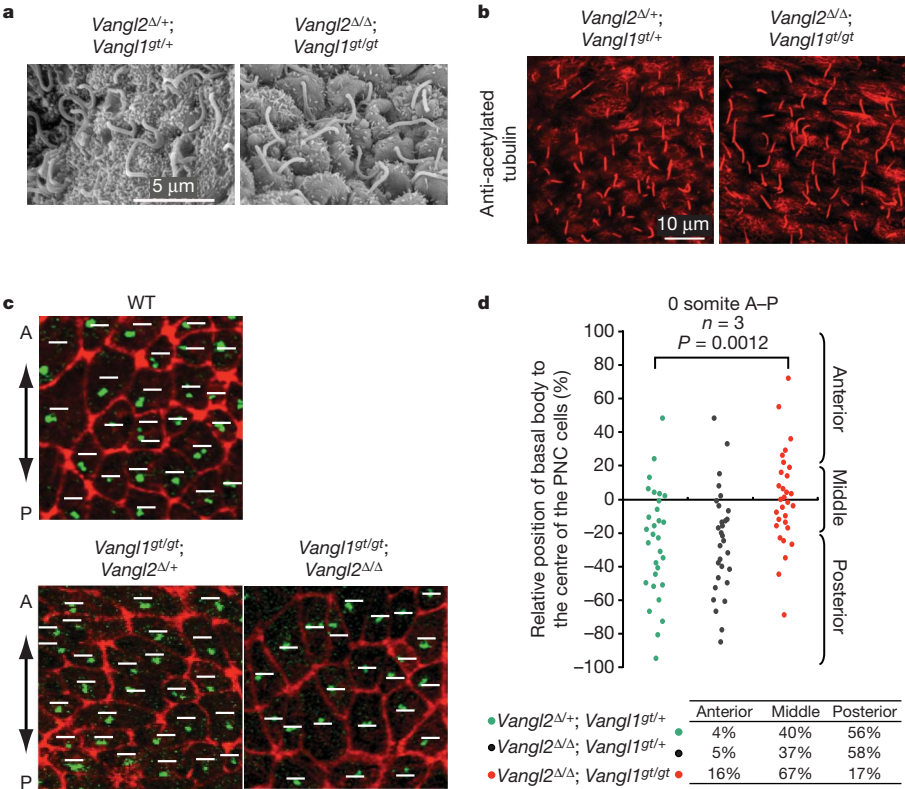


Figure 3 | Randomized ciliary positioning in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. **a**, Scanning electron microscopy of the PNC showing normal cilia at E8.0. **b**, Normal cilia (acetylated tubulin, red) in the PNC of the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryo at E8.0. **c**, Basal bodies (γ-tubulin, green) and cell boundaries (F-actin, red) of PNC cells at the 0-somite stage. White lines

indicate the middle position along the A–P axis of each PNC cell. WT, wild type. **d**, Plot of relative positions of basal bodies to the midline (along the A–P axis) of the PNC cells. Posteriorly biased basal body positioning was lost in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. The table shows percentages of PNC cells with indicated cilium localization.

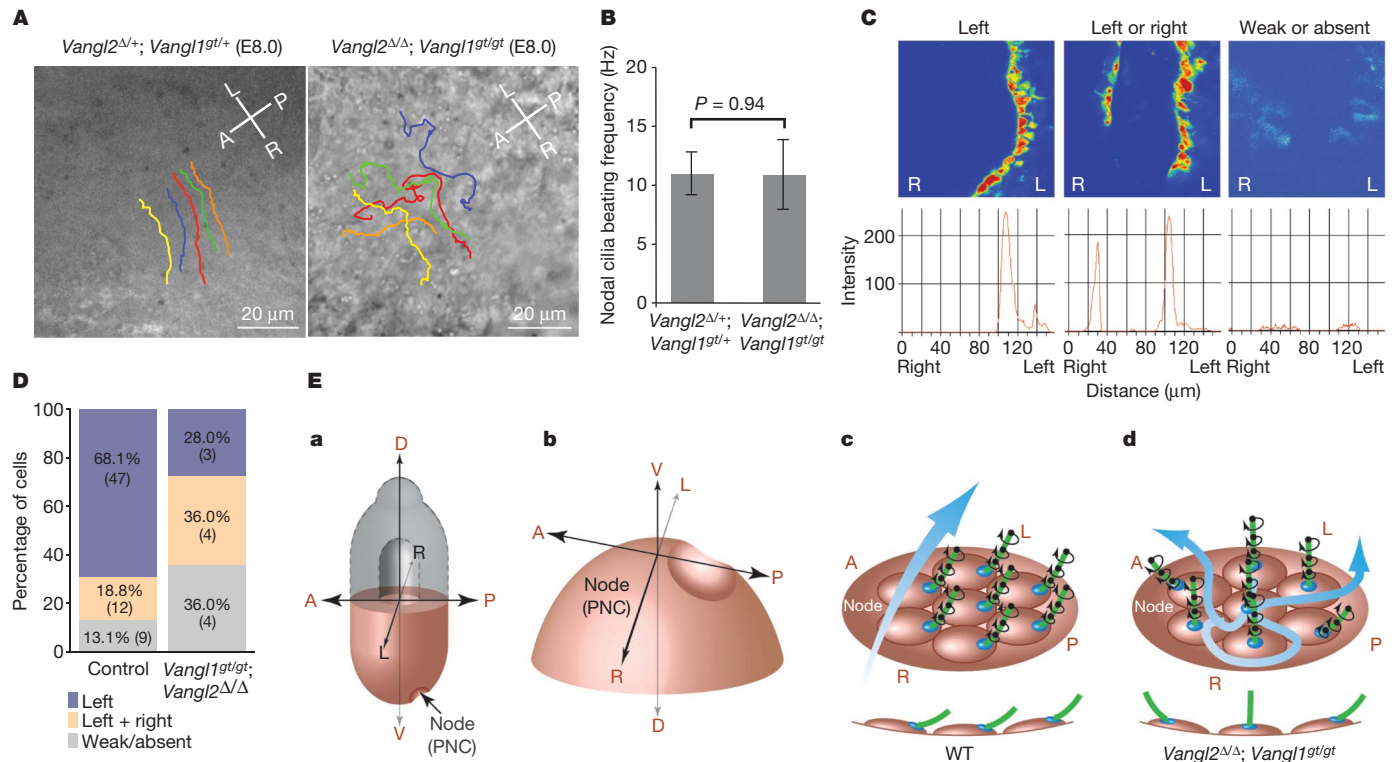


Figure 4 | *Vangl1* and *Vangl2* are required for normal leftward nodal flow.

A, Nodal flow shown by tracing latex bead movement across the PNC with lines in different colours (Supplementary Movie 1, 2). **B**, Ciliary beating was recorded by live imaging (Supplementary Movies 3 and 4) and the mean beating frequency was calculated from at least ten nodal cilia of three embryos with indicated genotypes. Results are means $\pm$ s.d. **C**, Intracellular calcium release around the node, intensity scales are shown in the lower panels.

compromised in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos. We have therefore identified PCP as the first step in breaking L–R symmetry by transmitting the A–P positional information to the PNC cells to enable posterior ciliary positioning (Fig. 4E).

Our data show that PCP is not required for ciliogenesis in the mouse (Fig. 1b and Supplementary Fig. 6c–f). In addition, *Vangl1* and *Vangl2* were not co-localized with cilia in any of the cells we examined, including the PNC cells, sensory hair cells, tracheal epithelial cells and floor plate cells (Supplementary Figs 2a, d and 6a, b). In the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos at E18.5, there were no defects in ciliogenesis or docking of the basal bodies along the apical–basal axis of the multiciliated tracheal epithelial cells (Supplementary Fig. 6c, d). Normal Hedgehog signalling in the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryos and cells provides further evidence that primary cilia are functionally normal (Supplementary Figs 6g, h and 7).

To promote a robust break in symmetry, vertebrates have evolved mechanisms of early asymmetry initiation followed by asymmetry amplification and stabilization²². Because PCP is a major conserved molecular mechanism by which directional cellular and tissue behaviours are controlled through reorganizing the cytoskeleton, it could lead to both cilia-dependent and cilia-independent regulation of L–R asymmetry by controlling asymmetrical ciliary localization in ciliated PNC or cell movement around the node in the chick, which lacks ciliated PNC^{5,6,23}. Involvement of PCP in regulating L–R asymmetry has been implicated previously: mutations in *Inversin* (*Inv*), *Bicaudal C* and *Seahorse* lead to L–R asymmetry defects upstream of *Nodal* expression, and these genes can interact with a core PCP gene, *Dishevelled* (*Dvl*)^{24–26}. *Dvl* has recently been found to regulate ciliary localization in the PNC and has been implicated in the regulation of L–R patterning²⁷. However, it is unlikely that *Inv* acts mainly through PCP because *Inv*, but not *Vangl1/2*, is localized to PNC cilia, and because the phenotypes of the *Inv* and *Vangl1/2* mutants are different

D, Distribution of embryos (percentages and numbers) according to calcium patterns. **E**, Schematic diagram showing the role of *Vangl1* and *Vangl2* in breaking L–R symmetry. D, dorsal; V, ventral. **a**, A mouse embryo at E8.0. **b**, Enlarged distal view of **a**. **c**, Normal nodal flow generated by posteriorly positioned cilia. A side view of ciliated cells is shown in the lower panel. WT, wild type. **d**, In the *Vangl2*^{Δ/Δ}; *Vangl1*^{gt/gt} embryo, randomly positioned cilia around the centre of the PNC cells resulted in aberrant, turbulent nodal flow.

in many aspects. Because abnormal L–R patterning is also associated with multiple congenital diseases²⁸, the availability of *Vangl1* and *Vangl2* conditional double mutants has opened a door to the investigation of the roles of PCP in many fundamental aspects of vertebrate development and diseases.

Note added in proof: While this work was in the press, two independent studies^{29,30} were published; both also found that planar cell polarity mediated by *Vangl2* is required for orienting and positioning motile cilia.

METHODS SUMMARY

Scanning electron microscopic analysis. Embryos were dissected at E7.75 and fixed overnight in 2% glutaraldehyde, 4% paraformaldehyde, 0.1 M sodium cacodylate, pH 7.4 at 4 °C. Standard procedures were followed to prepare samples. Samples were examined with a Hitachi S-570 scanning electron microscope.

Analysis of calcium signalling. Embryos were dissected at E7.75–E8.0 and analysed in accordance with the established procedure²¹.

Videomicroscopy. Embryos were dissected at E7.75–E8.0 in DMEM medium containing 20% FBS. The node was placed face down in a chamber with a coverslip on the bottom. Fluospheres (Invitrogen) were added into the medium to reveal the direction of nodal flow. Wide-field images were collected with a Personal DeltaVision system (Applied Precision Inc.) mounted on an inverted Olympus IX71 microscope with an oil-immersion Olympus Plan Apo N 60 $\times$ 1.42 numerical aperture objective lens. Images were acquired with a CoolSNAP ES2 camera with 2 $\times$ 2 binning and an imaging field of 128 pixels $\times$ 128 pixels.

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

Received 25 January; accepted 26 April 2010.

Published online 20 June 2010.

1. Wilhelm, H. Experimentelle Untersuchungen über situs inversus viscerum. *Arch. EntwMech.* **48**, 517–532 (1921).
2. Brown, N. A. & Wolpert, L. The development of handedness in left/right asymmetry. *Development* **109**, 1–9 (1990).

3. Nonaka, S. *et al.* Randomization of left–right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell* **95**, 829–837 (1998).
4. Blum, M. *et al.* Ciliation and gene expression distinguish between node and posterior notochord in the mammalian embryo. *Differentiation* **75**, 133–146 (2007).
5. Nonaka, S. *et al.* De novo formation of left–right asymmetry by posterior tilt of nodal cilia. *PLoS Biol.* **3**, e268 (2005).
6. Okada, Y., Takeda, S., Tanaka, Y., Belmonte, J. C. & Hirokawa, N. Mechanism of nodal flow: a conserved symmetry breaking event in left–right axis determination. *Cell* **121**, 633–644 (2005).
7. Seifert, J. R. & Mlodzik, M. Frizzled/PCP signalling: a conserved mechanism regulating cell polarity and directed motility. *Nature Rev. Genet.* **8**, 126–138 (2007).
8. Axelrod, J. D. & McNeill, H. Coupling planar cell polarity signaling to morphogenesis. *ScientificWorldJournal* **2**, 434–454 (2002).
9. Jessen, J. R. *et al.* Zebrafish trilobite identifies new roles for Strabismus in gastrulation and neuronal movements. *Nature Cell Biol.* **4**, 610–615 (2002).
10. Hamada, H., Meno, C., Watanabe, D. & Saijoh, Y. Establishment of vertebrate left–right asymmetry. *Nature Rev. Genet.* **3**, 103–113 (2002).
11. Lowe, L. A. *et al.* Conserved left–right asymmetry of nodal expression and alterations in murine situs inversus. *Nature* **381**, 158–161 (1996).
12. Collignon, J., Varlet, I. & Robertson, E. J. Relationship between asymmetric nodal expression and the direction of embryonic turning. *Nature* **381**, 155–158 (1996).
13. Nakamura, T. *et al.* Generation of robust left–right asymmetry in the mouse embryo requires a self-enhancement and lateral-inhibition system. *Dev. Cell* **11**, 495–504 (2006).
14. Meno, C. *et al.* *Lefty-1* is required for left–right determination as a regulator of *Lefty-2* and *nodal*. *Cell* **94**, 287–297 (1998).
15. Yoshioka, H. *et al.* *Pitx2*, a bicoid-type homeobox gene, is involved in a Lefty-signaling pathway in determination of left–right asymmetry. *Cell* **94**, 299–305 (1998).
16. Lawrence, P. A., Struhl, G. & Casal, J. Planar cell polarity: one or two pathways? *Nature Rev. Genet.* **8**, 555–563 (2007).
17. Park, T. J., Haigo, S. L. & Wallingford, J. B. Ciliogenesis defects in embryos lacking inturned or fuzzy function are associated with failure of planar cell polarity and Hedgehog signaling. *Nature Genet.* **38**, 303–311 (2006).
18. Poole, C. A., Flint, M. H. & Beaumont, B. W. Analysis of the morphology and function of primary cilia in connective tissues: a cellular cybernetic probe? *Cell Motil.* **5**, 175–193 (1985).
19. Lee, J. D. & Anderson, K. V. Morphogenesis of the node and notochord: the cellular basis for the establishment and maintenance of left–right asymmetry in the mouse. *Dev. Dyn.* **237**, 3464–3476 (2008).
20. Jones, C. *et al.* Ciliary proteins link basal body polarization to planar cell polarity regulation. *Nature Genet.* **40**, 69–77 (2008).
21. McGrath, J., Somlo, S., Makova, S., Tian, X. & Brueckner, M. Two populations of node monocilia initiate left–right asymmetry in the mouse. *Cell* **114**, 61–73 (2003).
22. Tabin, C. J. The key to left–right asymmetry. *Cell* **127**, 27–32 (2006).
23. Gros, J., Feistel, K., Viebahn, C., Blum, M. & Tabin, C. J. Cell movements at Hensen's node establish left/right asymmetric gene expression in the chick. *Science* **324**, 941–944 (2009).
24. Kishimoto, N., Cao, Y., Park, A. & Sun, Z. Cystic kidney gene seahorse regulates cilia-mediated processes and Wnt pathways. *Dev. Cell* **14**, 954–961 (2008).
25. Maisonneuve, C. *et al.* Bicaudal C, a novel regulator of Dvl signaling abutting RNA-processing bodies, controls cilia orientation and leftward flow. *Development* **136**, 3019–3030 (2009).
26. Simons, M. *et al.* Inversin, the gene product mutated in nephronophthisis type II, functions as a molecular switch between Wnt signaling pathways. *Nature Genet.* **37**, 537–543 (2005).
27. Hashimoto, M. *et al.* Planar polarization of node cells determines the rotational axis of node cilia. *Nature Cell Biol.* **12**, 170–176 (2010).
28. Sharma, N., Berbari, N. F. & Yoder, B. K. Ciliary dysfunction in developmental abnormalities and diseases. *Curr. Top. Dev. Biol.* **85**, 371–427 (2008).
29. Guirao, B. *et al.* Coupling between hydrodynamic forces and planar cell polarity orients mammalian motile cilia. *Nature Cell Biol.* **12**, 341–350 (2010).
30. Borovina, A., Superina, S., Voskas, D. & Ciruna, B. Vangl2 directs the posterior tilting and asymmetric localization of motile primary cilia. *Nature Cell Biol.* **12**, 407–412 (2010).

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

Acknowledgements We thank members of the Yang laboratory for stimulating discussion; M. Kuehn, A. Kumar and J. Regard for critical reading of the manuscript; J. Fekacs and D. Leja for preparing the graphics; D. Wu for teaching us to dissect the mouse cochlea; and M. Kelley for the anti-Vangl2 antibodies. This study is supported by the intramural research programme of National Human Genome Research Institute.

Author Contributions H.S. designed and performed the experiments, generated most of the data in the manuscript, and interpreted and presented the data in figures. J.H. designed and performed some of the experiments, conducted statistical analysis, interpreted the data and put it together in figures. W.C. analysed and maintained the *Vangl1* mutant mice. G.E. generated the *Vangl1* and *Vangl2* mutant mice by performing blastocyst injection and tetraploid aggregation. P.A. and B.G. provided some critical technical support and help in writing the manuscript. Y.Y. designed and supervised the entire project, interpreted all data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Y.Y. (yingzi@mail.nih.gov).

METHODS

Generation of and genotyping mouse lines. Generation of the floxed *Vangl2* allele is described in Supplementary Fig. 1e. *Vangl2^Δ* was generated by deleting exons 4 and 5 using the *Sox2-Cre* line³¹. *Vangl1^{8t}* allele was generated by using a Baygenomic gene-trap embryonic stem-cell line (XL802) (Supplementary Fig. 1a). Genotyping was performed with PCR using genomic DNA prepared from yolk sac or tail. *Vangl2P1* (5'-CCATCGATGCACATGTGGTATC-3') and *Vangl2P2* (5'-GACAC TGTCTCCATGTCCTTG-3') were used to genotype the *Vangl2^f* allele and *Vangl2⁺* allele. *Vangl2P1* and *Vangl2P3* (5'-GCTATGACCAGCTACCTGAAG TC-3') were used to genotype the *Vangl2^Δ* allele. *Vangl1P1* (5'-CCCTGGC TTTCTTGTGGTC-3') and *Vangl1P3* (5'-GGGCTGGCTCTTGGAGTCAT-3') were used to genotype the *Vangl1⁺* allele. *Vangl1P1* and *Vangl1P2* (5'-CCTGGG GTTCGTGTCCTACA-3') were used to genotype the *Vangl1^{8t}* allele.

Whole-mount *in situ* hybridization and LacZ staining. Whole-mount *in situ* hybridization and LacZ staining were performed as described previously³². Embryos were cryosectioned at 20 μm.

Immunofluorescence and antibodies. Embryos were fixed in 4% paraformaldehyde for 30 min at 4 °C, washed in PBS and incubated overnight in PBS containing 30% sucrose. Embryos were embedded in OCT compound (Tissue-Tek) and sectioned at 14 μm. Sections were permeabilized with 0.5% Triton X-100 in PBS and blocked for 1 h in 3% BSA, 0.1% Triton X-100 in PBS. Sections were incubated overnight with primary antibodies at 4 °C. Secondary antibodies were applied for 1 h at room temperature (22–25 °C). Primary antibodies used in immunofluorescence included mouse anti-Pax6, Pax7, Nkx2.2, Lim1/2, Nkx6.1, Hnf3β and Islet-1 (1:50 dilution; Developmental Studies Hybridoma Bank), rabbit anti-Olig2 (1:100; Abcam), mouse anti-acetylated-tubulin (1:5,000; Sigma), goat anti-γ-tubulin (1:200; Santa Cruz), goat anti-Vangl2 (1:100; Santa Cruz), rabbit anti-Vangl2 (1:50; gift from M. Kelley's laboratory) and goat anti-E-Cadherin (1:500; R&D). Vangl1 antibodies were generated against Vangl1 (residues 1–110) polypeptides fused to glutathione S-transferase in rabbit. Anti-Vangl1 antibodies were purified with the immunizing peptide coupled to Affigel (Bio-Rad). Alexa Fluor 488 and 568 of donkey anti-mouse, rabbit and goat secondary antibodies were from Invitrogen. Confocal images were acquired with a Zeiss LSM 510 NLO microscope. Excitation wavelengths of 488, 561 and 740 nm were used for the detection of secondary antibodies conjugated to Alexa Fluor 488, 568 and 4,6-diamidino-2-phenylindole, respectively.

Whole-mount immunostaining of cochlea or early E6–E8.5 embryos. Dissected cochlea or embryos were fixed for 30 min in 4% paraformaldehyde at 4 °C. Cochleae or embryos were permeabilized for 5 min with 0.5% Triton X-100 in PBS and blocked for 1 h in PBS with 3% BSA, 0.1% Triton X-100 at room temperature, followed by incubation overnight with primary antibodies at

4 °C. Cochleae or embryos were washed for 1 h in PBS with 0.1% Triton X-100 at room temperature, followed by incubation for 1 h with fluorescent secondary antibodies and Alexa Fluor-594-conjugated phalloidin (1:200; Invitrogen) at room temperature. The actin-based stereocilium bundles were shown by phalloidin staining, whereas the microtubule-based kinocilia or PNC cilia were shown by staining with the mouse anti-acetylated-tubulin antibody (1:5,000; Sigma). After being washed for 1 h with 0.1% Triton X-100 in PBS at room temperature, cochleae or the PNC region were flat mounted. Confocal images were acquired with a Zeiss LSM510 confocal microscope.

Immunoblot analysis. Embryonic tissues or cells were lysed in RIPA buffer. Cell lysates were analysed by immunoblotting. Primary antibodies used included goat anti-Vangl2 (1:2,000), rabbit anti-Vangl1 (1:1,000) and rabbit anti-Gli3 (1:5,000; gift from S. Mackem).

Analysis of calcium signalling. Embryos were dissected at E7.75–E8.0 and incubated for 20 min with the fluorescent calcium indicator Fluo-4 (Invitrogen) at 10 μM in DMEM with 20% FBS at 37 °C. Embryos were then washed three times in DMEM with 20% FBS and placed with the node facing upwards in a silicon chamber under a coverslip. Embryos were incubated for 20 min at 37 °C before confocal imaging. Confocal images were acquired with a Zeiss LSM510 confocal microscope. An excitation wavelength of 488 nm was used to detect Fluo-4. Embryos were genotyped after imaging. Data were analysed with Image-Pro 6.3 software. High intensity was shown in red and low intensity in blue. Intensity scales of the calcium signal were plotted from the left to the right side of the node.

Videomicroscopy. Embryos were dissected at E7.75–E8.0 in DMEM with 20% FBS. The node was placed facing downwards in a chamber with a coverslip bottom. For the nodal flow assay, Fluospheres (Invitrogen) were added into the medium to reveal the direction of nodal flow. Embryos were genotyped after imaging. Wide-field images were collected with a Personal DeltaVision system (Applied Precision Inc.) mounted on an inverted Olympus IX71 microscope with an Olympus Plan Apo N 60× 1.42 numerical aperture objective lens. All images were acquired with a CoolSNAP ES2 camera with 2 × 2 binning and an imaging field of 128 pixels × 128 pixels. Videos were created with SoftWoRx software package version 4.0.0 (Applied Precision). Individual bead movement was traced and shown as a line with a different colour. The mean cilia beating frequency was calculated from at least ten nodal cilia of embryos with indicated genotypes at E8.0. Results are means ± s.d. *P* was calculated with Student's *t*-test.

31. Hayashi, S., Tenzen, T. & McMahon, A. P. Maternal inheritance of Cre activity in a Sox2Cre deleter strain. *Genesis* **37**, 51–53 (2003).
32. Topol, L. *et al.* Wnt-5a inhibits the canonical Wnt pathway by promoting GSK-3-independent β-catenin degradation. *J. Cell Biol.* **162**, 899–908 (2003).

Members of the H3K4 trimethylation complex regulate lifespan in a germline-dependent manner in *C. elegans*

Eric L. Greer^{1,2}, Travis J. Maures¹, Anna G. Hauswirth¹, Erin M. Green⁴, Dena S. Leeman^{1,2}, Géraldine S. Maro³, Shuo Han¹, Max R. Banko¹, Or Gozani^{2,4} & Anne Brunet^{1,2}

The plasticity of ageing suggests that longevity may be controlled epigenetically by specific alterations in chromatin state. The link between chromatin and ageing has mostly focused on histone deacetylation by the Sir2 family^{1,2}, but less is known about the role of other histone modifications in longevity. Histone methylation has a crucial role in development and in maintaining stem cell pluripotency in mammals³. Regulators of histone methylation have been associated with ageing in worms^{4–7} and flies⁸, but characterization of their role and mechanism of action has been limited. Here we identify the ASH-2 trithorax complex⁹, which trimethylates histone H3 at lysine 4 (H3K4), as a regulator of lifespan in *Caenorhabditis elegans* in a directed RNA interference (RNAi) screen in fertile worms. Deficiencies in members of the ASH-2 complex—ASH-2 itself, WDR-5 and the H3K4 methyltransferase SET-2—extend worm lifespan. Conversely, the H3K4 demethylase RBR-2 is required for normal lifespan, consistent with the idea that an excess of H3K4 trimethylation—a mark associated with active chromatin—is detrimental for longevity. Lifespan extension induced by ASH-2 complex deficiency requires the presence of an intact adult germline and the continuous production of mature eggs. ASH-2 and RBR-2 act in the germline, at least in part, to regulate lifespan and to control a set of genes involved in lifespan determination. These results indicate that the longevity of the soma is regulated by an H3K4 methyltransferase/demethylase complex acting in the *C. elegans* germline.

Genome-wide RNAi screens for genes that regulate lifespan in *C. elegans* have been performed previously in worms in which progeny production was inhibited^{4,10,11}, which could mask the effect of some genes on lifespan. We performed a targeted RNAi screen in fertile worms by selecting genes that encode known worm methyltransferases, proteins containing the enzymatic domain of methyltransferases (SET domain), or orthologues of regulators of histone methylation. As reported previously⁴, *set-9* and *set-15* knockdown extended worm lifespan (Fig. 1a). We also found that *ash-2*, *set-2* and *set-4* knockdown extended fertile worm lifespan, with *ash-2* knockdown having the most significant effect (23.1–30.9%, $P < 0.0001$) (Fig. 1a, Supplementary Fig. 1a). ASH-2 is a member of an H3K4 trimethylation (H3K4me3) complex in yeast, flies and mammals^{9,12,13} and is important for the conversion of H3K4 dimethylation (H3K4me2) to H3K4me3 in mammals¹⁴. *ash-2* knockdown decreased global H3K4me3 levels at larval stage L3 (Fig. 1b), indicating that ASH-2 promotes histone H3 trimethylation at lysine 4 in *C. elegans*. WDR-5/TAG-125 is a WD40-repeat protein that interacts with ASH-2 in mammals, and is important for the mono-, di- and trimethylation of H3K4 in mammals and worms^{15,16}. *wdr-5/tag-125* knockdown or *wdr-5/tag-125* deletion also decreased H3K4me3

levels (Fig. 1b) and significantly extended *C. elegans* lifespan (~30%, $P < 0.0001$ and 16–28%, $P < 0.0001$, respectively) (Fig. 1c, Supplementary Fig. 1b). Thus, ASH-2 and WDR-5 promote H3K4 trimethylation and normally limit lifespan in *C. elegans*.

In mammals, ASH-2 and WDR-5 form a complex with several H3K4me3 methyltransferases of the SET1/mixed lineage leukaemia (MLL) family¹⁷. There are four SET1/MLL orthologues in *C. elegans*: SET-1, SET-2, SET-12 and SET-16 (Supplementary Fig. 1c). Of these, only SET-2, which is similar to mammalian SET1A/SET1B, regulated worm lifespan (*set-2* knockdown: ~20%, $P < 0.0001$; *set-2* deletion *set-2(ok952)*: 17–26%, $P < 0.0001$) (Fig. 1a, d; Supplementary Fig. 1d, e). As reported previously¹⁶, *set-2(ok952)* mutant worms or worms treated with *set-2* RNAi had reduced H3K4me3 levels (Fig. 1e). SET-2 was relatively specific in regulating lifespan and H3K4me3 in worms, as neither SET-9 nor SET-15 affected global H3K4me3 levels (Supplementary Fig. 1f), even though they both regulate lifespan⁴. The worm SET-2 methyltransferase domain expressed in bacteria directly methylated histone H3 at lysine 4 *in vitro* (Fig. 1f). SET2 generated H3K4me2 but not H3K4me3 *in vitro* (Supplementary Fig. 1g), consistent with the fact that ASH-2 is required for the conversion of H3K4me2 to H3K4me3 (ref. 18). To test if ASH-2, WDR-5 and SET-2 act together to regulate lifespan, we performed epistasis experiments. Lifespan extension by *ash-2* or *wdr-5/tag-125* knockdown was significantly less pronounced in *set-2(ok952)* mutants than in wild-type worms (Fig. 1g, combined two-way ANOVA $P < 0.0001$). Similarly, *ash-2* knockdown did not extend further the lifespan of *wdr-5/tag-125(ok1417)* mutant worms (Supplementary Fig. 1h, $P = 0.1534$). Thus, ASH-2, WDR-5 and SET-2 probably act in the same pathway, perhaps in a complex, to limit lifespan. As *ash-2* and *wdr-5/tag-125* knockdown slightly extended *set-2(ok952)* mutant worm lifespan (Fig. 1g, Supplementary Table 2), other methyltransferases may also complex with ASH-2 and WDR-5 to regulate lifespan.

RBR-2 is an H3K4me3 demethylase involved in vulva formation in worms¹⁹ and is homologous to human RBP2 and PLU-1 (also known as KDM5A and KDM5B, respectively), two H3K4me3 demethylases of the JARID family¹⁹. Consistent with a published report¹⁹, *rbr-2(tm1231)* mutant worms showed increased H3K4me3 levels (Fig. 2a). Both *rbr-2(tm1231)* mutation and *rbr-2* knockdown decreased lifespan significantly (15–25%, $P < 0.0001$ and 10–24.2%, $P < 0.0001$, respectively) (Fig. 2b, Supplementary Fig. 2a), indicating that RBR-2 is necessary for normal longevity. The decrease of H3K4me3 induced by *ash-2* knockdown was less pronounced in *rbr-2(tm1231)* mutant worms than in wild-type worms (Fig. 2c). In addition, *ash-2* knockdown no longer extended *rbr-2(tm1231)* mutant worm lifespan (Fig. 2d, $P = 0.4673$). Similarly, the lifespan of *wdr-5/tag-125(ok1415); rbr-2(tm1231)* and *set-2(ok952); rbr-2(tm1231)* double mutants was similar to that of

¹Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. ²Cancer Biology Graduate Program, Stanford University, Stanford, California 94305, USA. ³Departments of Biology and Pathology, Stanford University, Stanford, California 94305, USA. ⁴Department of Biology, Stanford University, Stanford, California 94305, USA.

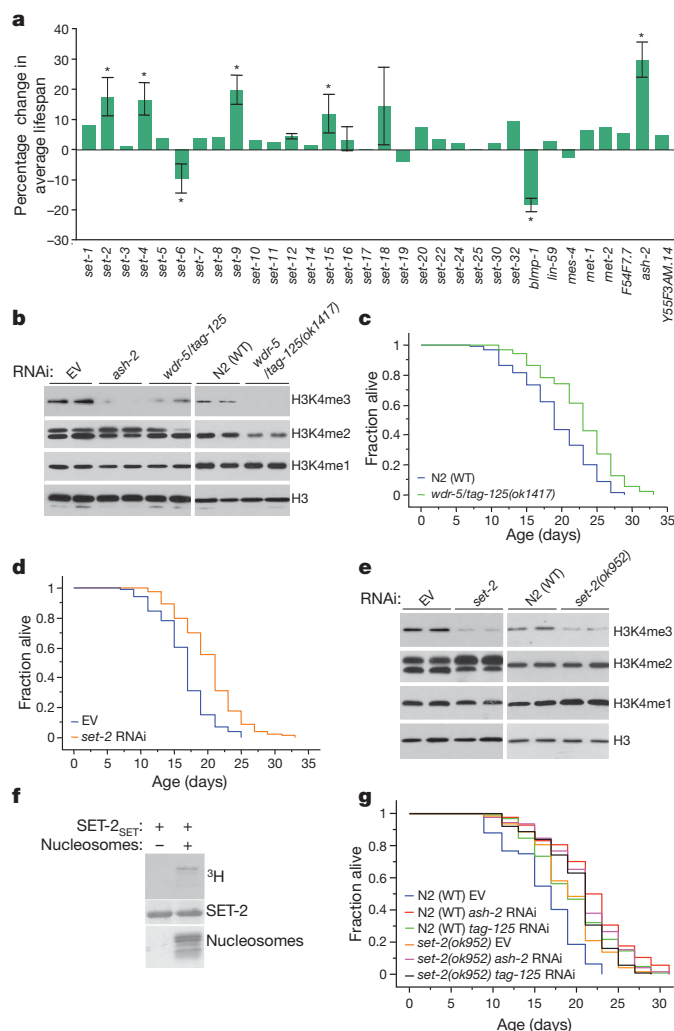


Figure 1 | ASH-2, WDR-5 and SET-2 function together to regulate H3K4me3 and lifespan in *C. elegans*. **a**, Percentage change in average lifespan in worms treated with RNAi to the indicated genes compared with empty vector. Mean $\pm$ s.d. of two or three independent experiments is shown. * $P < 0.05$. **b**, Western blots on L3 worm extracts (representative of four independent experiments). EV, empty vector; WT, wild type. **c**, *wdr-5/tag-125(ok1417)* mutant worms have an extended lifespan. **d**, *set-2* knockdown extends lifespan. **e**, Western blots of L3 worm extracts after *set-2* knockdown or deletion (representative of three independent experiments). **f**, The methyltransferase domain of worm SET-2 (SET-2_{SET}) methylates histone H3 *in vitro*. **g**, *ash-2* and *wdr-5/tag-125* knockdown slightly extend the lifespan of *set-2(ok952)* mutant worms, but significantly less than in the N2 (WT) background. *ash-2* mRNA was efficiently knocked-down by RNAi in *set-2(ok952)* mutant worms (Supplementary Fig. 3a). Statistics are presented in Supplementary Tables 1 and 2.

rbr-2(tm1231) single mutants (Supplementary Fig. 2b, c, $P = 0.2121$ and $P = 0.6943$, respectively). Together, these results indicate that RBR-2 counteracts the effects of ASH-2/WDR-5/SET-2 on H3K4me3 and lifespan.

Whole-mount immunocytochemistry with a monoclonal ASH-2 antibody (Supplementary Fig. 4a) showed that ASH-2 is expressed in the nuclei of cells from many tissues, and is highly expressed in the germline (Fig. 3a). Visualization of an ASH-2–GFP (green fluorescent protein) fusion driven by the endogenous *ash-2* promoter in low-copy or high-copy transgenic lines confirmed that ASH-2 is highly expressed in the germline and in newly formed eggs (Supplementary Fig. 4b, c). Visualization of an RBR-2–GFP fusion driven by the endogenous *rbr-2* promoter in low-copy or high-copy transgenic lines showed expression of RBR-2 in the nuclei of cells from many tissues, although RBR-2 was not particularly enriched in

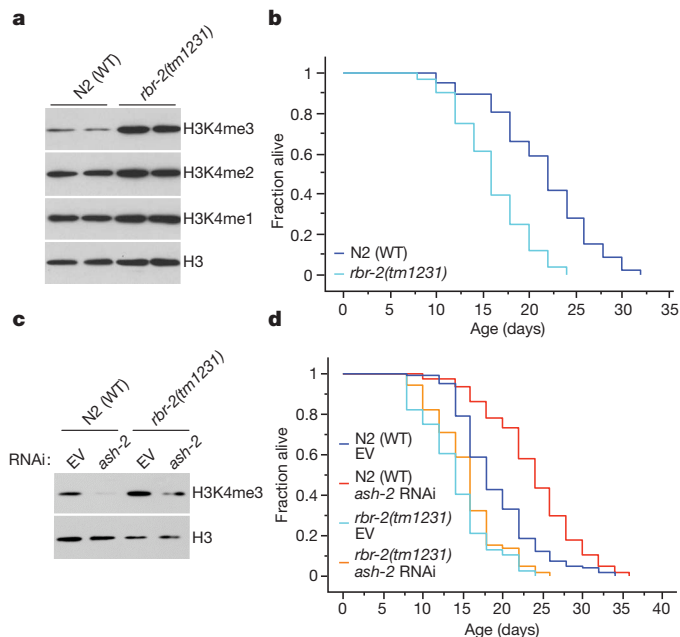


Figure 2 | RBR-2 is an H3K4me3 demethylase that counteracts the effect of the ASH-2 methyltransferase complex. **a**, Western blots of *rbr-2(tm1231)* mutant worms (representative of three independent experiments). **b**, *rbr-2(tm1231)* mutant worms have a decreased lifespan. **c**, Western blots of *rbr-2(tm1231)* mutant worms treated with *ash-2* RNAi (representative of two independent experiments). **d**, *ash-2* knockdown does not extend the lifespan of *rbr-2(tm1231)* mutant worms. *ash-2* mRNA was efficiently knocked down by RNAi in *rbr-2(tm1231)* mutant worms (Supplementary Fig. 3a). Statistics are presented in Supplementary Table 3.

the germline or eggs (Supplementary Fig. 4d, data not shown). The H3K4me3 mark was present in the nuclei of cells from all tissues, with high abundance in the germline in L3 and young adult worms (Fig. 3b, Supplementary Fig. 4e), consistent with the high levels of ASH-2 and SET-2 (ref. 20) in the germline. As expected, *wdr-5/tag-125(ok1417)* and *set-2(ok952)* mutant worms had markedly reduced H3K4me3 levels (Fig. 3b, Supplementary Fig. 4e). Conversely, *rbr-2(tm1231)* mutant worms showed increased H3K4me3 levels, particularly in the germline (Fig. 3b, Supplementary Fig. 4e). These results confirm that members of the ASH-2 complex and RBR-2 regulate H3K4me3, and that this mark is abundant in the germline.

To determine if the presence of an intact germline is necessary for lifespan regulation by ASH-2, we examined the effects of *ash-2* knockdown in *glp-1(e2141ts)* mutant worms, which develop only 5–15 meiotic germ cells instead of $\sim 1,500$ when shifted to the restrictive temperature at the L1 stage^{21,22}. *ash-2* knockdown did not extend further the long lifespan of *glp-1(e2141ts)* mutant worms (Fig. 3c). We verified that at the permissive temperature, *ash-2* knockdown extended *glp-1(e2141ts)* mutant worm lifespan (Supplementary Fig. 5a, b). Consistent with these observations, *ash-2* knockdown did not extend the lifespan of sterile *glp-4(bn2ts)* or *pgl-1(bn101ts)* mutant worms (Supplementary Fig. 5c, d). The presence of an intact germline was also necessary for lifespan regulation by WDR-5, SET-2 and RBR-2 (Fig. 3d, Supplementary Fig. 5e–g). In contrast, lifespan regulation by SET-9 and SET-15 did not depend on an intact germline (Supplementary Fig. 6a, b), indicating that not all methyltransferases require the germline to control lifespan.

We performed a genome-wide analysis to identify ASH-2-regulated genes that are dependent on the presence of an intact germline. *ash-2* knockdown led to changes in the expression of 220 genes at the L3 stage and 847 genes at day 8 of life (day 5 of adulthood) in wild-type worms (Fig. 3e, Supplementary Tables 6–9). The majority of ASH-2-controlled genes were regulated in a germline-dependent manner, as their expression was affected in wild-type worms, but not in *glp-1(e2141ts)* mutant worms (93.6% at L3 and 69.8% at day 8 of life)

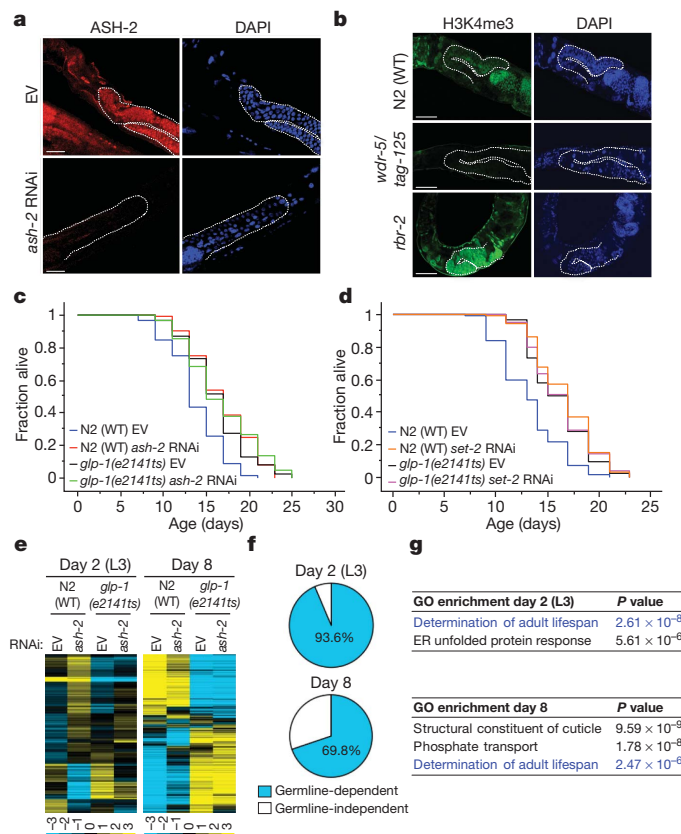


Figure 3 | An intact germline is necessary for longevity and gene expression control by the H3K4me3 regulatory complex. **a, b**, Whole-mount immunofluorescence of young adult worms stained with an ASH-2 antibody (**a**) or an H3K4me3 antibody (**b**). Dashed lines mark the germline. Scale bars, 50 μ m. **c**, *ash-2* knockdown does not extend further the long lifespan of *glp-1(e2141ts)* mutant worms that were shifted to the restrictive temperature at the L1 stage. **d**, *set-2* knockdown does not extend further the long lifespan of germline-deficient *glp-1(e2141ts)* mutant worms that were shifted to the restrictive temperature at the L1 stage. Statistics are presented in Supplementary Table 4. **e**, Microarray clusters of genes that change significantly on *ash-2* knockdown in wild-type worms, but not in *glp-1(e2141ts)* mutants. Experimental values are presented in Supplementary Tables 8 and 9. **f**, Percentage of genes regulated by ASH-2 only in wild-type worms, but not in *glp-1(e2141ts)* mutants. **g**, Top gene ontology (GO) terms for genes regulated by ASH-2 in wild-type worms, but not in *glp-1(e2141ts)* mutants.

(Fig. 3f). Interestingly, one of the top gene ontology (GO) enrichment categories for genes regulated by ASH-2 at L3 and day 8 was 'determination of adult lifespan' (Fig. 3g, $P = 2.61 \times 10^{-8}$ and $P = 2.47 \times 10^{-6}$, respectively). Furthermore, at both L3 and day 8 of life, ASH-2-regulated genes were also significantly enriched (twofold, $P = 5 \times 10^{-6}$ and 1.5-fold, $P = 6 \times 10^{-6}$, respectively) for genes whose expression changed during ageing²³.

To determine if members of the ASH-2 complex function in the germline itself to regulate lifespan, we used *rrf-1(pk1417)* mutant worms in which RNAi is efficient in the germline, but not in somatic cells²⁴. *ash-2* or *set-2* knockdown still extended lifespan in *rrf-1(pk1417)* mutants (23.6%, $P < 0.0001$ and 22.5%, $P < 0.0001$, respectively) (Fig. 4a, Supplementary Fig. 7a), indicating that ASH-2 and SET-2 function in the germline to regulate lifespan. Conversely, to test if RBR-2 expression in the germline was sufficient to extend lifespan, we generated low-copy integrants of RBR-2-GFP driven by the germline-specific *pie-1* promoter (*P_{pie-1}::rbr-2::gfp*). Three independent *P_{pie-1}::rbr-2::gfp* lines had an extended lifespan compared to two independent control *P_{pie-1}::gfp* lines (17.9–30.7%, $P < 0.005$) (Fig. 4b), indicating that expression of RBR-2 in the germline is sufficient to prolong worm lifespan. High-copy *P_{rbr-2}::rbr-2::gfp* transgenic worms

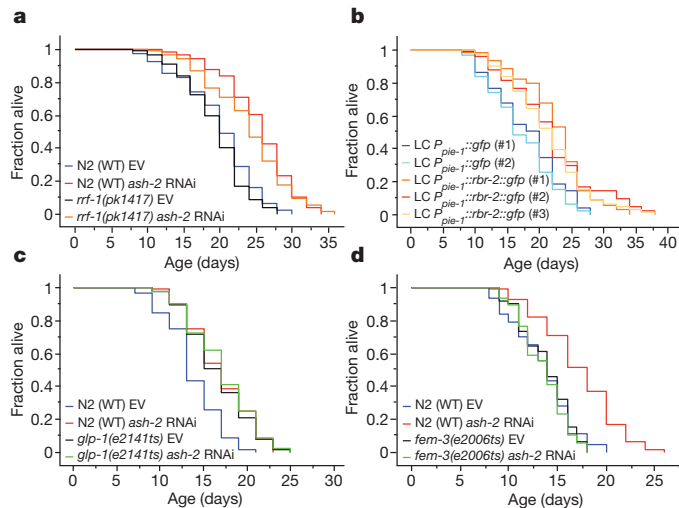


Figure 4 | ASH-2/RBR-2 function primarily in the germline to regulate lifespan and require the continuous production of mature eggs for lifespan extension. **a**, *ash-2* RNAi extends the lifespan of *rrf-1(pk1417)* mutant worms, which are deficient for RNAi in the soma, to a similar extent as in N2 (WT) worms. **b**, Three independent lines of low-copy (LC) integrant *P_{pie-1}::rbr-2::gfp* transgenic worms have increased lifespan compared to two independent lines of *P_{pie-1}::gfp* transgenic worms. **c**, *ash-2* knockdown does not extend further the long lifespan of *glp-1(e2141ts)* mutant worms that were shifted to the restrictive temperature at the young adult stage. **d**, *ash-2* knockdown does not extend the lifespan of *fem-3(e2006ts)* mutant worms, which do not produce mature eggs. *ash-2* mRNA was efficiently knocked down by RNAi in *fem-3(e2006ts)* mutants (Supplementary Fig. 3b). Statistics are presented in Supplementary Tables 4 and 5.

also had a modest increase in lifespan (10–14%, $P < 0.005$) (Supplementary Fig. 7b). Because high-copy transgenic worms tend to have suppression of transgene expression in the germline²⁵, these results indicate that expression of RBR-2 in somatic cells may also contribute to lifespan extension, or that very low amounts of RBR-2 expressed in the germline in these high-copy transgenics may be sufficient to extend lifespan. Thus, the ASH-2 methyltransferase complex and the RBR-2 demethylase act in the germline, at least in part, to regulate lifespan.

The ASH-2 complex and RBR-2 do not seem to regulate lifespan by simply affecting progeny production. *ash-2* knockdown worms were fertile and did not have significant differences in number of eggs laid (Supplementary Fig. 8a), number of live progeny (Supplementary Fig. 8b), germline cell number (Supplementary Fig. 8c) and germline morphology (data not shown). Similarly, *set-2(ok952)*, *wdr-5/tag-125(ok1417)* and *rbr-2(tm1231)* mutant worms did not have significant differences in egg number and were mostly fertile (Supplementary Fig. 8d).

We next investigated whether ASH-2 and RBR-2 require an intact adult or developing germline to regulate lifespan. Knocking-down *ash-2* and *rbr-2* in young adults, after the germline has developed, still affected lifespan (*ash-2*: +26%, $P < 0.0001$; *rbr-2*: –13.5%, $P < 0.0001$) (Supplementary Fig. 9a, b). Furthermore, *ash-2* knockdown no longer extended lifespan in *glp-1(e2141ts)* mutant worms switched to the restrictive temperature at the young adult stage (Fig. 4c), whereas *ash-2* knockdown extended lifespan in *glp-1(e2141ts)* mutant worms that were switched to the restrictive temperature after the egg laying period (Supplementary Fig. 9c). Similar observations were made with *glp-4(bn25)* mutant worms (Supplementary Table 4). Consistently, *ash-2* knockdown no longer extended lifespan in another *glp-1* mutant, *glp-1(e2142ts)*, which has a fully developed germline, but no adult germline stem cells or eggs (Supplementary Fig. 9d). Finally, *ash-2* knockdown did not extend further the long lifespan of worms treated with 5-fluorodeoxyuridine (FUDR), a drug that inhibits the proliferation of germline stem cells and the production of intact eggs in adults²⁶ (Supplementary Fig. 9e). Similarly, *rbr-2* knockdown did not decrease—and

even slightly extended—worm lifespan in the presence of FUDR (Supplementary Fig. 9f), consistent with a previous study²⁷. Thus, ASH-2 regulates lifespan in adult worms by a mechanism that depends on the presence of adult germline stem cells and/or the continuous production of eggs.

To distinguish if the germline stem cells or the continuous production of eggs is required for the regulation of lifespan by ASH-2, we used *fem-3(e2006ts)* mutant worms, which have functional germline stem cells, but do not produce fertilized eggs²⁸. *ash-2* knockdown no longer extended the lifespan of *fem-3(e2006ts)* mutant worms (Fig. 4d), indicating that the production of mature eggs, rather than germline stem cells, is necessary for *ash-2* knockdown to extend lifespan. Furthermore, *ash-2* knockdown still significantly extended the lifespan of *daf-16(mu86)* and *daf-12(m20)* mutant worms, two genes involved in signalling pathways that are critical for germline stem cells to regulate lifespan^{21,22,29} (Supplementary Fig. 10a, b). Consistently, *set-2(ok952)* mutant worms displayed lifespan extension even in the presence of *daf-16* RNAi (Supplementary Table 2). These results indicate that the ability of members of the ASH-2 complex to regulate lifespan depends on the continuous production of mature eggs, but not on the germline stem cell signalling pathway.

Here we show that members of an H3K4me3 methyltransferase complex have a pivotal role in the regulation of longevity in *C. elegans*. Our results indicate that the ASH-2 complex and RBR-2 regulate ageing at least in part by controlling trimethylation of H3K4, although they might also act by controlling the methylation of non-histone proteins. ASH-2 regulates the expression of a specific subset of genes enriched for lifespan determination genes, but more global changes in chromatin state and gene expression—and therefore quantity of protein produced—may also be responsible for the modulation of lifespan by ASH-2. The observation that the continuous production of mature eggs throughout adulthood is necessary for ASH-2 to regulate lifespan in *C. elegans* may explain why the ASH-2 trimethylase complex was not identified in earlier screens for ageing genes in *C. elegans*, in which the production of mature eggs was inhibited^{4,10,11}. ASH-2/RBR-2 may function in the germline—perhaps in the maturing eggs—to control endocrine hormones that would, in turn, regulate longevity of the soma. Alternatively, ASH-2/RBR-2 may control the expression of genes involved in lifespan determination, and additional signals from the germline/mature eggs may be required to allow lifespan extension. The regulation of H3K4me3 by specific methyltransferase complexes may regulate ageing in a germline-dependent manner in other species. Male flies heterozygous for the H3K4me3 methyltransferase *trx* had a normal lifespan⁸, but *trx* may not be the specific H3K4me3 methyltransferase that regulates lifespan, or the H3K4me3 regulatory complex may be more important in females/hermaphrodites than in males for controlling longevity. The finding that ageing can be regulated epigenetically raises the intriguing possibility that aspects of the ageing process could be reversed.

METHODS SUMMARY

Lifespan assays. Worm lifespan assays were performed at 20 °C, without FUDR, as described previously³⁰ unless noted otherwise. For each lifespan assay, 90 worms per condition were used in three plates (30 worms per plate), except for the initial RNAi screen, which was performed with 30–60 worms per condition.

Statistical analysis. Statistical analyses of lifespan were performed on Kaplan–Meier survival curves in StatView 5.0.01 by Logrank (Mantel–Cox) tests. For statistical comparison of independent replicates, Fisher's combined probability tests were performed. To compare the interaction between genotype and RNAi, two-way ANOVA tests were performed.

Microarray analysis. N2 (wild type) and *glp-1(e2141ts)* mutant worms were hatched at 16 °C and switched to bacteria with control empty vector or *ash-2* RNAi and moved to the restrictive temperature of 25.5 °C at the L1 stage. RNA was isolated at larval stage L3 or day 8 of life. Each condition was done in independent triplicate with ~1,000 worms per triplicate for L3 worms, and ~200 worms picked every day after adulthood to fresh plates per triplicate for day 8 worms. Total RNA was isolated using an RNAqueous kit (Ambion). Microarray hybridization was performed at the Stanford Protein and Nucleic Acid facility with oligonucleotide arrays (Affymetrix, GeneChip *C. elegans*

Genome Arrays). Background adjustment and normalization was performed with RMA (Robust Multiarray Analysis).

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

Received 3 December 2009; accepted 28 May 2010.

Published online 16 June 2010.

- Blander, G. & Guarente, L. The Sir2 family of protein deacetylases. *Annu. Rev. Biochem.* **73**, 417–435 (2004).
- Dang, W. *et al.* Histone H4 lysine 16 acetylation regulates cellular lifespan. *Nature* **459**, 802–807 (2009).
- Nottke, A., Colaiacovo, M. P. & Shi, Y. Developmental roles of the histone lysine demethylases. *Development* **136**, 879–889 (2009).
- Hamilton, B. *et al.* A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev.* **19**, 1544–1555 (2005).
- Li, J. *et al.* *Caenorhabditis elegans* HCF-1 functions in longevity maintenance as a DAF-16 regulator. *PLoS Biol.* **6**, e233 (2008).
- McColl, G. *et al.* Pharmacogenetic analysis of lithium-induced delayed aging in *Caenorhabditis elegans*. *J. Biol. Chem.* **283**, 350–357 (2008).
- Chen, S. *et al.* The conserved NAD(H)-dependent corepressor CTBP-1 regulates *Caenorhabditis elegans* life span. *Proc. Natl Acad. Sci. USA* **106**, 1496–1501 (2009).
- Siebold, A. P. *et al.* Polycomb repressive complex 2 and trithorax modulate *Drosophila* longevity and stress resistance. *Proc. Natl Acad. Sci. USA* **107**, 169–174 (2010).
- Wysocka, J., Myers, M. P., Laherty, C. D., Eisenman, R. N. & Herr, W. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3–K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev.* **17**, 896–911 (2003).
- Lee, S. S. *et al.* A systematic RNAi screen identifies a critical role for mitochondria in *C. elegans* longevity. *Nature Genet.* **33**, 40–48 (2003).
- Hansen, M., Hsu, A. L., Dillin, A. & Kenyon, C. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS Genet.* **1**, e17 (2005).
- Miller, T. *et al.* COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc. Natl Acad. Sci. USA* **98**, 12902–12907 (2001).
- Papoulas, O. *et al.* The *Drosophila* trithorax group proteins BRM, ASH1 and ASH2 are subunits of distinct protein complexes. *Development* **125**, 3955–3966 (1998).
- Dou, Y. *et al.* Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nature Struct. Mol. Biol.* **13**, 713–719 (2006).
- Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121**, 859–872 (2005).
- Simonet, T., Dulermo, R., Schott, S. & Palladino, F. Antagonistic functions of SET-2/SET1 and HPL/HP1 proteins in *C. elegans* development. *Dev. Biol.* **312**, 367–383 (2007).
- Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25**, 15–30 (2007).
- Schneider, J. *et al.* Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Mol. Cell* **19**, 849–856 (2005).
- Christensen, J. *et al.* RBP2 belongs to a family of demethylases, specific for tri- and dimethylated lysine 4 on histone 3. *Cell* **128**, 1063–1076 (2007).
- Xu, L. & Strome, S. Depletion of a novel SET-domain protein enhances the sterility of *mes-3* and *mes-4* mutants of *Caenorhabditis elegans*. *Genetics* **159**, 1019–1029 (2001).
- Arantes-Oliveira, N., Apfeld, J., Dillin, A. & Kenyon, C. Regulation of life-span by germ-line stem cells in *Caenorhabditis elegans*. *Science* **295**, 502–505 (2002).
- Berman, J. R. & Kenyon, C. Germ-cell loss extends *C. elegans* life span through regulation of DAF-16 by *kri-1* and lipophilic-hormone signaling. *Cell* **124**, 1055–1068 (2006).
- Budovskaya, Y. V. *et al.* An *elt-3/elt-5/elt-6* GATA transcription circuit guides aging in *C. elegans*. *Cell* **134**, 291–303 (2008).
- Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
- Kelly, W. G., Xu, S., Montgomery, M. K. & Fire, A. Distinct requirements for somatic and germline expression of a generally expressed *Caenorhabditis elegans* gene. *Genetics* **146**, 227–238 (1997).
- Mitchell, D. H., Stiles, J. W., Santelli, J. & Sanadi, D. R. Synchronous growth and aging of *Caenorhabditis elegans* in the presence of fluorodeoxyuridine. *J. Gerontol.* **34**, 28–36 (1979).
- Lee, S. S., Kennedy, S., Tolonen, A. C. & Ruvkun, G. DAF-16 target genes that control *C. elegans* life-span and metabolism. *Science* **300**, 644–647 (2003).
- Haag, E. S., Wang, S. & Kimble, J. Rapid coevolution of the nematode sex-determining genes *fem-3* and *tra-2*. *Curr. Biol.* **12**, 2035–2041 (2002).
- Gerisch, B., Weitzel, C., Kober-Eisermann, C., Rottiers, V. & Antebi, A. A hormonal signaling pathway influencing *C. elegans* metabolism, reproductive development, and life span. *Dev. Cell* **1**, 841–851 (2001).
- Greer, E. L. *et al.* An AMPK–FOXO pathway mediates longevity induced by a novel method of dietary restriction in *C. elegans*. *Curr. Biol.* **17**, 1646–1656 (2007).

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

Acknowledgements We are grateful to A. Fire, K. Helin, S. Kim, K. Shen, T. Stiernagle and the *Caenorhabditis* Genetics Center, M. W. Tan and A. Villeneuve for gifts of strains, reagents and antibodies. We thank S. Kim, G. Seydoux, C. Slightam and K. Shen for advice on worm transgenesis, and S. Kim, A. Morgan and Y. Kobayashi for help with microarray analysis. We thank A. Fire, S. Kim, G. Seydoux, M. W. Tan and J. Wysocka for discussions. We thank members of the Brunet laboratory, M. Kaeberlein, J. Lieberman, J. Sage and J. Wysocka for critical reading of the manuscript. This work was supported by NIH grant R01-AG31198 to A.B.; E.L.G. was supported by NIH training grant T32-CA009302, an NSF graduate fellowship and by NIH ARRA-AG31198. T.J.M. was supported by NIH grant T32-HG000044. D.S.L. and E.M.G. were supported by NIH training grant T32-CA009302. S.H. was supported by a Stanford graduate fellowship. G.S.M. was supported by a Human Frontier Science Program post-doctoral fellowship. M.R.B. was supported by NIH fellowship F31-AG032837. O.G. was supported by a Searle Scholar award.

Author Contributions E.L.G. and A.B. conceived and planned the study. E.L.G. performed the experiments and wrote the paper with the help of A.B.; T.J.M. completed Fig. 3a, b, Supplementary Fig. 4c, e and Supplementary Fig. 8c, and generated all low-copy integrant transgenic worm lines. A.G.H. helped with Fig. 1b, e and Supplementary Fig. 1f. E.M.G. was advised by O.G. and completed Fig. 1f and Supplementary Fig. 1g. D.S.L. helped with Fig. 2a. G.S.M. generated all high-copy transgenic worm lines. S.H. helped with Supplementary Fig. 8c. M.R.B. generated the *P_{rbr-2}::rbr-2::gfp* construct. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.B. (anne.brunet@stanford.edu).

METHODS

C. elegans strains. N2 and *daf-16(mu86)* strains were provided by M. W. Tan. *glp-1(e2142ts)* and *rff-1(pk1417)* strains were provided by A. Fire. *unc-119(ed3)* strain was provided by S. Kim. *wdr-5/tag-125(ok1417)*, *set-2(ok952)*, *rbr-2(tm1231)*, *fem-3(e2006ts)*, *pgl-1(bn101ts)*, *glp-4(bn2ts)*, *glp-1(e2141ts)* and *daf-12(m20)* strains were provided by T. Stiernagle and the *Caenorhabditis* Genetics Center. Double mutants were constructed by crossing *rbr-2(tm1231)* males with *glp-1(e2141ts)* hermaphrodites, *set-2(ok952)* males with *rbr-2(tm1231)* hermaphrodites, *wdr-5/tag-125(ok1417)* males with *rbr-2(tm1231)* or *pgl-1(bn101ts)* hermaphrodites. *P_{rbr-2::rbr-2::gfp}* high-copy or low-copy integrant transgenic lines contain 2.3 kilobases of the *rbr-2* promoter driving the full-length *rbr-2* cDNA fused to GFP at the carboxy-terminus. *P_{ash-2::ash-2::gfp}* high-copy or low-copy integrant transgenic lines contain 1.6 kb of the *ash-2* promoter driving the full-length *ash-2* cDNA fused to GFP at the C terminus. *P_{pie-1::rbr-2::gfp}* low-copy integrant transgenic line contains 2.4 kb of the germline-specific *pie-1* promoter³¹ driving the full-length *rbr-2* cDNA fused to GFP at the C terminus. *P_{pie-1::gfp}* low-copy integrant transgenic line contains 2.4 kb of the germline-specific *pie-1* promoter driving GFP.

RNA interference. *Escherichia coli* HT115 (DE3) transformed with vectors expressing dsRNA of the genes of interest were all obtained from the Ahringer library (a gift from M. W. Tan), except RNAi to *rbr-2*, *blmp-1*, *set-4* and *set-18* that were from the Open Biosystems library (a gift from K. Shen) and were grown at 37 °C and seeded on standard nematode growth medium (NGM) plates containing ampicillin (100 mg ml⁻¹) and isopropylthiogalactoside (IPTG; 0.4 mM). Adult worms were placed on standard NGM plates and removed after 4–6 h to obtain synchronized populations of worms. L1 worms obtained from these synchronized populations were placed on NGM plates containing ampicillin (100 mg ml⁻¹) and IPTG (0.4 mM) seeded with the respective bacteria. Worms placed on RNAi at different time points were placed on empty vector control RNAi at the L1 stage and shifted to the respective bacteria at the appropriate time.

Constructs. pSM, a derivative of the pPD49.26 plasmid³² with additional cloning sites was used to generate transgenic constructs. *P_{rbr-2::rbr-2::gfp}* transgenic plasmid included 2.3 kb of *rbr-2* promoter driving the full-length *rbr-2* cDNA fused to GFP at the C terminus. The *rbr-2* promoter was amplified by PCR from wild-type (N2) genomic DNA using the following primers, F: 5'-ATAGCGGCGGCC CAACATTCTCTGGTGTATG-3' and R: 5'-TATGGATCCTTTTCAGTTTG GCTTATGCTTCATGAG-3' and subcloned into pSM between NotI and BamHI. The *rbr-2* cDNA was amplified by PCR from *rbr-2* cDNA¹⁹ with the following primers F: 5'-ATAGCTAGCATGCGTGCACGTCGTCAGAGAAG AG-3' and R: 5'-TATACCGGTAAATCGGTGGAACACTCGAAGTTGGAGA GTCC-3' and subcloned into pSM containing the *rbr-2* promoter, between NheI and AgeI.

P_{ash-2::ash-2::gfp} transgenic plasmid included 1.6 kb of *ash-2* promoter driving the full-length *ash-2* cDNA fused to GFP at the C terminus. The *ash-2* promoter was amplified by PCR from wild-type (N2) genomic DNA using the following primers, F: 5'-ATAGCTAGCATCTCGTTCTTCTCTGCTCAGCG-3' and R: 5'-TATGTCGACAGTGTGAAATCTAGTTTTTTAG-3' and subcloned into pSM between NheI and AclI. The *ash-2* cDNA isoform A (*C. elegans* ORFeome library, Open Biosystems) was amplified by PCR with the following primers, F: 5'-ATAGTCGACATGAGAAGCTCAAAAGGAGGTCGGGG-3' and R: 5'-TA TGGTACCATAATTTCTGTTTAATTTCTTTC-3' and cloned into pSM containing the *ash-2* promoter, between AclI and KpnI. To generate the *P_{pie-1::rbr-2::gfp}* biolistic bombardment plasmid, *rbr-2::gfp* was subcloned between SpeI and XmaI into the pIC26 plasmid³³ containing 2.4 kb of the *pie-1* promoter³¹. To generate the *P_{ash-2::ash-2::gfp}* and *P_{rbr-2::rbr-2::gfp}* biolistic bombardment plasmids, the *unc-119* rescue fragment and a PacI linker: 5'-GGGGTTAA TTAACCCCGC-3' were subcloned between SpeI and PacI in pSM *P_{ash-2::ash-2::gfp}* and pSM *P_{rbr-2::rbr-2::gfp}*. To generate a bacterially expressed form of SET-2, the SET domain of the *set-2* cDNA isoform B (*C. elegans* ORFeome library, Open Biosystems) was amplified by PCR with the following primers, F: 5'-ATAGGATCCCAACGTCGATTGCTACGTCACCTGGTG-3' and R: 5'-TATCTCGAGTCAATTAAGATATCCACGACACGCTCTTCGC-3' and cloned into pGEX 4T1 (Promega) between BamHI and XhoI. To generate a bacterially-expressed fusion protein between maltose-binding protein (MBP) and ASH-2, full-length *ash-2* cDNA was amplified by PCR with the following primers, F: 5'-ATAGAATTCATGAGAAGCTCAAAAGGAGGTCGGGGACG-3' and R: 5'-TATTTCTAGATTAAATTTCTGTTTAATTTCTTTCTTG-3' and then subcloned between EcoRI and XbaI into the pMCH vector³⁴. All the fragments generated by PCR were entirely sequenced to verify that there were no mutations introduced by the PCR amplification steps.

High-copy transgenic worm generation. Extrachromosomal-array-carrying transgenic strains were generated using a microinjector, injecting 10 ng µl⁻¹ of the transgene plasmid and 60 ng µl⁻¹ of the co-injection plasmid *P_{odr-1::dsred}*

(ref. 35). Two or three independent transgenic lines for each transgene were examined for fluorescence expression and lifespan assays. Fluorescence images were obtained using a confocal microscope Zeiss LSM 510 META.

Low-copy integrant transgenic worm generation by biolistic transformation. Transgenic worm strains were generated using standard biolistic transformation³⁶ by bombarding *unc-119(ed3)* worms with the constructs containing the indicated cDNA fused to GFP driven by either endogenous promoters or by the *pie-1* germline specific promoter, as well as the *unc-119* rescue gene. Approximately 2 weeks after bombardment, worms in which the uncoordinated phenotype was rescued were picked to individual NGM plates. Progeny were kept at 25 °C for 3–4 days and were examined for GFP genomic integration by PCR. GFP-positive worms were selected and expanded³⁷. Two to three independent lines were generated per construct.

Quantitative RT-PCR. Two hundred worms were picked to NGM plates with *E. coli* OP50 overnight two days in a row. Worms were then picked to NGM plates without bacteria and washed three times with M9 buffer (22 mM KH₂PO₄, 34 mM K₂HPO₄, 86 mM NaCl, 1 mM MgSO₄). Worm pellets were resuspended in TRIzol (Invitrogen), followed by six freeze-thaw cycles in liquid N₂. One microgram of total RNA was reverse-transcribed with oligo dT primers using Superscript II reverse transcriptase (Invitrogen) according to the manufacturer's protocol. Real-time PCR was performed on a Bio-Rad iCycler using iQ SYBR green (Bio-Rad) with the following primers: pan-actin F: 5'-TCGGTATGG GACAGAAGGAC-3', pan-actin R: 5'-CATCCAGTTGGTGACGATA-3', *ash-2* F: 5'-CGATCGAAACACGGAACGA-3', *ash-2* R: 5'-TGCCGGAATCTGC AGTTTTT-3'. The experiments were conducted in duplicate and the results were expressed as 2^{-(target gene number of cycles - pan-actin number of cycles)}.

Microarray analysis. One-class analysis in significance analysis of microarrays (SAM)³⁸ was performed comparing empty vector to *ash-2* RNAi using a 5% false discovery rate (FDR). Resulting SAM lists of genes regulated by ASH-2 were compared for genes that were specifically changed in N2, but not in *glp-1(e2141ts)* mutant worms. A complete linkage hierarchical clustering was performed using Gene Cluster 3.0 (ref. 39). Clustering results were analysed further with Java Treeview⁴⁰. GOstat⁴¹ was used to identify overrepresented gene ontology categories in the list of ASH-2 regulated genes that were dependent on the presence of the germline.

Antibodies. The monoclonal ASH-2 antibody was generated by injection of a fusion protein between GST and amino acids 285–572 of *C. elegans* ASH-2 and was affinity-purified by Abmart (Shanghai, China). The H3K4me3 antibody (8580), the H3K4me1 antibody (8895), the H3K9me3 antibody (8898) and the H3 antibody (1791) were obtained from Abcam. The H3K4me2 antibody (07-030) was obtained from Upstate Biotechnology. The affinity-purified monoclonal GFP antibody was a gift from A. Villeneuve.

Whole-mount immunofluorescence. Worms were washed several times to remove bacteria and resuspended in fixing solution (160 mM KCl, 100 mM Tris-HCl pH 7.4, 40 mM NaCl, 20 mM Na₂EGTA, 1 mM EDTA, 10 mM spermidine HCl, 30 mM PIPES pH 7.4, 1% Triton X-100, 50% methanol, 2% formaldehyde) and subjected to two rounds of snap freezing in liquid N₂. The worms were fixed at 4 °C for 30 min and washed briefly in T buffer (100 mM Tris-HCl pH 7.4, 1 mM EDTA, 1% Triton X-100) before a 1 h incubation in T buffer supplemented with 1% β-mercaptoethanol at 37 °C. The worms were washed with borate buffer (25 mM H₃BO₃, 12.5 mM NaOH pH 9.5) and then incubated in borate buffer containing 10 mM DTT for 15 min. Worms were blocked in PBST (PBS pH 7.4, 0.5% Triton X-100, 1 mM EDTA) containing 1% BSA for 30 min and incubated overnight with H3K4me3 antibody (1:100), the monoclonal ASH-2 antibody (1:100), or with an affinity-purified GFP antibody (1:1,000) and with Alexa Fluor 594 secondary antibody (1:25–1:100). 4',6-diamidino-2-phenylindole (DAPI; 2 mg ml⁻¹) was added to visualize nuclei. The worms were mounted on a microscope slide and visualized using a Leica SP2 confocal system or a Zeiss Axioskop2 plus fluorescence microscope.

Protein analysis by western blot. Worms were grown synchronously to appropriate stages and washed off plates with M9 buffer (22 mM KH₂PO₄, 34 mM K₂HPO₄, 86 mM NaCl, 1 mM MgSO₄). Worms were washed several times in M9 buffer and snap frozen in liquid N₂. Sample buffer (2.36% SDS, 9.43% glycerol, 5% β-mercaptoethanol, 0.0945 M Tris-HCl pH 6.8, 0.001% bromophenol blue) was added to worm pellets and they were repeatedly snap frozen in liquid N₂. Worm extracts were sonicated three times for 30 s at ~15 W (VirSonic 600) and boiled for 2 min before being resolved on SDS-PAGE (14%) and transferred to nitrocellulose membranes. The membranes were incubated with primary antibodies (H3K4me3, 1:1,000; H3K4me2, 1:2,000; H3K4me1, 1:500; H3K9me3, 1:1,000; H3, 1:2,000) and the primary antibodies were visualized using horseradish peroxidase-conjugated anti-rabbit secondary antibody (Calbiochem 401393) and ECL Plus (Amersham Biosciences).

In vitro methylation assays. Methyltransferase assays were performed as described previously⁴². Briefly, 10 µg of glutathione S-transferase (GST)-purified SET-2_{SET} was incubated with nucleosomes purified from HeLa cells in the presence of either 0.1 mM S-adenosyl-methionine (SAM) or 2 µCi [³H]SAM at 20 °C overnight in a

methyltransferase reaction buffer (50 mM Tris-HCl pH 8.0, 10% glycerol, 20 mM KCl, 5 mM MgCl₂, 1 mM dithiothreitol (DTT) and 1 mM phenylmethylsulphonyl fluoride (PMSF)). Reactions were subjected to SDS-PAGE and either autoradiography or western blot as described above.

Fertility assays. Wild-type (N2) worms treated with empty vector or *ash-2* RNAi bacteria or *set-2(ok952)*, *wdr-5(tag-125(ok1417))*, or *rbr-2(tm1231)* mutant worms were synchronized. From day 3 to day 8 of life, 10 worms were placed on NGM plates with OP50-1 in triplicate (30 worms total per condition). After 24 h, the adult worms were removed from each plate and the number of eggs was counted. For *ash-2* RNAi, the number of live progeny was also counted after 48 h. Statistical analyses of fertility were performed using two-way ANOVA tests with Bonferroni post-tests, or *t*-tests using the mean and standard error values.

Germline cell morphology. N2 worms were grown until either day 4 or day 8 on plates with bacteria expressing empty vector or *ash-2* RNAi. Worms were then fixed, permeabilized and DAPI-stained as described above. Worms were imaged on a Leica SP2 confocal microscope. Z-sections of germ cell nuclei were imaged, compiled to stacks and manually counted in a blinded manner using MetaView imaging software. For each condition, distal gonad germ cell numbers from three worms were counted and the results of two independent experiments were analysed using a two-tailed Student's *t*-test.

31. Mello, C. C. *et al.* The PIE-1 protein and germline specification in *C. elegans* embryos. *Nature* **382**, 710–712 (1996).
32. Fire, A., Harrison, S. W. & Dixon, D. A modular set of *lacZ* fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene* **93**, 189–190 (1990).
33. Cheeseman, I. M. *et al.* A conserved protein network controls assembly of the outer kinetochore and its ability to sustain tension. *Genes Dev.* **18**, 2255–2268 (2004).
34. Venteicher, A. S., Meng, Z., Mason, P. J., Veenstra, T. D. & Artandi, S. E. Identification of ATPases pontin and reptin as telomerase components essential for holoenzyme assembly. *Cell* **132**, 945–957 (2008).
35. Klassen, M. P. & Shen, K. Wnt signaling positions neuromuscular connectivity by inhibiting synapse formation in *C. elegans*. *Cell* **130**, 704–716 (2007).
36. Praitis, V., Casey, E., Collar, D. & Austin, J. Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics* **157**, 1217–1226 (2001).
37. Merritt, C. & Seydoux, G. In *WormBook* (ed. The *C. elegans* Research Community) doi:10.1895/wormbook.1.148.1 (<http://www.wormbook.org/>) (2010).
38. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
39. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
40. Saldanha, A. J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
41. Beissbarth, T. & Speed, T. P. GStat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465 (2004).
42. Shi, X. *et al.* Modulation of p53 function by SET8-mediated methylation at lysine 382. *Mol. Cell* **27**, 636–646 (2007).

LETTERS

Relationship between nucleosome positioning and DNA methylation

Ramakrishna K. Chodavarapu^{1*}, Suhua Feng^{1,2*}, Yana V. Bernatavichute^{1,3}, Pao-Yang Chen¹, Hume Stroud¹, Yanchun Yu⁴, Jonathan A. Hetzel¹, Frank Kuo¹, Jin Kim¹, Shawn J. Cokus¹, David Casero¹, Maria Bernal⁵, Peter Huijser⁶, Amander T. Clark^{1,7}, Ute Krämer^{5†}, Sabeeha S. Merchant^{3,8}, Xiaoyu Zhang⁴, Steven E. Jacobsen^{1,2,3} & Matteo Pellegrini^{1,3}

Nucleosomes compact and regulate access to DNA in the nucleus, and are composed of approximately 147 bases of DNA wrapped around a histone octamer^{1,2}. Here we report a genome-wide nucleosome positioning analysis of *Arabidopsis thaliana* using massively parallel sequencing of mononucleosomes. By combining this data with profiles of DNA methylation at single base resolution, we identified 10-base periodicities in the DNA methylation status of nucleosome-bound DNA and found that nucleosomal DNA was more highly methylated than flanking DNA. These results indicate that nucleosome positioning influences DNA methylation patterning throughout the genome and that DNA methyltransferases preferentially target nucleosome-bound DNA. We also observed similar trends in human nucleosomal DNA, indicating that the relationships between nucleosomes and DNA methyltransferases are conserved. Finally, as has been observed in animals, nucleosomes were highly enriched on exons, and preferentially positioned at intron–exon and exon–intron boundaries. RNA polymerase II (Pol II) was also enriched on exons relative to introns, consistent with the hypothesis that nucleosome positioning regulates Pol II processivity. DNA methylation is also enriched on exons, consistent with the targeting of DNA methylation to nucleosomes, and suggesting a role for DNA methylation in exon definition.

To investigate the position of nucleosomes in *Arabidopsis thaliana*, we sequenced micrococcal nuclease (MNase)-digested nucleosomal DNA using an Illumina GAII sequencer to achieve a roughly 68-fold coverage of nucleosomes (see Supplementary Methods). The data are displayed in a modified UCSC genome browser (<http://epigenomics.mcdb.ucla.edu/Nuc-Seq/>) along with nucleosome density tracks that allow easy visualization of nucleosome positions throughout the genome (Fig. 1a).

To obtain a chromosomal view of nucleosome content, we plotted reads in bins of 100 kilobases tiling the chromosomes. As a control for biases in mapping and sequencing we also sequenced a library of randomly sheared *Arabidopsis* genomic DNA. We then normalized the nucleosome counts by the number of uniquely mapping genomic DNA counts within each bin along the chromosomes. Nucleosome content was relatively uniform throughout the euchromatic regions of chromosomes, but showed significant enrichment in pericentromeric heterochromatin regions, (Fig. 1b, Supplementary Fig. 1). To confirm these results, we performed chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) using an

antibody against unmodified histone H3 (Fig. 1b and Supplementary Fig. 1). These observations indicate that nucleosomes are more densely packed in pericentromeric regions that are transcriptionally silent, are rich in transposons and contain heavily methylated DNA, than in the euchromatic arms³.

By examining the relationship between reads that map to opposite strands of DNA, we observed a peak of reads on the reverse strand that occurred approximately 145–170 bases downstream of the reads on the forward strand (Supplementary Fig. 2a). This broad peak in the general vicinity of the accepted size of a nucleosome indicates that many nucleosomes are well positioned, leading to the repeated sequencing of both the forward and reverse complement of the corresponding nucleosome regions. Similarly, when we plotted the correlation between positive strand reads with other positive strand reads we saw a progressively decreasing correlation from the start of the nucleosome (Fig. 1c and Supplementary Fig. 2b), with smaller peaks spaced at 174 and 355 base pairs from the starting position of the reference reads indicating some preference for regular spacing of nucleosomes genome-wide. Assuming that nucleosomes are wrapped around 147 base pairs of DNA, the average length of the linkers of regularly spaced nucleosomes is about 30 base pairs. Consistent with the higher content of nucleosomes in pericentromeric heterochromatin regions, we found that the peaks at 174 and 355 were more pronounced in these regions than in the euchromatic arms (Fig. 1c), indicating that pericentromeric nucleosomes have a higher tendency to be in regularly spaced arrays. These results are consistent with earlier evidence for more regular spacing of nucleosomes in *Drosophila* heterochromatin⁴.

Similar to findings in animal and fungal high-throughput nucleosome sequencing studies^{5–8}, we found 10-base periodicities in WW (W = A or T) dinucleotides, and SS (S = G or C) dinucleotides that were 5 bases out of phase with the WW dinucleotides (Fig. 1d and Supplementary Figs 3–5). WW dinucleotides are favoured at sites where the minor groove faces the histone core, whereas SS dinucleotides are favoured at sites where the minor groove faces away from the histone core^{9,10}. The out of phase peaks in the frequencies of these dinucleotides leads to optimal bending of DNA, as A/T nucleotides cause a negative base roll and G/C nucleotides cause a positive base roll^{9,10}.

To study the relationship of DNA methylation with nucleosome positions, we used our single-nucleotide resolution whole-genome bisulphite-sequencing data³. *Arabidopsis* cytosines are methylated by

¹Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California 90095, USA. ²Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, California 90095, USA. ³Molecular Biology Institute, University of California Los Angeles, Los Angeles, California 90095, USA. ⁴Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA. ⁵University of Heidelberg, BIOQUANT 23, R. 645, Im Neuenheimer Feld 267, D-69120 Heidelberg, Germany. ⁶Department of Molecular Plant Genetics, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. ⁷Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California 90095, USA. ⁸Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, USA. [†]Present address: Department of Plant Physiology, University of Bochum, Universitätsstrasse 150, D-44801 Bochum, Germany.

*These authors contributed equally to this work.

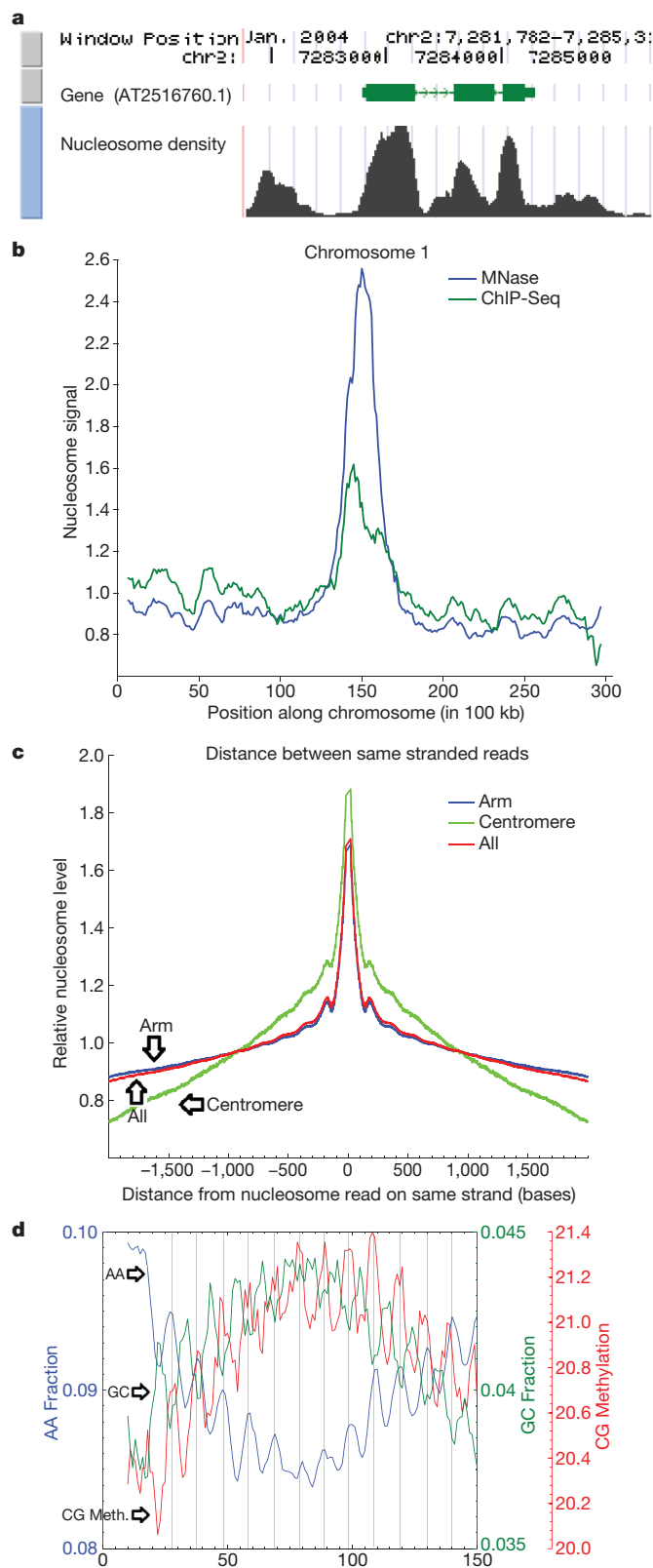


Figure 1 | Characterization of *A. thaliana* nucleosome data.

a, Representative UCSC Browser screenshot of a gene showing calculated nucleosome densities. **b**, Chromosomal view of nucleosome read counts and ChIP-seq read counts in 100-kb tiles along chromosome 1 shows nucleosome enrichment in pericentromeric regions. **c**, Autocorrelation of positive stranded reads shows local peaks at positions 174 and 355 bases. **d**, AA and GC dinucleotide and CG methylation profiles show a 10-base periodicity over nucleosome-bound DNA, with DNA methylation profiles in phase with WW dinucleotide profiles, and out of phase with SS dinucleotide profiles.

one of three different DNA methyltransferases depending on their sequence context. CG sites are methylated by MET1, and CHG sites (where H is A, C or T) are methylated by CMT3. Finally, CHH sites are methylated by DRM2, a *de novo* methyltransferase that is targeted by small RNAs¹¹. Notably, all three types of methylation showed a 10-base periodicity on nucleosomal DNA, which was in phase with the WW dinucleotides and out of phase with the SS dinucleotides (Figs 1d, 2a–c and Supplementary Fig. 6). These methylation preferences were not correlated with preferences for CG, CHG or CHH sequences at these locations (Supplementary Fig. 7). Because DNA methyltransferases access the major groove, this methylation would be on DNA that is on the outside of the nucleosome (minor groove facing the histones) and thus more accessible to the DNA methyltransferases. This in turn indicates that DNA might be in part methylated as the DNA is still bound to nucleosomes, leading to the observed 10-base pair periodicity. It has been proposed that chromatin remodelling enzymes are required for DNA methyltransferases to gain access to the DNA, and indeed DRD1, DDM1 and LSH1 are remodellers known to be important factors controlling DNA methylation¹². However, our data indicate that nucleosomal DNA may also be a substrate for DNA methyltransferases *in vivo*, demonstrating a prominent role of the nucleosome in determining methylation patterning throughout the genome.

We previously reported a 10-nucleotide periodicity in CHH methylation data when performing autocorrelation analysis on the methylation pattern of the whole genome³. Our previous interpretation was that the structure of the DRM2 enzyme might be responsible for this pattern, because the orthologous Dnmt3 enzymes in mammals are known to act as heteromeric complexes in which two methyltransferase active sites have a spacing equivalent to roughly 10 nucleotides of DNA¹³. However, the current data in which we see 10-base pair periodicities for all types of methylation indicate a more general explanation: that nucleosomes are to some extent dictating access to the DNA and therefore setting the register of methylation for all DNA methyltransferases. Nucleosomal preferences could also partially explain the sequence preferences that we observed previously for CHG and CHH methylation³. Highly methylated cytosines tended to be followed immediately by A/T but not C, consistent with our finding that DNA methylation is out of phase with CC dinucleotides (Supplementary Fig. 6).

On a larger scale, we also observed that levels of all three types of DNA methylation were higher in nucleosome-spanning DNA than in flanking DNA, indicating that nucleosome-bound DNA is enriched for DNA methylation (Fig. 2 and Supplementary Fig. 8). This finding supports the view that nucleosomes are preferentially targeted by DNA methyltransferases. In the case of CMT3, it is predicted that this enzyme is recruited or activated by histone H3 lysine 9 dimethylation, because a knockout of histone H3 lysine 9 methyltransferase mimics a knock out of CMT3, and because the CMT3 chromodomain can bind to methylated histones¹⁴. However, our data show that all types of methylation are enriched on nucleosome-bound DNA, indicating that nucleosomes or histone modifications may assist in the recruitment of all of the *Arabidopsis* DNA methyltransferases.

To test whether the patterns of DNA methylation on nucleosomal DNA are also found in humans, we used previously published MNase nucleosome sequencing data¹⁵ together with our single-nucleotide resolution bisulphite sequencing data on the human embryonic stem cell line HSF1 (see Supplementary Methods). We found that human nucleosome-bound DNA also showed a 10-base periodicity in its CG, CHG and CHH methylation status (Fig. 3). Furthermore, as in *Arabidopsis*, the overall level of methylation was higher on nucleosome-bound DNA than in flanking regions. This indicates that nucleosome architecture has a role in shaping DNA methylation patterning in the human genome, and is consistent with the recent finding of stable anchoring of Dnmt3 DNA methyltransferases to mammalian chromatin¹⁶ and more generally with the conservation of DNA methyltransferase function in plants and animals¹⁷. We also

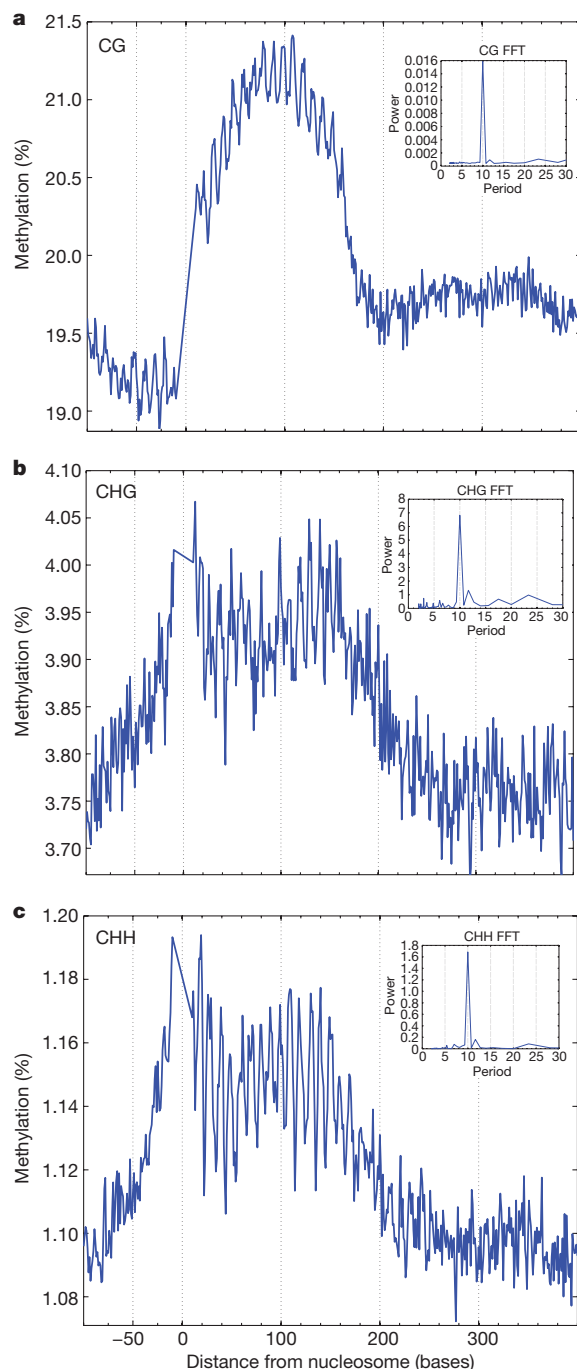


Figure 2 | DNA methylation profiles of nucleosome-bound DNA in *Arabidopsis*. The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a–c**, CG methylation (**a**), CHG methylation (**b**) and CHH methylation (**c**) each show a 10-base periodicity over nucleosome-bound DNA (1–147 bases). Fast Fourier transforms (FFTs) can be used to deconstruct inherent frequencies in a complex signal. The FFTs calculated over the region of the nucleosome (insets) demonstrate this periodicity in CG (**a**), CHG (**b**) and CHH (**c**) contexts.

analysed DNA methylation patterning on nucleosomes in different regions of the *Arabidopsis* and human genomes, including genes, promoters, pericentromeric regions and euchromatic arm regions (Supplementary Figs 9–17) and found that the 10-nucleotide periodicity was found in all cases, indicating that the relationship between nucleosome positioning and DNA methylation is general.

We observed that nucleosomes were much more abundant in exons than in introns (see Fig. 1a for an example), consistent with

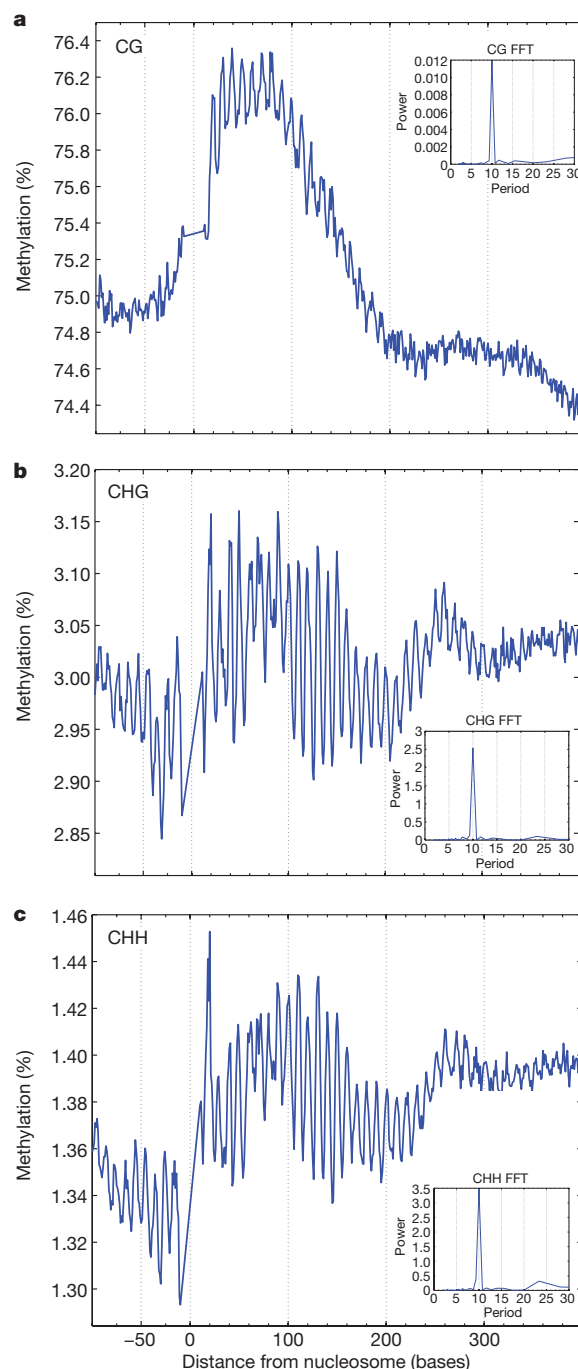


Figure 3 | DNA methylation profiles of nucleosome-bound DNA in humans. The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a–c**, CG methylation (**a**), CHG methylation (**b**) and CHH methylation (**c**) each show a 10-base periodicity over nucleosome-bound DNA (1–147 bases). The FFTs calculated over the region of the nucleosome (insets) demonstrate this periodicity in CG (**a**), CHG (**b**), and CHH (**c**) contexts.

findings from several recent nucleosome positioning studies in other organisms^{18–23}. The number of nucleosomes per base pair in introns was only 63% of the level found in exons. Furthermore, we found a strong peak of nucleosome start sites at intron–exon junctions and at exon–intron junctions (Fig. 4, Supplementary Fig. 18). The enrichment of nucleosomes in exons was not solely owing to higher G+C content (Supplementary Fig. 19a–c), or because of consensus splice site sequences (Supplemental Fig. 19d–f), and was confirmed using independent chromatin immunoprecipitation methods (Supplementary Fig. 20). Longer exons contained a higher number of

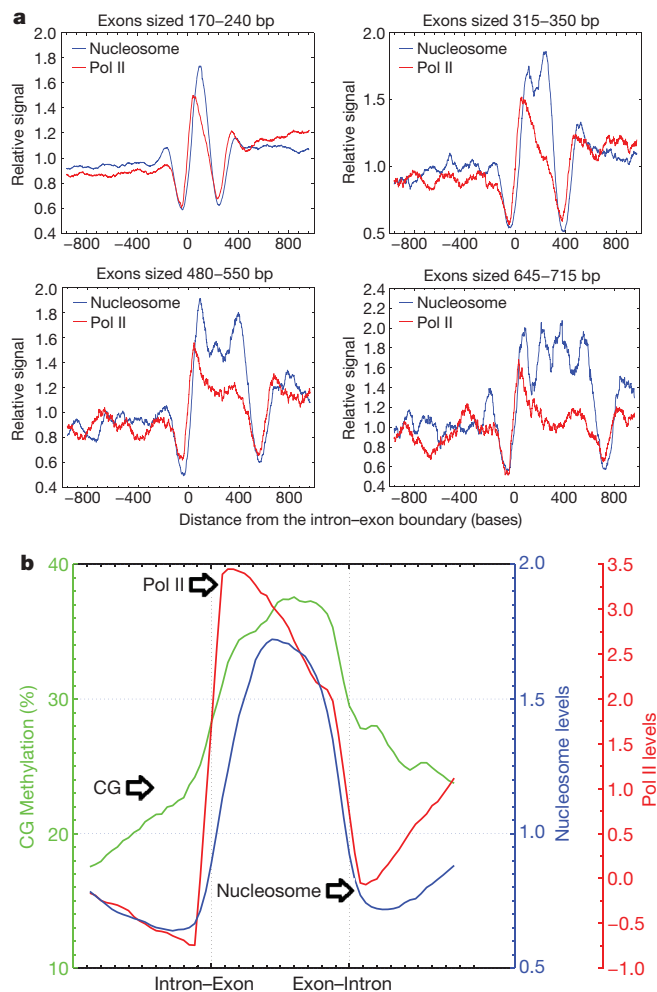


Figure 4 | Nucleosome and Pol II levels in exons. We performed chromatin immunoprecipitation with an antibody against RNA Pol II, and hybridized the resulting DNA to a whole genome Affymetrix microarray. We normalized these data to randomly sheared genomic DNA to control for probe efficiencies. **a**, Nucleosomes are phased in exons. Exon size limits were selected such that the exon could house no more than 1, 2, 3 or 4 nucleosomes. Nucleosome midpoints are plotted over these intron-exon boundaries. **b**, Each exon was divided into 25 equal sized bins and the nucleosome midpoints, Pol II and CG methylation levels are plotted over the exon and flanking introns.

nucleosomes. Examination of exons in the size ranges of 170–240, 315–350, 480–550 and 645–715 base pairs revealed peaks of one, two, three or four well-positioned nucleosomes respectively (Fig. 4a), indicating that nucleosomes are phased within exons.

Strong nucleosome enrichment on exons was detected both for genes that are highly expressed and for those not expressed (Supplementary Fig. 21 and Supplementary Table 1), suggesting that DNA sequences position nucleosomes in the absence of active transcription and splicing. Consistent with this hypothesis, using a nucleosome positioning prediction algorithm²⁴ we found similar patterns of nucleosome positioning in introns and exons between the experimental and theoretical data sets (Supplementary Fig. 22). Similarly, using theoretically predicted nucleosome positions, we observed similar patterns of DNA methylation on nucleosome-bound DNA (Supplementary Fig. 23), as well as enrichment of predicted nucleosomes in pericentromeric regions (Supplementary Fig. 24).

Because nucleosomes present a barrier to RNA Pol II transcription, we tested for Pol II occupancy in exons using a chromatin immunoprecipitation microarray approach. We observed significant enrichment of Pol II in exons relative to introns, consistent with the

hypothesis that Pol II is paused on exonic DNA (Fig. 4). One possibility is that Pol II stalling on exons could enhance accurate splicing of upstream introns, thus reducing exon skipping and aiding in the fidelity of exon definition²⁰. This is consistent with the finding of Pol II enrichment on human exons and indicates that Pol II enrichment on exons might be a common eukaryotic feature²⁰. Furthermore, particular histone modifications have been recently shown to recruit splicing regulators, providing additional possible mechanisms for the regulation of splicing by nucleosome positioning²⁵.

Because of the enhancement of DNA methylation over nucleosomal DNA, and the enrichment of nucleosomes on exons, a prediction is that DNA methylation should be enhanced on exons relative to introns. Indeed, we found depletion of DNA methylation in introns and enrichment in exons in a pattern that was similar to nucleosome occupancy (Fig. 4b). This is consistent with recent findings of enhanced exonic methylation in other plant species, as well as in the *Chlamydomonas*, honeybee, *Ciona* and human genomes^{26,27}. This suggests the possibility that DNA methylation, which frequently exists in the transcribed regions of active genes^{26,28}, could have a conserved role in exon definition or splicing regulation. These findings also reinforce the view that nucleosomal positions have an important role in shaping the methylation landscape of eukaryotic genomes.

METHODS SUMMARY

Approximately 300 ng of MNase-digested mononucleosome DNA were used for Illumina library generation following manufacturer instructions. Libraries were sequenced using an Illumina Genome Analyzer II following manufacturer instructions. Resulting reads were mapped to the TAIR7 annotation of the *Arabidopsis* genome and reads that mapped to multiple sites were eliminated. A total of 24.2 million reads were mapped to the forward strand and 23.9 million reads to the reverse strand, generating a 68-fold coverage of nucleosome space.

Pol II ChIP-chip was performed as described previously²⁹. Briefly, crosslinked chromatin was extracted, sonicated and used in ChIP with an antibody against Pol II CTD (ab817; Abcam). Pol II-bound DNA and input genomic DNA were extracted, amplified, labelled and hybridized to Affymetrix tiling microarrays as described previously²⁹. Three biological replicates were performed and the log₂ ratios of Pol II-bound DNA over input DNA were calculated using the Tiling Analysis Software (Affymetrix).

Received 5 March; accepted 7 May 2010.

Published online 30 May 2010.

- Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
- Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
- Cokus, S. J. et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
- Sun, F. L., Cuaycong, M. H. & Elgin, S. C. Long-range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. *Mol. Cell. Biol.* **21**, 2867–2879 (2001).
- Albert, I. et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
- Mavrich, T. N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).
- Valouev, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
- Mavrich, T. N. et al. Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).
- Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34**, 269–324 (2001).
- Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
- Henderson, I. R. & Jacobsen, S. E. Epigenetic inheritance in plants. *Nature* **447**, 418–424 (2007).
- Kobor, M. S. & Loricz, M. C. H2A.Z and DNA methylation: irreconcilable differences. *Trends Biochem. Sci.* **34**, 158–161 (2009).
- Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature* **449**, 248–251 (2007).
- Lindroth, A. M. et al. Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *EMBO J.* **23**, 4146–4155 (2004).
- Schones, D. E. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).

16. Jeong, S. *et al.* Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. *Mol. Cell. Biol.* **29**, 5366–5376 (2009).
17. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11**, 204–220 (2010).
18. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
19. Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* **16**, 996–1001 (2009).
20. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).
21. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genet.* **41**, 376–381 (2009).
22. Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* **36**, 245–254 (2009).
23. Ponts, N. *et al.* Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res.* **20**, 228–238 (2010).
24. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
25. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
26. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).
27. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20**, 320–331 (2010).
28. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* advance online publication doi:10.1126/science.1186366 (15 April 2010).
29. Zhang, X. *et al.* Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biol.* **5**, e129 (2007).

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

Acknowledgements We thank members of our laboratories for input and guidance, and the Department of Energy, Institute for Genomics and Proteomics for support. R.K.C. is supported by NIH training grant GM07104. S.F. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. P.-Y.C. is funded by Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA. M.B. was supported by an Alexander-von-Humboldt Fellowship and U.K. by a Heisenberg Fellowship. X.Z. was supported by a Faculty Research Grant (JR-040) from the University of Georgia. Research was supported by an Innovation Award from the Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA, NIH grants GM60398 and GM42143, and NSF Plant Genome Research Program #0701745. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions S.E.J. and M.P. designed the research. S.F., Y.V.B., H.S., Y.Y., J.A.H., M.B., P.H., A.T.C., U.K., S.S.M. and X.Z. performed experiments. R.K.C., P.-Y.C., F.K., J.K., S.J.C. and D.C. analysed data. R.K.C., S. F., S.E.J. and M.P. wrote the paper.

Author Information All sequencing files have been deposited in GEO under accession code GSE21673. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.E.J. (jacobsen@ucla.edu) or M.P. (matteop@mcdb.ucla.edu).

Structure of the gating ring from the human large-conductance Ca^{2+} -gated K^+ channel

Yunkun Wu^{1,2}, Yi Yang¹, Sheng Ye³ & Youxing Jiang^{1,2}

Large-conductance Ca^{2+} -gated K^+ (BK) channels are essential for many biological processes such as smooth muscle contraction and neurotransmitter release^{1–4}. This group of channels can be activated synergistically by both voltage and intracellular Ca^{2+} , with the large carboxy-terminal intracellular portion being responsible for Ca^{2+} sensing^{5–13}. Here we present the crystal structure of the entire cytoplasmic region of the human BK channel in a Ca^{2+} -free state. The structure reveals four intracellular subunits, each comprising two tandem RCK domains, assembled into a gating ring similar to that seen in the MthK channel¹⁴ and probably representing its physiological assembly. Three Ca^{2+} binding sites including the Ca^{2+} bowl are mapped onto the structure based on mutagenesis data. The Ca^{2+} bowl, located within the second RCK domain, forms an EF-hand-like motif and is strategically positioned close to the assembly interface between two subunits. The other two Ca^{2+} (or Mg^{2+}) binding sites, Asp 367 and Glu 374/Glu 399, are located on the first RCK domain. The Asp 367 site has high Ca^{2+} sensitivity and is positioned in the groove between the amino- and carboxy-terminal subdomains of RCK1, whereas the low-affinity Mg^{2+} -binding Glu 374/Glu 399 site is positioned on the upper plateau of the gating ring and close to the membrane. Our structure also contains the linker connecting the transmembrane and intracellular domains, allowing us to dock a voltage-gated K^+ channel pore of known structure onto the gating ring with reasonable accuracy and generate a structural model for the full BK channel.

BK channels, also called MaxiK or Slo1, are a vital class of K^+ channels found throughout the animal kingdom and expressed in a variety of cell types^{1–4}. Several unique biophysical properties, including a large conductance and a gating mechanism involving both voltage and intracellular Ca^{2+} , distinguish them from other K^+ channels. The S1–S4 transmembrane region of BK channels forms a voltage sensing domain as in other voltage-gated K^+ channels, whereas the large C-terminal intracellular ligand-binding part, accounting for two-thirds of the full channel, is responsible for sensing Ca^{2+} (refs 5–13) as well as other intracellular stimuli¹⁵. The latter has been predicted to contain two RCK domains, first identified in an *Escherichia coli* K^+ channel and conserved in the majority of prokaryotic ligand-gated K^+ channels^{14,16} as well as in the bacterial K^+ uptake and efflux machinery (the KTN domain)^{17,18}. Although the first RCK domain of BK shares high sequence similarity with its prokaryotic counterparts, the less homologous second RCK domain is not as well defined. Even though the intracellular ligand-gating properties of BK channels have been extensively studied, little is known about the structural details underlying their ligand gating mechanism beyond the proposed models based on MthK^{16,19,20} and a recent low-resolution structure from electron cryo-microscopy²¹. The structure of the intracellular ligand-binding domain presented here therefore provides essential structural information to guide future studies aimed at understanding how these physiologically

essential channels integrate multiple cellular stimuli to modulate membrane excitability.

A human BK (also called hslol1) intracellular domain construct starting from Arg 329, located right after the pore-lining inner helix, to the C-terminal end (Supplementary Fig. 1) was chosen for structural characterization. Additionally, an extra 32-amino-acid-long sequence from a mutant transcription factor GCN4 leucine zipper (GCN4_LI), known to form a four-helical coiled-coil domain, was inserted before the N terminus of the BK intracellular domain. The protein was also treated with Lambda protein phosphatase before crystallization. Both modifications proved to be essential to obtaining good-quality crystals. The crystals, which were of space group *I*422 and contained one subunit per asymmetric unit, revealed four crystallographically related subunits assembled as a tetramer with the molecular four-fold coinciding with the crystallographic tetrad. The final model was refined to 3.1 Å with R_{work} of 23.8% and R_{free} of 28.9% (see Methods and Supplementary Table 1).

Despite low sequence identity and large differences in sequence length between BK channels and prokaryotic RCK-regulated K^+ channels, the intracellular portion of human BK contains two tandem RCK domains as predicted¹⁴ and its structure is strikingly similar to that of its prokaryotic counterpart, the MthK channel (Fig. 1 and Supplementary Figs 1 and 2). For comparative purposes, secondary structure elements of each BK RCK domain are labelled following the same nomenclature used for its prokaryotic counterparts (Supplementary Fig. 1).

Each RCK domain can be divided into three subdomains: an N-terminal Rossmann-folded subdomain (from βA to βF) that forms the central core of the gating ring; the intermediate helix-crossover domain (αF -turn- αG) that interlocks the two RCK domains; and a C-terminal subdomain that interacts with its counterpart from a neighbouring RCK domain within the same subunit (Fig. 1a, b and Supplementary Fig. 2). The two RCK domains form a bi-lobed architecture with extensive inter-RCK interactions at the helix crossover and C-terminal subdomain, reminiscent of the RCK homodimer in MthK (Fig. 1c, d). Notably, the deep cleft between the two RCK domains of MthK as well as other prokaryotic RCK proteins is occupied in the BK channel mainly by residues from the extended loops between βD and αD of both RCKs (Fig. 1c). Whereas this wide open cleft allows for large conformational changes at the flexible interface at its base in MthK upon ligand binding, its absence in BK raises the question of whether similar conformational changes also occur. Further studies are needed to address this issue.

Four BK intracellular subunits assemble as a gating ring of eight RCK domains. The Rossmann-folded subdomains from eight RCKs form the central core of the gating ring whereas the C-terminal subdomains form four protruded knobs at its periphery (Fig. 2a). The inter-subunit interactions occur at the assembly interface formed by

¹Department of Physiology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9040, USA. ²Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9040, USA. ³Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang 310058, China.

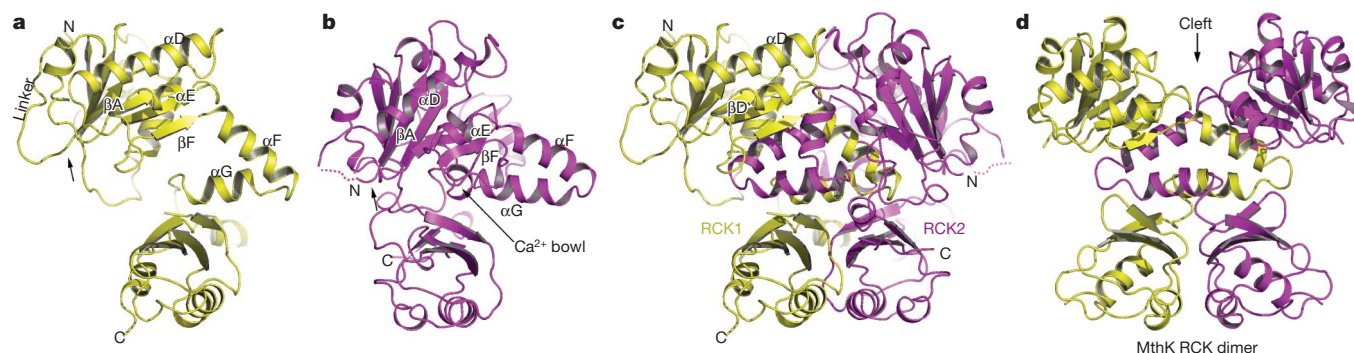


Figure 1 | The BK intracellular subunit contains two RCK domains. **a, b,** The structure of the individual RCK domain is shown with RCK1 in yellow (**a**) and RCK2 in magenta (**b**). Only selected secondary structural elements are labelled for clarity. The arrow indicates the starting point of each RCK, and the linker connects RCK1 to the channel pore. The disordered loop between α D and β E on RCK2 is drawn as a dotted line. **c,** The BK

intracellular subunit is comprised of RCK1 and RCK2, and the orientation of RCK1 shown in the subunit assembly is the same as in **a**. The long loop connecting the C terminus of RCK1 and N terminus of RCK2 is disordered. **d,** The structure of the RCK homodimer in MthK is shown with RCK domains coloured yellow and magenta, respectively. Stereo diagrams of BK domain structures are shown in Supplementary figures.

helices α D and α E from each RCK domain in a head-to-tail manner; that is, RCK1 contacts RCK2 from the neighbouring subunit. The protein–protein contacts at the assembly interface are mainly hydrophobic, involving residues Ile 441, Met 442 and Ile 445 on α D and His 468 on α E of RCK1 and residues Ile 821, Leu 822, Leu 825 and Phe 890 on the equivalent helices of RCK2 from the neighbouring subunit (Fig. 2b and Supplementary Fig. 3). The three residues on α D that form the core of the assembly interface are conserved as hydrophobic in most RCK domains^{14,16}. Interestingly, α Es from both BK RCK domains seem to be 3_{10} helices. Compared with MthK, the BK gating ring is more expanded, as expected for a protein of much larger size. The diagonal distance of the gating ring is about 82 Å, measured between the C α atoms of Arg 342 residues located at the outer rim of the ring (Fig. 2c). The BK gating ring also has a wide central hole with a width of about 20 Å, allowing ions to move freely between the ion conduction pore and the intracellular side. The crystals were obtained in the absence of Ca^{2+} and no bound divalent cations were

observed in the structure. This led us to suggest that the BK gating ring structure shown here might represent a closed conformation.

Although the BK intracellular domain is responsive to various chemical stimuli inside the cell, Ca^{2+} activation has so far been best characterized and will be the focus of the discussion here. Micromolar concentrations of Ca^{2+} can activate voltage-dependent gating of BK channels^{4,22,23}. This high Ca^{2+} sensitivity can be partly attributed to a stretch of well-conserved aspartate residues within the second RCK domain, known as the Ca^{2+} bowl^{6,7}. The Ca^{2+} bowl is located right after α E in the second RCK followed by a short helix (α EF), and is physically in close proximity to the assembly interface (Fig. 3a, b). Our structure was determined in the absence of Ca^{2+} and the stretch of acidic residues in the Ca^{2+} bowl does not form a well-ordered structure. The short helix following the Ca^{2+} bowl is not seen on BK RCK1 and other RCK proteins, and was assigned as α EF because it is between α E and β F. Together with the Ca^{2+} bowl and α E, this helix forms a helix–loop–helix motif (α E-loop- α EF) very similar to Ca^{2+} -binding

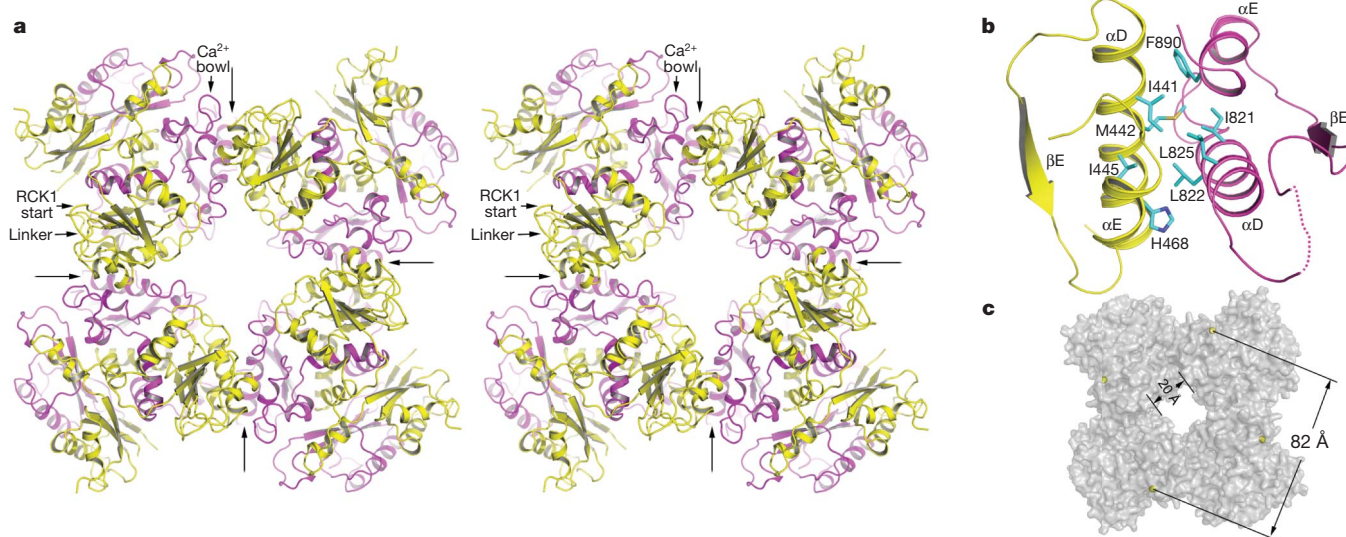


Figure 2 | Four BK intracellular subunits assemble into a gating ring. **a,** Stereo diagram of the gating ring structure viewed down the four-fold axis from the extracellular side. Four long arrows indicate the position of inter-subunit assembly interfaces and define the boundary of each subunit. RCK1 and RCK2 in each subunit are coloured yellow and magenta, respectively. The linker between RCK1 and the channel pore, the starting point of RCK1, and the position of the Ca^{2+} bowl are labelled on one subunit (upper left

corner). **b,** An enlarged view of the assembly interface formed by helices α D and α E from both RCK1 (yellow) and RCK2 (magenta). The side chains of those hydrophobic residues important for protein–protein contacts are also shown. The disordered loop between α D and β E is drawn as a dotted line. **c,** Overall shape of the BK gating ring in surface representation. Yellow spheres represent the C α atoms from Arg 342 residues used to define the diagonal length of the gating ring.

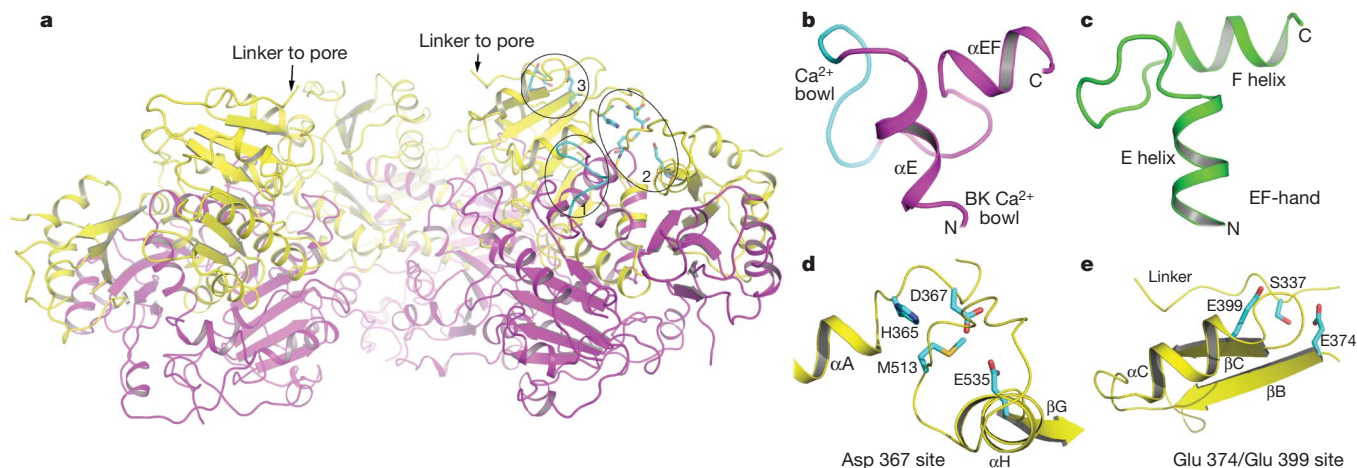


Figure 3 | Ca^{2+} binding sites in the BK intracellular subunit. **a**, A side view of the BK gating ring with the top surface facing the membrane. The front subunit is removed for clarity. The circled positions represent the three Ca^{2+} binding sites on one subunit, and are labelled 1 for the Ca^{2+} bowl, 2 for the Asp 367 site and 3 for the Glu 374/Glu 399 site. RCK1 and RCK2 are coloured yellow and magenta, respectively. **b**, αE , the Ca^{2+} bowl and αEF

form an EF-hand-like motif. The structure of the Ca^{2+} bowl (cyan loop) is disordered in the absence of Ca^{2+} . **c**, The EF-hand structure of a Ca^{2+} binding protein (a human calmodulin molecule with Protein Data Bank code 1CLL). **d**, **e**, Structural details of the Asp 367 and Glu 374/Glu 399 Ca^{2+} binding sites, respectively.

EF-hand proteins such as calmodulin^{24,25} (Fig. 3c), despite αEF being much shorter than the exiting helix (F helix) of most EF-hand proteins. It is likely that the disordered Ca^{2+} bowl structure represents an *apo* state and becomes ordered upon Ca^{2+} binding, as commonly seen in EF-hand proteins which are also rich in acidic residues at the Ca^{2+} binding loop²⁵. The strategic location of the Ca^{2+} bowl beside the inter-subunit assembly interface led us to postulate that Ca^{2+} binding triggers conformational changes at the assembly interface, changing the size of the gating ring and leading to channel activation.

In addition to the Ca^{2+} bowl, recent studies have also identified at least two other independent Ca^{2+} binding sites within the first RCK domain^{8–12}. One such site also has high Ca^{2+} sensitivity and involves an acidic residue, Asp 367, located on the loop between αA and βB . In the context of the gating ring, this site is positioned in the groove between the N- and C-terminal subdomains of RCK1, exposed to the open space between the membrane and the peripheral C-terminal subdomain (Fig. 3a). Figure 3d shows the chemical environment surrounding Asp 367, comprising the side chains of His 365, Met 513 and Glu 535. His 365 has been shown to be involved in pH activation of BK and was proposed to interact electrostatically with Asp 367 upon protonation²⁶, mimicking the effect of Ca^{2+} . The location of Met 513 at this site explains the profound effect on Ca^{2+} sensitivity observed upon mutation⁹. Even though the side chain of Met 513 may not directly bind Ca^{2+} , it may have a key role in maintaining the local structural integrity of the Ca^{2+} binding site. The chemistry of the Asp 367 ion binding site with a limited number of ligands does not seem suitable for high-affinity Ca^{2+} binding. It is likely that residues from other parts of the channel may participate in Ca^{2+} binding. One possibility is residue Glu 535 on the C-terminal subdomain of RCK1, the side chain of which seems to be perfectly positioned to chelate Ca^{2+} together with Asp 367, and may well be involved in Ca^{2+} binding, a hypothesis that warrants further investigation through mutagenesis.

The second Ca^{2+} binding site on RCK1 has lower affinity and has been suggested to be physiologically regulated by Mg^{2+} (refs 10, 11, 27). This low-affinity site formed by Glu 374 and Glu 399 located on two neighbouring β strands, βB and βC , respectively (Fig. 3e), is very close to the linker between the gating ring and the channel pore. The hydroxyl group of Ser 337 on the linker is oriented towards the ion binding site and could also serve as a ligand for Ca^{2+} (or Mg^{2+}). This site is positioned on the upper plateau of the gating ring near the membrane surface (Fig. 3a), indicating that residues from the intracellular-facing part of the membrane-spanning channel may participate

in forming ion-binding sites. Indeed, a recent study has pointed to an acidic residue from the large S0–S1 loop as being important for Mg^{2+} binding²⁷.

The protein construct used in our structural study contains the linker region thought to connect directly the gating ring to the pore-lining inner helices of the transmembrane pore. This linker has an extended strand-like structure pointing towards the centre of the gating ring (Fig. 2a and 4) and does not seem to form any specific non-covalent interactions with other parts of the protein. A stretched conformation probably allows the four linkers of a functional BK tetramer to more effectively couple gating ring conformational changes to pore opening and closing. Indeed, insertion and deletion mutations at the linker region have a profound effect on BK channel gating²⁸.

The formation of a four-helical coiled-coil by the attached GCN4_{LI} in effect mimics the pore-lining inner helices in constraining the linker. We believe that it does not adversely affect the linker structure as we observe a similar orientation and extended strand-like structure of the linker connecting the pore and transmembrane ligand-binding domain in a recent structure of a bacterial K^+ channel from *Geobacter sulfurreducens*, which contains two tandem RCK domains as in BK channels, solved recently by our group (our own unpublished data). Taken together, these data provide plausible evidence indicating that the linker structure seen in the BK gating ring is physiologically relevant.

The presence of the linker between the BK gating ring and its channel pore allowed us to generate a full channel model by docking a voltage-gated K^+ channel onto the gating ring (Supplementary Fig. 4). The S1–S6 portion of the $\text{K}_{\text{v}}1.2$ structure²⁹ was used as an approximate model for the BK transmembrane pore and was positioned onto the BK gating ring assembly with its C-terminal end of the S6 inner helix right at the N terminus of the BK linker. Even though specific architectural features of BK and K_{v} channels may be different, we nevertheless believe that this full channel model provides useful information about the relative positioning/orientation of the BK ligand-binding domain with respect to the pore.

A 3 Å monomeric crystal structure of the human BK intracellular domain along with a tetrameric gating ring model based on the 6 Å cytoplasmic subunit structure of the chicken Slo2 channel, a Na^+ -activated BK homologue, was published immediately preceding this paper³⁰. The overall architecture of the individual subunits and the deduced gating ring model match very well with the structure of our tetrameric gating ring assembly. Moreover, in the presence of 50 mM

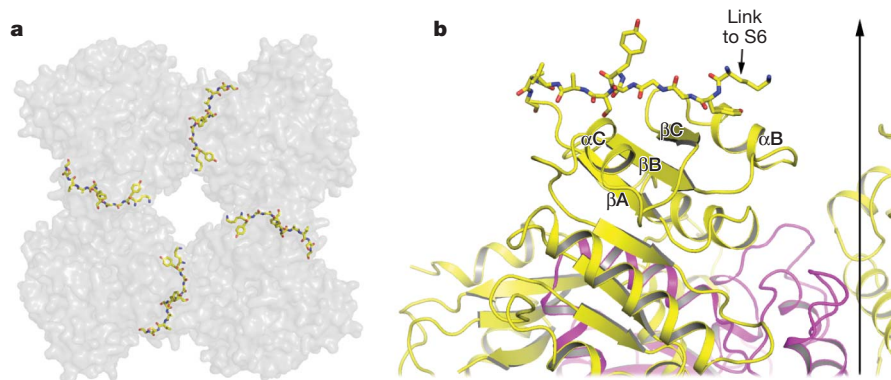


Figure 4 | Structure of the linker between the gating ring and the channel pore. **a**, Top view of the four linkers (shown in ball and stick representation) between the gating ring (grey surface representation) and the channel pore.

Ca^{2+} , the authors suggest that their structures represent the ligand-bound, open conformation compared to our ligand-free structure. Taken together, these related studies provide the first glimpse of two possible gating-ring conformations of BK channels.

METHODS SUMMARY

The intracellular fragment of the human *SLO1* (also called *KCNMA1*) gene was subcloned into the pFastBac-HTa vector and expressed in Sf9 insect cells using a baculovirus system. The construct contains an N-terminal His tag, TEV protease cleavage site and an extra 32-amino-acid stretch of sequence (RMKQIEDKLE EILSKLYHIENELARIKKLLGE) from the mutant transcription factor GCN4 leucine zipper (GCN4_LI) immediately before the N terminus of the BK intracellular domain. Expressed protein was first purified on Co^{2+} affinity resin followed by overnight TEV cleavage at 4 °C to remove the His tag. Protein was concentrated and then treated with Lambda protein phosphatase before further purification on a gel filtration column. Purified protein was concentrated to about 7 mg ml⁻¹ and crystallized at 4 °C using the sitting-drop vapour diffusion method by mixing equal volumes of concentrated protein and well solution containing 3.5 M sodium formate and 0.1 M Tris-HCl, pH 8.5. The crystals were of space group *I*422 with cell dimensions of $a = b = 134.1$ Å, $c = 231.6$ Å, $\alpha = \beta = \gamma = 90^\circ$, and contained one subunit per asymmetric unit. The structure was determined by single wavelength anomalous dispersion (SAD) using selenomethionine substituted protein crystals. Model building was aided by the Se sites and conserved structure features of RCK domains. The final model was refined to 3.1 Å with R_{work} of 23.8% and R_{free} of 28.9%, and contained residues 330–616, 684–833 and 872–1059. The density for the N-terminal GCN4_LI coiled-coil is weak and is not modelled in the final structure.

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

Received 5 May; accepted 7 June 2010.

Published online 23 June 2010.

- Kaczorowski, G. J., Knaus, H. G., Leonard, R. J., McManus, O. B. & Garcia, M. L. High-conductance calcium-activated potassium channels; structure, pharmacology, and function. *J. Bioenerg. Biomembr.* **28**, 255–267 (1996).
- Latorre, R., Oberhauser, A., Labarca, P. & Alvarez, O. Varieties of calcium-activated potassium channels. *Annu. Rev. Physiol.* **51**, 385–399 (1989).
- Cui, J., Yang, H. & Lee, U. S. Molecular mechanisms of BK channel activation. *Cell. Mol. Life Sci.* **66**, 852–875 (2008).
- Magleby, K. L. Gating mechanism of BK (Slo1) channels: so near, yet so far. *J. Gen. Physiol.* **121**, 81–96 (2003).
- Wei, A., Solaro, C., Lingle, C. & Salkoff, L. Calcium sensitivity of BK-type KCa channels determined by a separable domain. *Neuron* **13**, 671–681 (1994).
- Schreiber, M. & Salkoff, L. A novel calcium-sensing domain in the BK channel. *Biophys. J.* **73**, 1355–1363 (1997).
- Bao, L., Kaldany, C., Holmstrand, E. C. & Cox, D. H. Mapping the BKCa channel's "Ca²⁺ bowl": side-chains essential for Ca²⁺ sensing. *J. Gen. Physiol.* **123**, 475–489 (2004).
- Zhang, X., Solaro, C. R. & Lingle, C. J. Allosteric regulation of BK channel gating by Ca²⁺ and Mg²⁺ through a nonselective, low affinity divalent cation site. *J. Gen. Physiol.* **118**, 607–636 (2001).
- Bao, L., Rapin, A. M., Holmstrand, E. C. & Cox, D. H. Elimination of the BKCa channel's high-affinity Ca²⁺ sensitivity. *J. Gen. Physiol.* **120**, 173–189 (2002).

b, An enlarged side view of the extended linker (ball and stick representation) on top of RCK1 (yellow). The vertical arrow indicates the central four-fold axis of the gating ring.

- Xia, X. M., Zeng, X. & Lingle, C. J. Multiple regulatory sites in large-conductance calcium-activated potassium channels. *Nature* **418**, 880–884 (2002).
- Shi, J. *et al.* Mechanism of magnesium activation of calcium-activated potassium channels. *Nature* **418**, 876–880 (2002).
- Zeng, X. H., Xia, X. M. & Lingle, C. J. Divalent cation sensitivity of BK channel activation supports the existence of three distinct binding sites. *J. Gen. Physiol.* **125**, 273–286 (2005).
- Xia, X. M., Zhang, X. & Lingle, C. J. Ligand-dependent activation of Slo family channels is defined by interchangeable cytosolic domains. *J. Neurosci.* **24**, 5585–5591 (2004).
- Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
- Hou, S., Heinemann, S. H. & Hoshi, T. Modulation of BKCa channel gating by endogenous signaling molecules. *Physiology* **24**, 26–35 (2009).
- Jiang, Y., Pico, A., Cadene, M., Chait, B. T. & MacKinnon, R. Structure of the RCK domain from the *E. coli* K⁺ channel and demonstration of its presence in the human BK channel. *Neuron* **29**, 593–601 (2001).
- Albright, R. A., Ibar, J. L., Kim, C. U., Gruner, S. M. & Morais-Cabral, J. H. The RCK domain of the KtrAB K⁺ transporter: multiple conformations of an octameric ring. *Cell* **126**, 1147–1159 (2006).
- Roosild, T. P., Miller, S., Booth, I. R. & Choe, S. A mechanism of regulating transmembrane potassium flux through a ligand-mediated conformational switch. *Cell* **109**, 781–791 (2002).
- Kim, H. J., Lim, H. H., Rho, S. H., Eom, S. H. & Park, C. S. Hydrophobic interface between two regulators of K⁺ conductance domains critical for calcium-dependent activation of large conductance Ca²⁺-activated K⁺ channels. *J. Biol. Chem.* **281**, 38573–38581 (2006).
- Yusifov, T., Savalli, N., Gandhi, C. S., Ottolia, M. & Olcese, R. The RCK2 domain of the human BKCa channel is a calcium sensor. *Proc. Natl Acad. Sci. USA* **105**, 376–381 (2008).
- Wang, L. & Sigworth, F. J. Structure of the BK potassium channel in a lipid membrane from electron cryomicroscopy. *Nature* **461**, 292–295 (2009).
- Horrigan, F. T. & Aldrich, R. W. Coupling between voltage sensor activation, Ca²⁺ binding and channel opening in large conductance (BK) potassium channels. *J. Gen. Physiol.* **120**, 267–305 (2002).
- Lingle, C. J. Setting the stage for molecular dissection of the regulatory components of BK channels. *J. Gen. Physiol.* **120**, 261–265 (2002).
- Falke, J. J., Drake, S. K., Hazard, A. L. & Peersen, O. B. Molecular tuning of ion binding to calcium signaling proteins. *Q. Rev. Biophys.* **27**, 219–290 (1994).
- Gifford, J. L., Walsh, M. P. & Vogel, H. J. Structures and metal-ion-binding properties of the Ca²⁺-binding helix-loop-helix EF-hand motifs. *Biochem. J.* **405**, 199–221 (2007).
- Hou, S., Xu, R., Heinemann, S. H. & Hoshi, T. Reciprocal regulation of the Ca²⁺ and H⁺ sensitivity in the Slo1 BK channel conferred by the RCK1 domain. *Nature Struct. Mol. Biol.* **15**, 403–410 (2008).
- Yang, H. *et al.* Activation of Slo1 BK channels by Mg²⁺ coordinated between the voltage sensor and RCK1 domains. *Nature Struct. Mol. Biol.* **15**, 1152–1159 (2008).
- Niu, X., Qian, X. & Magleby, K. L. Linker-gating ring complex as passive spring and Ca²⁺-dependent machine for a voltage- and Ca²⁺-activated potassium channel. *Neuron* **42**, 745–756 (2004).
- Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. Atomic structure of a voltage-dependent K⁺ channel in a lipid membrane-like environment. *Nature* **450**, 376–382 (2007).
- Yuan, P., Leonetti, M. D., Pico, A. R., Hsiung, Y. & MacKinnon, R. Structure of the human BK channel Ca²⁺-activation apparatus at 3.0 Å resolution. *Science*. doi:10.1126/science.1190414 (27 May 2010).

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

Acknowledgements We thank A. Alam and M. Derebe for manuscript preparation; A. Pico for discussion in the early stages of this study; and X. Zhang for help in structure determination. Use of the Advanced Photon Source (APS) was supported by the US Department of Energy, Office of Energy Research. We thank the beamline (23ID and 19ID) staff for assistance in data collection. This work was supported by Howard Hughes Medical Institute and by grants from the NIH/NIGMS (RO1 GM071621) and Welch Foundation.

Author Contributions Y.W. performed sample preparation and structure determination. Y.Y. performed protein expression and purification. S.Y. performed

model building and refinement. Y.W. and Y.J. designed the research, analysed data and prepared the manuscript.

Author Information The atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession number 3NAF. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Y.J. (youxing.jiang@utsouthwestern.edu).

METHODS

Protein preparation. The intracellular fragment of the human *SLO1* gene (GI:26638650) starting from residue 329 to the C terminus was subcloned into the pFastBac-HTa vector containing an N-terminal His tag and TEV protease cleavage site. An extra 32-amino-acid stretch of sequence (RMKQIEDKLEE ILSKLYHIENELARIKKLLGE) from the mutant transcription factor GCN4 leucine zipper (GCN4_LI)^{31,32} was inserted immediately before the N terminus of the BK intracellular domain. The attachment of GCN4 (LI) does not affect tetramer formation of the BK intracellular subunit, as the same protein construct without GCN4 (LI) attached remains tetrameric in solution as indicated by size-exclusion chromatography. We have also been able to obtain diffracting crystals of an identical construct lacking the GCN4 (LI) attachment. However, the crystals are twinned, making structure determination problematic. The attachment serves to alleviate this twinning issue.

Recombinant bacmid was produced in *E. coli* DH10Bac by site-specific transposition using the Bac-to-Bac expression system. Isolated bacmid DNA was used to transfect Sf9 insect cells and a high titre stock of recombinant baculovirus was obtained through several rounds of amplification. Suspended culture of Sf9 cells was then infected with high titre recombinant baculovirus, with protein expression peaking at approximately 56 h after infection. Expressed protein was first purified on a Co²⁺ affinity resin followed by overnight TEV cleavage at 4 °C to remove the His tag. Protein was concentrated and then treated with 100 units of Lambda protein phosphatase per mg of protein at room temperature for 1 h, before further purification on a gel filtration column. Phosphatase treatment does not affect the biochemical behaviour of the protein but is essential to obtain diffracting crystals. The selenomethionine-substituted protein was obtained following the same protocol as described³³.

Crystallization. Purified protein was concentrated to about 7 mg ml⁻¹ and crystallized at 4 °C using the sitting-drop vapour diffusion method by mixing equal volumes of concentrated protein and well solution containing 3.5 M sodium formate and 0.1 M Tris-HCl, pH 8.5. Before being flash frozen in liquid nitrogen, the crystals were equilibrated overnight in the crystallization drop with sodium formate concentration in the well solution being increased to 4.8 M. The crystals were of space group *I*422 with cell dimensions of *a* = *b* = 134.1 Å, *c* = 231.6 Å,

$\alpha = \beta = \gamma = 90^\circ$, and contained one subunit per asymmetric unit. The four-fold axis of the channel tetramer coincides with the crystallographic four-fold.

Structure determination. All data were collected at the Advanced Photon Source (APS) and processed with HKL2000³⁴. The structure was determined by single wavelength anomalous dispersion (SAD) using selenomethionine-substituted protein crystals. Seventeen Se sites were determined with SHELXD³⁵ and initial phases were calculated using autoSHARP³⁶ and improved by solvent flattening in DM³⁷. Model building was aided by the Se sites and conserved structure features of RCK domains. The model was constructed using Coot³⁸ and was refined in PHENIX³⁹ to 3.1 Å with *R*_{work} of 23.8% and *R*_{free} of 28.9%, and contained residues 330–616, 684–833 and 872–1059. The density for the N-terminal GCN4_LI coiled-coil is weak and is not modelled in the final structure. We believe that GCN4 (LI) is not disordered like a loop but mobile as a rigid unit as it is not involved in protein packing. The main disordered regions are the long loops between RCK1 and RCK2, between α D and β E of RCK2, and the C-terminal 54 residues. All structure figures were generated in Pymol (<http://www.pymol.org>).

31. Zerangue, N., Jan, Y. N. & Jan, L. Y. An artificial tetramerization domain restores efficient assembly of functional Shaker channels lacking T1. *Proc. Natl Acad. Sci. USA* **97**, 3591–3595 (2000).
32. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
33. Cronin, C. N., Lim, K. B. & Rogers, J. Production of selenomethionyl-derivatized proteins in baculovirus-infected insect cells. *Protein Sci.* **16**, 2023–2029 (2007).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
35. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58**, 1772–1779 (2002).
36. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* **364**, 215–230 (2007).
37. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
39. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).

CORRIGENDUM

[doi:10.1038/nature09280](https://doi.org/10.1038/nature09280)***Skp2* targeting suppresses tumorigenesis by *Arf-p53*-independent cellular senescence**

Hui-Kuan Lin, Zhenbang Chen, Guocan Wang, Caterina Nardella, Szu-Wei Lee, Chia-Hsin Chan, Wei-Lei Yang, Jing Wang, Ainara Egia, Keiichi I. Nakayama, Carlos Cordon-Cardo, Julie Teruya-Feldstein & Pier Paolo Pandolfi

Nature 464, 374–379 (2010)

In this Article, author Chia-Hsin Chan was incorrectly listed as Chan-Hsin Chan.

CORRIGENDUM

doi:10.1038/nature09281

Tumour angiogenesis is reduced in the Tc1 mouse model of Down's syndrome

Louise E. Reynolds, Alan R. Watson, Marianne Baker, Tania A. Jones, Gabriela D'Amico, Stephen D. Robinson, Carine Joffre, Sarah Garrido-Urbani, Juan Carlos Rodriguez-Manzaneque, Estefania Martino-Echarri, Michel Aurrand-Lions, Denise Sheer, Franca Dagna-Bicarelli, Dean Nizetic, Christopher J. McCabe, Andrew S. Turnell, Stephanie Kermorgant, Beat A. Imhof, Ralf H. Adams, Elizabeth M. C. Fisher, Victor L. J. Tybulewicz, Ian R. Hart & Kairbaan M. Hodivala-Dilke

Nature 465, 813–817 (2010)

In this Letter, the address for author Ralf H. Adams was listed incorrectly. The correct address is: Department of Tissue Morphogenesis, Max-Planck-Institute for Molecular Biomedicine, and Faculty of Medicine, University of Münster, D-48149 Münster, Germany.

Q&A

Michael Ehlers will become the chief scientific officer for neuroscience research at drug firm Pfizer in August. He tells *Nature* why he decided to trade in his post in academia to explore a career in industry.

Did your career path aim straight at neuroscience?

No. I began my science career as a chemistry undergraduate at the California Institute of Technology in Pasadena. A close friend was a biology major and recommended a neurobiology course that she was taking, because it was so interesting. The class used electrical-engineering principles and even a theory of neural networks, and I was hooked. I went to the store and got all the books about the brain I could find, and never looked back.

Although you became a full-time basic-science researcher, are you glad you did an MD-PhD?

Yes, it has been incredibly useful. At the time, I didn't know what to do with my interest in medicine. MD-PhDs are typically good at multiple things yet undecided about where to focus their careers. But at the end of the day, it was a great decision and is one of the reasons that I am making this move now. I will get to use my background in chemistry and pharmacology, and my medical training. To pull it all together on a scale that isn't feasible in an academic lab is a great opportunity.

What does industry offer you?

There is an enormous range of unmet medical needs in neurological and psychiatric diseases. For example, there are no good therapies for schizophrenia or bipolar disorder. It will take a full suite of skills and knowledge to develop them. I am drawn to industry's approach — synthesizing different aspects of basic and translational science to develop new innovative



medicines. I think it is a significant intellectual challenge.

Is it hard to give up tenure?

If someone had told me years ago that I would be a Howard Hughes investigator and endowed professor of neurobiology at Duke University, I wouldn't have believed them. The only idea I would have found crazier is that I would willingly give that up. People have questioned why I'm making this move. I tell them that this is exactly the right time to do it. I feel as if I have reached the limit of what I can do in the medium-sized academic lab setting. I have the good fortune to have an opportunity to re-challenge myself and do something new, to work on a larger scale to help translate interesting biology into innovative new medicines.

How did you balance the potential pitfalls and perks of moving to industry?

Industry is more dynamic than academia — programmes can change and positions disappear. But when I served on Pfizer's neuroscience advisory panel, I was struck by the extremely high level of science there. The scientists have incredible dedication. They are passionate about developing medicines. I wanted to work on a larger scale, in a setting in which people are charged with collaborating in teams and there is strong

incentive for that teamwork. People in academia talk about collaborating and occasionally do, but it is not as structured.

Do you think you will have the freedom to follow basic research interests despite the applied bent of big pharma?

There is a perception that industry doesn't offer as much freedom, but the constraints or variables make drug development more intellectually challenging. I realize that I'll need an end goal in mind and business will be based on opportunity and priorities, but I've got a broad range of interests and I think the full scope of my scientific curiosity will be occupied. One of my goals is to develop a culture of innovation and exploration.

What will be your first industry project?

Within the area of neurological and psychiatric diseases there are a number of exciting drug candidates. One important development is recent research indicating that antibody technologies may deplete the brain of beta amyloid, the toxic plaques that cause Alzheimer's disease. If we can get a hook into Alzheimer's, it may open up a whole target space and disease strategy for confronting other neurodegenerative disorders.

Is a neuroscience revolution afoot?

Yes. Neuroscience is at a clear inflection point at which human and molecular genetics and systems biology are merging with an understanding of how circuits work in the brain. The next decade will bring a continued revolution — and a lot of it will happen in industry. ■

Interview by Virginia Gewin

IN BRIEF

Graduates wanted in UK

Employers in the United Kingdom's chemical and engineering sectors are seeking more graduates this year than last year, says a survey released on 30 June by High Fliers Research, a market-research firm in London. This could restore half the jobs cut in 2008 and 2009, according to the survey. Vacancies have increased by 24.1% at pharmaceutical and chemical employers, and 9.5% at engineering employers. Scientists may also want to consider applying for jobs in investment banking and accountancy, as these sectors are recruiting about 30% more graduates. Banks like to hire scientists, says High Fliers managing director Martin Birchall, "because they are good problem solvers."

Ireland cuts awards

About 600 postdoctoral and PhD awards will not be offered in Ireland this year, as fiscal belt tightening continues there. The cuts, made by Science Foundation Ireland (SFI), Ireland's main research agency, came to light on 6 July. "These losses are short-lived, inevitable consequences of budgetary restrictions, but the cuts were made across a broad spread of skills to avoid damaging any specific areas," says SFI's director general, Frank Gannon. He predicts that research positions will be among the first employment offers to emerge next year, as they are a priority for Ireland's economic recovery. In April, SFI reduced the number of grants and the amount of funding for new principal investigators (*Nature* 465, 255; 2010).

Balancing gender

Steps can be taken to address gender inequities in European science, according to a report released last month by the European Commission-supported genSET project. The report calls for gender diversity in grant-selection panels, better awareness of salary-negotiation tactics, procedures that cater to dual-career couples, and institutional public-relations activities that consistently include women (for example, in promotional campaigns for new jobs). The report also recommends reducing teaching for those with heavy administrative workloads, focusing assessments on quality of research rather than quantity, and developing better ways to encourage women to apply for scientific posts. The €1.03 million (US\$1.3 million) genSET project will be completed in 2012.

THE WAY BACK

With traditional taboos receding, and the biotechnology and pharmaceutical sectors in flux, some researchers are moving from industry back to academia. **Quirin Schiermeier** looks at the trade-offs.

Taro Hitosugi's family did not want him to join the chemistry department at the University of Tokyo in the early 2000s. After finishing his PhD in applied chemistry in 1999, the young Japanese scientist embarked on a promising career with Sony, the Tokyo-headquartered consumer-electronics company. He gained solid research and marketing expertise, and by 2003 his annual salary had reached some US\$80,000. As an assistant professor at the University of Tokyo he would earn around \$20,000 less, a differential that gave his wife pause. But seeking more autonomy, Hitosugi wanted to make the move back to academia, and his wife and family agreed to make the best of it.

Now an associate professor at Tohoku University in Sendai, Hitosugi says he wanted to publish papers and work on subjects of his choosing. "But in a company research department you can't choose your topics," he says. "It was frustrating to find interesting science angles but not having time to pursue them further." In the corporate world, managers, not researchers, usually dictate research and development priorities. Publishing opportunities are limited and some researchers lament a lack of scientific freedom.

Contrasts like these often cause young scientists to agonize over whether to stay in academia or pursue a job in industry. Moving from industry to academia can be equally difficult — yet is potentially rewarding. This path is still not common in most disciplines, but there are high-profile examples who have made the switch — such as Lee Rubin, director of translational medicine at the Harvard Stem Cell Institute in Cambridge, Massachusetts, who was formerly chief scientific officer of biotech company Curis, also in Cambridge; and physicist Cherry Murray who moved from Bell Laboratories to Harvard. The two worlds are no longer as separate as they were, and career paths are more of a two-way street. Those looking to make the change should consider the salary implications, a potentially painful new start with a new team and the relentless demands of getting funding.



IMAGES.COM/CORBIS

Ten years ago, it would have been rare for a private-sector scientist to move back to an academic position, given the then-booming biotechnology sector. Many had left academic positions to pursue the plentiful opportunities in the private sector. But the biotech downturn and the pharmaceutical industry's struggles have made more researchers keen to return to academia. And in some places, notably the United States, university departments have become more welcoming to scientists, especially experienced ones, with a non-academic background. The industry-to-academia trend is most pronounced in the life sciences — particularly in systems biology and drug

discovery — where publicly funded research institutes and universities are increasingly involved in the development of new drugs and therapies for diseases that have been neglected by the pharmaceutical industry.

A potent mixture

Many universities like this intermingling of expertise and the addition of researchers with industrial experience. It can prove an asset for an academic institute when it wants to set up collaborations with industry, and make use of advanced technology with which industry researchers have more experience. Industry researchers also offer a valuable perspective on teamwork and efficient group management,

project management, high-throughput data analysis and quality control, says Paul Wyatt, head of the drug-discovery unit at the University of Dundee, UK, who took up his current post in 2006 after working as a project leader at several drug companies, including Astex Therapeutics in Cambridge, UK. Scientists with industry experience can also help to exploit a lab's research commercially, facilitating spin-offs, for example.

"Here in Dundee we now work very much as in a biotech company," says Wyatt. "Having experienced industry people among their faculty can help an academic unit like ours attract real world-class scientists." As many biology departments move towards drug discovery, Wyatt expects quite a few ex-industry people to move to academia. His unit is looking for experienced drug-discovery researchers to work on neglected diseases such as malaria and African sleeping sickness.

Those who successfully bridge the two sectors bring a valuable perspective. Hitosugi, for example, has set up collaborations with several leading companies, including Toyota. He says that the car-maker appreciates his ability to identify novel technologies with real market potential. "I understand the company logic," he says. "I know that managers need to think about product planning and marketing early on." Technology development, says Hitosugi, is only part of the story; getting products to market is just as important. "It's an advantage for me that I do know how to sell," he says, referring to his industry experience. "Thanks to my experience in marketing I can imagine the whole process from basic research to product planning, production and sales."

At the Harvard Stem Cell Institute, Rubin says that he used his industry experience to pursue stem-cell research, for which industry has little appetite. "Ironically, the academy is now often a better place to pursue therapeutic discovery than industry," says Rubin. The university community has been "amazingly receptive" to his work, he says.

Researchers can become valuable conduits of knowledge and expertise between the two sectors. "Drug development will in the future become a big joint industry-academic project," says computational biologist Igor Goryanin, chair of computational systems biology at the University of Edinburgh, UK, and a former industry researcher.

"I see myself as an interpreter who, connected to two different worlds and mentalities, speaks both languages," Goryanin was a senior research scientist with drug maker GlaxoSmithKline, based in London, UK. He moved to Edinburgh in 2005 after becoming dissatisfied with the company's changing research focus. Budgets decreased and the company began, he says, to focus more on production and marketing of existing products rather than on developing innovative drugs. Goryanin says that he now has ample financial support, thanks to a large



Taro Hitosugi does not regret leaving industry.

European Union grant on biosimulation and further funding from the UK and Scottish governments.

These interpreters benefit students as well, notes Richard DiMarchi, a biochemist at Indiana University in Bloomington who spent more than 20 years with global drug maker Eli Lilly, headquartered in Indianapolis. DiMarchi's lab has three senior members with a combined 75 years of experience in industry. "For an academic lab, these people provide a diversity of experience which no single professor could ever supply," he says. "It is a wonderful environment for the education of PhD students and postdocs." He and his industry-veteran lab mates share their insights into the stages of the drug-discovery and development process, from the early experimental phase to proof-of-concept to clinical trials.

Tough transition

The transition back to academia will not be easy, however. "It was a shock," says Steven Kliewer, a molecular endocrinologist at the University of Texas Southwestern Medical Center in Dallas, who oversaw GlaxoSmithKline's research on nuclear receptors and protein orientation at its North Carolina research centre until 2002. His new position came with tenure, but he still needed research results to secure grants. "It was a very tense time — lots of sleepless nights, extremely frustrating from an ego perspective," he says.

After the merger of Glaxo Wellcome and SmithKline Beecham in 2000, the company started supporting less basic research, and Kliewer, who had up to then declined overtures from several US universities, started considering his options. "I saw the writing on the wall," he says. "I had once thought I'd stay forever in industry where pay is good and so are the hours. But I was more passionate about doing basic research than I had realized."

Industry, he notes, requires people to tackle whatever projects are needed as they come up. Kliewer, however, wanted to focus on one line of research — nuclear receptors.

When the right opportunity came along, he seized it. But getting reacquainted with academia proved much harder than the transition to industry ten years earlier. Having come from a 20-member team with a basic-research focus, Kliewer already had several publications to his name. But he soon realized that references and reputation were not enough to secure money from grant agencies. It took him more than a year to build a top-notch team, and two years to generate the experimental results necessary to earn a grant from the National Institutes of Health. In the interim, he worked with a small team supported by small grants and bridge funding. He says that his department supported him, knowing that the transition would be difficult.

The challenge of competing for grants and keeping funders happy awaits most scientists who return to academia. "Sustaining your research group is a big issue, particularly in the current climate of crisis," says Goryanin. Many industry scientists may have published little or nothing, having been free of the need to secure grants. A sparse publication record can be an obstacle to returning to an academic position.

Salary differential is another potential downside. Industry researchers, in particular those working for large companies where generous bonus payments are common, will usually earn less when they return to academia. And industry researchers starting a new group from scratch often have to reacquaint themselves with the system,

learning how and where to submit grants or get animal protocols approved — tasks an academic might take for granted.

Nevertheless, more professional exchanges between the worlds of industry and academia would be beneficial for individual career development and for research at large, says Adriano Henney, programme director of the Virtual Liver Network (HepatoSys), a multi-year, €46-million (US\$57.8-million) German systems-biology initiative. In a world of

increasing overlap and collaboration between academic and industry research, company managers should know more about academic hierarchies, and universities could learn a lot from industry about professional project and programme management. "People with expertise on both sides can bring together the two communities to the benefit of science and society," he says.

Quirin Schiermeier is Nature's Germany correspondent.



"I had once thought I'd stay forever in industry."

— Steven Kliewer

Gravity's whispers

Proof positive.

Gregory Benford

"The best is the enemy of the good," Sam said over my shoulder.

I whirled around, knowing the voice, smiling. "What —?"

He sauntered in, grinning in his lopsided way. "At 11 p.m. you're still working. Know your limits. The data can't get better when you're tired, y'know."

I threw down my pencil. "Right. Pursue the good. Let's get a beer."

At the Very Large Array, this meant a long drive back to Socorro. Our offices were there, but I liked spending time out among the big radio dishes, too. On the way back I rolled down the window to smell the tangy spring sagebrush and wondered whether Sam the Slow had finally decided to make a date with me, in his odd way. I'd been waiting half a year.

Then he said: "I was just passing by, thought I'd follow up on that puzzle I sent last week."

He had sent through a noise-dominated file. I had run one of my custom programs, got interested, and wasted a day pulling out a pattern. "You know me too well. I cracked it, yeah." I gave him a smile he didn't notice. "Not a very interesting solution."

"You'd be surprised," Sam said, watching the desert slide by.

"It's you guys who surprised the world — the first gravity waves, wow."

"Yeah, decades of work on LIGO paid off."

Sam was also modest, a trait that gave him gal problems in the fanatic tech crowd more than once. Getting a gravitational wave to tweak a cavity, and detect that with interfering waves, had burned 20 years of his life. He shrugged. "We thought it was a signal from a rotating neutron star with a deformed crust. Say, you have that solution handy?"

I flipped open my laptop. "It's a string of numbers, turns out to be the zeroes of the Riemann zeta function."

"Uh huh. Which is —?"

"A famous function of complex argument. It analytically continues the sum of

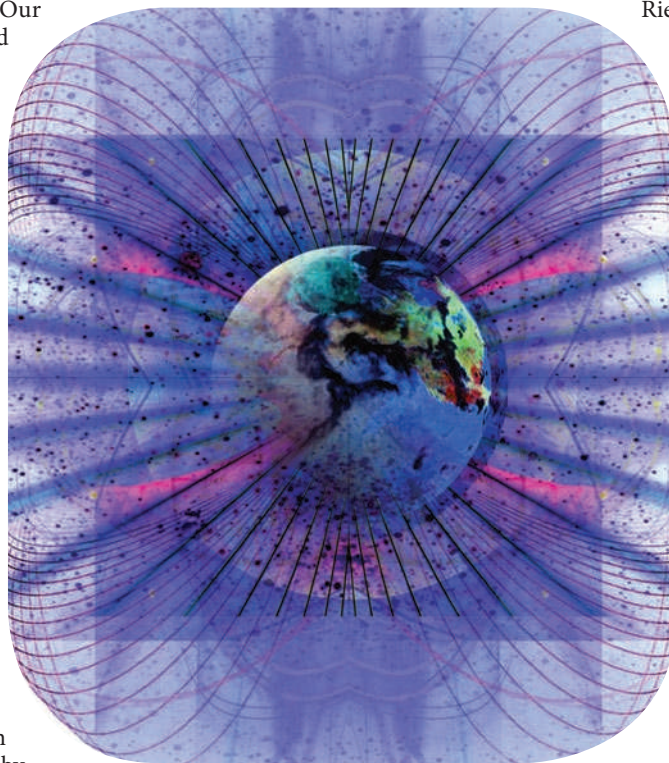
an infinite series."

"Sounds boring."

"Not so." At least he was looking at me now. "It's a big deal in analytic number theory, plenty of applications in physics, probability theory, Bose–Einstein condensates, spin waves —"

"Useful, good." Sam was usually sharp, focused, but now he gazed pensively at the stars.

"So how'd you get the detection?" It would help if I got him started about his



work — that is, his life. "You guys got rid of the noise from that road traffic and logging at the Louisiana site?"

"Yeah, took years. The signal we finally got had plenty of chirps and bursts in it, a bitch to clean up."

I grinned. Sam had worked for decades on LIGO, and now the milestone was here. "Now that you've got LIGO sensitive enough, there'll be plenty of signals. Supernovae in other galaxies, maybe rattling cosmic strings —"

"I want to understand this one. It's not a neutron star crust vibration, I think."

"Huh?" I was already tasting the beer in my mind.

"That decoding you did? That was our signal."

I blinked. "Can't be. No natural system —"

"Exactly." Sam hooked an eyebrow at me.

"What? A tunable gravitational wave with a signal? That's im—"

"—possible, I know. Unless you can sling around neutron stars and make them sing in code."

Maybe, just maybe, this could be more important than at last getting Sam to date me. "Then ... you should know that it's not just a list of numbers. After 20 of the Riemann zeros, there's something like a proof of the Riemann hypothesis."

He frowned. "Uh, so?"

"It's one of the greatest unsolved problems in mathematics. It says that any non-trivial zero has its real part exactly equal to $1/2$."

He shook his head. "And that's the attention-catching opener to a SETI signal?"

"So you see, it can't be. Opening up with π or e , prime numbers, the fine structure constant — that makes sense."

"Sense to the likes of us."

"So I must've made some mistake."

"No you didn't." Sam looked at me with a warm smile. "You're the only one I could run to with this analysis — the rest of 'em would laugh. You're good, really good."

I leaned over and kissed him.

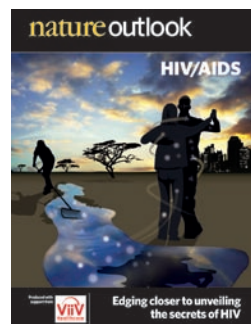
"Congratulations on the Nobel."

He kissed back, his eyes flickered, he grinned — but he didn't look happy. He grasped the steering wheel and peered ahead into the starlit darkness. In the high desert you can see stars above the headlights. I knew him enough to see that he was thinking about something that could whisper across the galaxies with gravitation, not using obvious means like radio or lasers. "Any mind that thinks the Riemann numbers are a calling card — and can throw around stars ..."

I got it. "Yeah. Know your limits. Maybe it's good, really good, that we can't possibly answer them."

Gregory Benford is a professor of physics at the University of California, Irvine, and the author of *Timescape*.

JACEY



COVER ART: NIK SPENCER

Supplement Editors

Apoorva Mandavilli,
Michelle Grayson

Art & Production

Martin Harrison,
Sarah Lorient, Karl Smart

Production Controllers

Susan Gray, Emilia Orviss

Web Production Editor

Terhi Raukko

Sub Editor

Victoria Kitchener

Marketing

Elena Woodstock,
Hannah Phipps

Managing Editor

Nick Campbell

Magazine Editor

Tim Appenzeller

Project Manager

Claudia Deasy

Editor in Chief,

Nature Publications
Philip Campbell

Sponsorship

Gerard Preston, Reya Silao,
Stephen Russell



nature publishing group

If you are only a casual follower of the trends in HIV/AIDS research, you would be forgiven for thinking that there is nothing new to say about this field.

The miraculous drugs that keep so many HIV-positive people alive have, ironically, blunted the urgency with which the press, policy makers and the general public used to talk about the epidemic.

In fact, as this *Nature Outlook* shows, there is a renaissance afoot in the field. There is, for instance, hushed talk of something most people had given up on: a cure for HIV/AIDS. With more powerful new drugs and a clearer understanding of the virus' ways, some researchers are quietly exploring approaches to flush the virus out of the body completely.

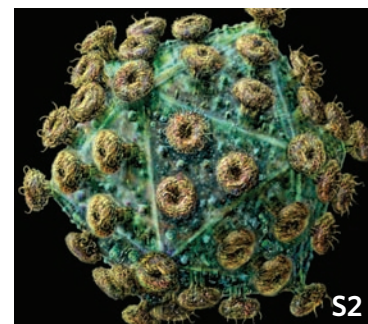
In the past few years, researchers have also made remarkable progress in understanding how HIV literally exhausts the immune system by keeping it on high alert, and how exquisitely different each individual's immune response is to the virus. Scientists are learning, slowly but surely, how to build antibodies that can neutralize the virus, and vaccines that can prevent it from infecting people in the first place.

Some of these lessons come from unexpected places — ranging from monkeys in Asia to sex workers in Africa and haemophiliacs in the United States. They are also the result, as is often the case with big advances in science, of multidisciplinary collaborations.

Outside the laboratory, too, there is new hope. Experts are combining every tool at hand — such as mobile telephones, male circumcision and microbicides laced with antiretroviral drugs — to prevent new infections. Even South Africa, that hotbed of AIDS denialism, is finally facing up to its epidemic.

We are pleased to acknowledge financial support from ViiV Healthcare in producing this Outlook. As always, *Nature* carries sole responsibility for all editorial content.

Apoorva Mandavilli, Consultant Editor



contents

- S2 On high alert**
Emma Marris
- S4 Learning from the elite**
Bijal Trivedi
- S5 The primate connection**
Bijal Trivedi
- S6 Dancing with an escape artist**
Sarah DeWeerd
- S8 Tiny steps towards an HIV vaccine**
Cassandra Willyard
- S9 Joining forces**
Cassandra Willyard
- S11 The outlook for a cure**
Virginia Hughes
- S14 Forgotten lessons**
Paroma Basu
- S16 Developing solutions**
T. V. Padma
- S18 Fighting the monster**
Amy Maxmen
- S20 Grassroots initiatives**
Amy Maxmen
- S21 A call for collaboration**
Unmesh Kher
- collection**
- S23 AIDS research must link to local policy**
Salim S. Abdool Karim & Quarraisha Abdool Karim
- S25 Public database for HIV drug resistance in southern Africa**
Tulio de Oliveira, Robert W. Shafer & Christopher Seebregts
- S26 Is the high virulence of HIV-1 an unfortunate coincidence of primate lentiviral evolution?**
Frank Kirchhoff
- S36 HIV-1 exploits innate signaling by TLR8 and DC-SIGN for productive infection of dendritic cells**
Sonja I. Gringhuis *et al.*
- S44 Retroviral intasome assembly and inhibition of DNA strand transfer**
Stephen Hare *et al.*

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* editorial guidelines available at http://www.nature.com/advertising/resources/pdf/outlook_guidelines.pdf

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol XXX, No. XXXX Suppl, Sxx-Sxx (2010).

To cite previously published articles from the collection,

please use the original citation, which can be found at the start of each article.

VISIT THE OUTLOOK ONLINE

The *Nature Outlook HIV/AIDS* supplement can be found at www.nature.com/nature/outlook/HIV-AIDS/. All featured articles will be freely available for six months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe (excluding Japan):
Nature Publishing Group, Subscriptions,
Brunel Road, Basingstoke, Hants, RG21 6XS, UK.

Tel: +44 (0) 1256 329242.

Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada).

CUSTOMER SERVICES

Feedback@nature.com

Copyright © 2010 Nature Publishing Group

On high alert

HIV keeps the immune system in a hyperactive state, gradually leading to its ruin, reports **Emma Marris**.

The last time you had a cold, you probably sniffled, coughed and stayed at home in bed for a couple of days. All the while, your immune system fought a quick battle against the cause of the cold: a virus. It was a battle with a foregone conclusion, as many of them tend to be. In most cases, the immune system is triumphant, and mops the floor with the virus within a couple of weeks.

HIV is no ordinary virus, however. After a quick skirmish with the immune system — what researchers call the ‘acute phase’ of infection — HIV escapes eradication, often living on quietly for years and only much later dragging the immune system to utter ruin.

In the past few years, researchers studying the long road from infection to AIDS have been particularly interested in ‘immune activation’ — a state in which the immune system stays on high alert and prepared to fight.

In this condition, the immune system marshals many types of cell into an active mode, ready to sense and attack pathogens. The body produces abundant amounts of signalling

proteins, known as cytokines and chemokines, and a host of other proteins.

This ready and roused state is short lived in the case of most viruses; however, HIV triggers chronic immune activation. “Practically every arm of the immune system that has been investigated has been shown to be in a hyperactive state,” says Jason Brechley, an investigator at the US National Institutes of Health.

As immune cells are activated, they become targets for HIV, so an active immune system paradoxically amounts to a higher viral load. Cells that constantly replicate when activated eventually become depleted, causing a premature ageing of the immune system. In a process that is still mysterious, chronic immune activation seems to wreck the tissues that produce immune cells¹.

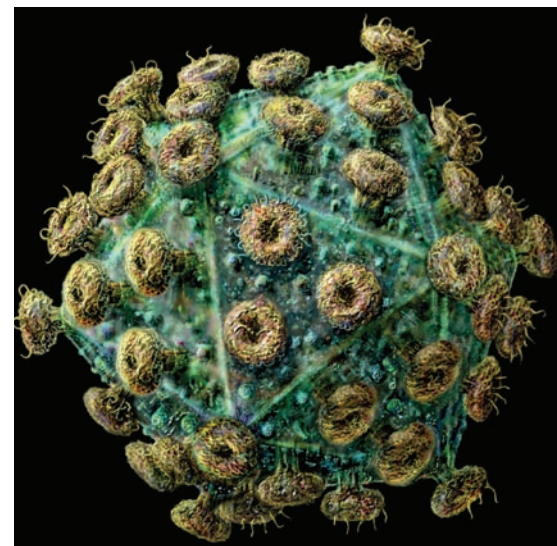
This set of effects has been linked to conditions ranging from heart disease to cancer. Even on its own, immune activation “makes you feel lousy,” says Daniel Douek, chief of human immunology at the US National Institute of Allergy and Infectious Diseases’ Vaccine Research Center.

Long road

The basic outline of HIV pathogenesis is well chronicled. The virus enters the body usually through mucosal tissue, such as the lining of the vagina or anus. Once there, it encounters CD4 cells, which orchestrate the immune response. The virus gains entry into these cells, where it insinuates its DNA into host nuclear DNA. For a day or less, each CD4 cell essentially becomes a factory for viruses, and then it dies.

This is what allows HIV to persist where other viruses beat a hasty retreat. Once the acute infection is past, HIV continues to replicate and diversify by mutating into many different forms, each one a new challenge to the immune system (see page S6).

Over years of infection, the average person’s CD4 cells slowly decline. HIV takes up residence in quiescent immune cells and possibly in other hideouts (see page S11). The depletion of CD4 cells cripples the immune system, leaving the host vulnerable to any attack, and turning HIV infection into full-blown AIDS.



A key element in this story is pace. The rate of disease progression varies enormously between people, and some ‘elite controllers’ never reach the milestones that mark AIDS (see page S4). Scientists can roughly predict disease course by looking at virus levels in the blood. As far back as 1993, however, researchers found that elevated levels of CD38⁺ CD8⁺ cells — a subset of immune cells that correlate with immune activation — are a better harbinger of AIDS’ arrival².

“Immune activation is one of the oldest ideas in HIV research,” notes Carl Dieffenbach, director of the Division of AIDS at the US National Institute of Allergy and Infectious Diseases. “We have just rediscovered this more times than I care to think about.”

“Practically every arm of the immune system has been shown to be in a hyperactive state.”

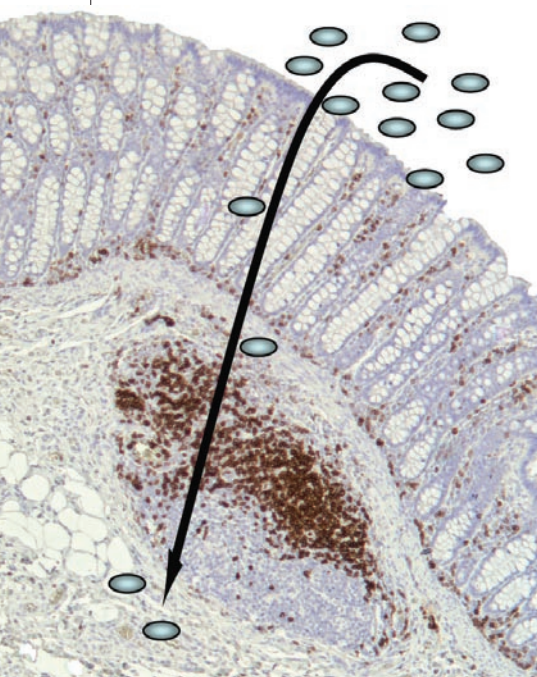
Recurring theme

When researchers first began thinking about immune activation, they saw it primarily as a symptom of the illness.

“At the time before we had good drugs, or any drugs at all, I think that most people thought that [if] we got a good drug and knocked out the virus, that immune activation would go away,” says Susan Plaeager, director of basic research in Dieffenbach’s division. As therapies that control the virus became available, however, it was clear that immune activation lingers even when viral numbers are down.

Several new findings have brought immune activation back to centre stage, according to researchers.

First, scientists discovered, about a decade ago, that sooty mangabeys and African green monkeys, ‘natural hosts’ of the HIV-like simian immunodeficiency virus (SIV), do not show immune activation and do



HIV damages the gut and makes it leaky, allowing other pathogens to enter and wreak havoc.

not get sick. Rhesus macaques, the experimental model for HIV vaccines, do have immune activation, however, and develop AIDS (see page S5).

Second, researchers reported in 2008 that elite controllers, despite their seemingly good health and negligible viral loads, show some abnormal immune activation³. Finally, many HIV-infected individuals taking antiretroviral drugs have low viral loads but suffer from inflammation-related diseases of ageing — heart disease, liver disease, cancer and perhaps even dementia — earlier than do their uninfected peers.

One trial, called Strategies for Management of Anti-Retroviral Therapy (SMART), which was designed to study drug scheduling in HIV-positive people, found that protein markers of inflammation were higher in those who died during the trial⁴. The implication is that immune activation contributes to mortality from conditions not typically classified as AIDS-related, says Douek.

Overall, evidence from natural host monkeys and clinical observations suggest that viral infection and immune activation are both needed for progression to AIDS.

Persistence of memory

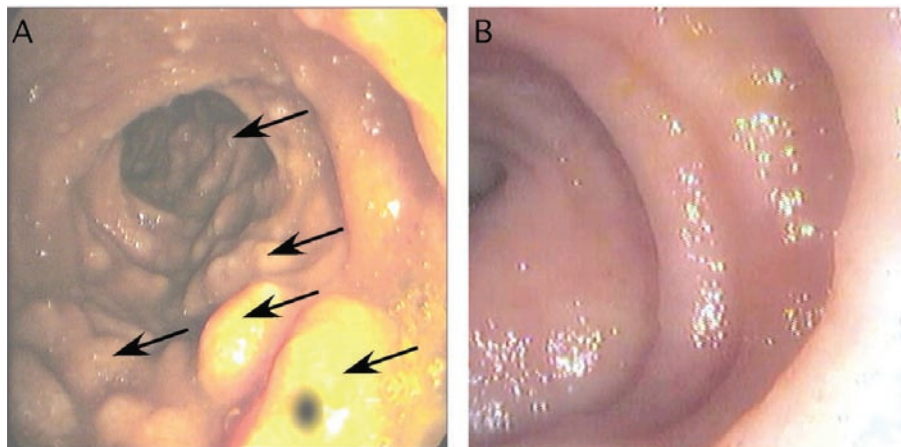
What causes immune activation? Most obviously, it is the virus. Like any invader, it causes immune cells to wake up and gird their loins for battle. It is not clear, however, why the state of activation should persist for decades.

It might be that the kind of cells that are infected are so important to the immune system that the host has no choice but to try to get rid of every last bit of the virus, suggests Guido Silvestri, a pathologist at the Hospital of the University of Pennsylvania in Philadelphia.

In particular, the presence of virus in lymph nodes, where ‘central memory’ CD4 cells live, could keep the immune system engaged in fighting the virus, Silvestri suggests. These cells specialize in retaining information about a pathogen’s identity and creating armies of rank-and-file ‘effector’ memory cells for when the host re-encounters the threat. Tellingly, sooty mangabeys and African green monkeys do not lose as many central memory cells or show chronic immune activation.

“It doesn’t really matter how much virus you produce in the body, but what kind of cells are killed,” says Silvestri. “Not all CD4 cells are created equal. Some are really quite expendable.”

Evidence in the past seven years of a great decline of CD4 cells in the gut has prompted two more theories about how HIV causes chronic immune activation.



In the acute phase of infection, HIV depletes the gut’s CD4 immune cells (seen in the bumps on left), and their numbers never bounce back (right).

In the acute phase, the virus depletes CD4 cells in the mucosal tissue of the gut, and their numbers never bounce back^{5,6}. In the process, the virus also destroys the gut mucosa’s structural cells. “The gut has holes, it is leaky,” says Douek.

Damage to the gut’s immune system allows other pathogens to make their way to lymph nodes and penetrate the body, where the rest of the immune system must wake up and fend them off, says Satya Dandekar, a microbiologist at the University of California, Davis.

This immune activation persists even when drugs knock the virus down to undetectable levels because the gut has lost its ability to repair itself. Early treatment might prevent this damage, Dandekar says. “A lot of benefit that one sees in early treatment may really stem from repair of the mucosal sites.”

Another, potentially complementary, theory suggests that the weakened gut defence allows pieces of the omnipresent — and generally benign — gut bacteria to enter the bloodstream, where they activate the immune system⁷.

This ‘microbial translocation’ has been shown in HIV-infected people in several studies, mostly in the later stages of disease. In May 2010, Douek, Brechley and colleagues reported in *Science Translational Medicine* that HIV infection increases the numbers of regulatory T cells — a subtype of CD4 cells that modifies the immune response — and lowers the numbers of another immune cell that helps defend the gut mucosa⁸. The resulting imbalance might lead to a leaky gut and a dysfunctional immune system, the researchers suggest.

“It doesn’t really matter how much virus you produce in the body, but what kind of cells are killed.”

Some researchers argue that microbial translocation is the consequence, not cause, of immune activation, but Douek dismisses the criticism. “It’s a circle — immune activation causes leaky gut which causes immune activation which causes leaky gut,” he says. “In a sense it doesn’t matter [which came first].”

Given that immune activation is intricately tied to disease progression, measuring viral load or even the number of CD4 cells might not be the best way to chart the disease, Douek adds. Doctors could instead count central memory CD4 cells or markers of immune activation and inflammation, and combine several measurements into a standardized algorithm that predicts the course of infection.

Meanwhile, some groups are trying to calm the chronically alert immune system by adapting therapies for other inflammatory conditions — some as basic as aspirin or vitamin D — and designing drugs to stop microbial translocation.

“Now that we are able, we believe, to suppress virus for many, many years with antiretroviral drugs,” says Douek, “immune activation and inflammation in treated individuals may become the greatest obstacle to long-term health.”

Emma Marris writes for *Nature* from Columbia, Missouri.

1. Sodora, D. L. & Silvestri, G. *AIDS* **22**, 439–446 (2008).
2. Giorgi, J. V. et al. *J. Acquir. Immune Defic. Syndr.* **6**, 904–912 (1993).
3. Hunt, P. W. et al. *J. Infect. Dis.* **197**, 126–133 (2008).
4. Kuller, L. H. et al. *PLoS Med.* **5**, e203 (2008).
5. Guadalupe, M. et al. *J. Virol.* **77**, 11708–11717 (2003).
6. Brechley, J. M. et al. *J. Exp. Med.* **200**, 749–759 (2004).
7. Brechley, J. M. et al. *Nature Med.* **12**, 1365–1371 (2006).
8. Favre, D. et al. *Sci. Transl. Med.* **2**, 32ra36 (2010).

Learning from the elite

Researchers hope to unlock the secrets of the select few who rein in, or even resist, HIV infection, says **Bijal Trivedi**.

Bob Hoff has been HIV-positive since 1984. He has never taken antiviral drugs, yet his immune cells are numerous and healthy, and the virus is undetectable in his blood.

Hoff is an 'elite controller' who can do what big pharma and vaccine developers cannot: hold HIV in check and keep AIDS at bay without medications.

More than 33 million individuals are infected with HIV, and about 2.7 million join those ranks each year. However, it is estimated that 1 in every 200 people infected never succumbs to the virus¹, whereas another 6 or so 'long-term non-progressors' (LTNPs) control HIV, albeit with a higher viral load, and remain symptom-free for at least seven years.

Figuring out how these people control HIV could inspire vaccines and therapies that mimic their immune systems. "[When] you meet someone who's been infected for 25 years, looks entirely healthy and has never taken any medications, you just feel like the answer is standing there for us to just fish out," says Bruce Walker, director of the Ragon Institute in Charlestown, Massachusetts.

Since 2003, the International HIV Controllers Study has examined the factors that keep elite controllers and LTNPs healthy. Other research groups are studying the few people with natural immunity to HIV, who resist infection even after multiple exposures.

An infected individual carries, on average, about 30,000 copies of HIV RNA per millilitre of blood. Those with AIDS have more than 100,000 copies, LTNPs have between 50 and 2,000, and elite controllers have fewer than 50, which is well below the detection limit of most conventional tests.

Last year, Walker showed that the immune system in elite controllers cripples the virus,



Some Kenyan sex workers, like Dorothy Mumbe (above), resist HIV infection even after being exposed hundreds of times to HIV-positive men.

forcing it to mutate and evolve in a way that renders it less fit². "These people didn't get infected with a wimpy virus — they made their viruses wimpy," he says.

Hints from the genome

Together with researchers at the Broad Institute in Cambridge, Massachusetts, Walker also led a genome-wide association study including about 1,000 elite controllers. Earlier this year his team announced, at a conference in Banff, Canada, that 350 single-nucleotide polymorphisms (SNPs) clustered in the major histocompatibility complex (MHC) — a region on chromosome 6 rich with immune-response genes — seem to protect elite controllers from the effects of the virus.

One clue to the role of these SNPs comes from a study published in *Nature* in May this year³. Walker and colleagues reported that a variant of the human leukocyte antigen B (*HLA-B*) gene located in the MHC boosts the immune response by producing powerful killer T cells, which recognize and destroy infected cells. T cells from those who carry the *HLA-B57* variant recognize many parts of the HIV virus and more HIV mutants than do T cells from controls.

About 65% of elite controllers carry the *HLA-B57* allele, says Mark Connors, chief of the HIV-Specific Immunity Section at the US National Institute of Allergy and Infectious Diseases. The link between the allele and T-cell function is "an attractive idea," but the computer models need to be backed up experimentally, Connors notes.

Quiescent systems

Rare individuals with natural immunity to HIV could also yield game-changing discoveries. These people — including sex workers, haemophiliacs who have received contaminated blood, and individuals with an infected

partner — are repeatedly exposed to the virus, but never get infected.

"Some women we've been studying for years have been exposed hundreds of times to men we know are HIV-positive. They weren't using a condom, despite counselling, and still they weren't infected," says microbiologist Keith Fowke of the University of Manitoba who has been studying a cohort of commercial sex workers in Nairobi for more than 20 years. "That's well beyond the realm of luck — it's something really biological," he says.

HIV-resistant women appear to have 'quiescent' or resting immune systems, Fowke says. They show lower levels of cytokines, which are signalling proteins that ramp up the immune response, and the activity of many non-immune genes is also lower than in controls.

The women have more regulatory T cells, which calm the immune system, and fewer activated CD4 immune cells. However, their ability to fight other infections and viruses is not compromised.

HIV infects and replicates better in activated cells. Fewer targets allow greater opportunity for the evolution of immune responses that would eliminate the virus, says Frank Plummer, scientific director general of the National Microbiology Laboratory at the Public Health Agency of Canada. The idea that a quiet immune system could be the key is "potentially paradigm shifting," Plummer adds.

Some reported data contradict this hypothesis. Plummer has proposed launching a global consortium to study individuals with natural immunity. He and others are also trying to identify the genes that endow the Kenyan women with resistance to HIV.

Bijal Trivedi is a freelance writer in Washington DC.

1. Okulicz, J. F. et al. *J. Infect. Dis.* **200**, 1714–1723 (2009).
2. Miura, T. et al. *J. Virol.* **83**, 2743–2755 (2009).
3. Košmrlj, A. et al. *Nature* **465**, 350–354 (2010).

The primate connection

Studies comparing HIV infection and its simian counterpart in different monkey species are filling gaps in knowledge, explains **Bijal Trivedi**.

Male African green monkeys are notorious for their striking red penis, bright-blue scrotal area and white abdomen. To AIDS researchers, however, these monkeys and another African species, the sooty mangabey, are known for a less visible trait.

Both are natural hosts of the simian immunodeficiency virus (SIV), which is the primate counterpart and ancestor of HIV. Although infected monkeys have more than 100,000 copies of SIV per millilitre of blood, they do not get sick or develop anything resembling AIDS. By contrast, Asian macaques infected with SIV mimic the trajectory of human HIV infection.

Understanding what allows the African monkeys to coexist with SIV could help develop drugs or vaccines against HIV that ape their response.

“Comparative AIDS research is essential to figure out what protects us from transmission, and what protects us from pathogenesis,” says Guido Silvestri, professor of pathology at Emory University.

SIV infections in the natural host monkeys share similarities with human HIV and macaque SIV infections; however, it is the differences that intrigue researchers. For example, unlike humans and macaques, female natural hosts rarely pass the virus on to their offspring.

In people, there is a general trend: the more virus in the body, the faster the progression to AIDS. This is probably why elite controllers, whose immune systems can whittle down viral loads to barely detectable levels, rarely get AIDS (see page S4). By contrast, sooty mangabeys and African green monkeys with high viral loads remain healthy.

What is more, the natural hosts maintain relatively healthy numbers of CD4 cells — white blood cells that orchestrate the immune response and are targets of HIV — in both blood and the mucus lining of the genital tract.

Evasive targets

One key to the monkeys’ resilience might be that they have fewer cells that HIV can infect.

CD4 cells in sooty mangabeys have lower levels of the receptor CCR5, which HIV needs to enter the cell. African green monkeys have fewer surface CD4 receptors, which are required for HIV infection.

“They’ve chosen two different ways of removing target cells of the virus,” notes Daniel Douek, who leads the Human Immunology Section of the US National Institute of Allergy and Infectious Diseases’ Vaccine Research Center. “I really think that is the root of why they don’t progress [to AIDS].”

Viral load is not the only factor influencing progression. People with low levels of HIV sometimes show serious inflammation and a slow decline in CD4 cells. Some such cases slowly progress to AIDS, says Steven Deeks, professor of medicine at the University of California, San Francisco.

“There are people who think it is only the virus, others think it is only inflammation, and we think it is a little of both,” adds Deeks.

When the virus first infects the body, humans, macaques and natural hosts mount a dramatic immune response. In sooty mangabeys and African green monkeys, this reverts to baseline levels after about three months, whereas it remains high in humans and macaques throughout the chronic infection.

“It is amazing that [the natural hosts] are able to switch off these pathways of immune response even when virus replication remains consistently high. And it seems that they benefit from doing this switch off,” says Silvestri.

One potential trigger for chronic immune activation in humans

and macaques could be the loss of T-helper 17 (T_H17) cells — a subset of CD4 cells that is key for antibacterial defence. Douek and others have shown that SIV-infected sooty mangabeys and African green monkeys have healthy levels of T_H17 cells in the mucus lining of the gastrointestinal tract. Loss of these cells is a hallmark of progressing infection in humans and macaques (see page S2).

Model monkeys

In their search for a surrogate for HIV in humans, researchers use about 20 different SIV strains and modified SIVs, dubbed SHIVs, to infect different species of macaques. Each virus–host combination has strengths and limitations.

The main problem with the SIV–macaque model, says Paul Bieniasz, a virologist at the Aaron Diamond AIDS Research Center, New York, is that the SIV genome differs by about 50% from HIV-1, which is the most prevalent strain worldwide.

To create a better model for vaccine tests, Bieniasz and colleagues are modifying HIV-1 to grow in monkeys. So far, they have created a version that will infect pigtail macaques, initially causing virus levels to reach up to one million copies per millilitre of blood — just like in the initial stages of human infection.

These monkeys do not become sick and, after six months, their infection mimics that seen in long-term non-progressors, who naturally control HIV levels and slow their progression to AIDS (see page S4).

“We basically have to find a way to make the virus more robust in pig-tail macaques,” says Bieniasz. “We need it to trigger AIDS.”

Bijal Trivedi is a freelance writer in Washington DC.

“It is amazing that [the monkeys] are able to switch off these pathways.”



African green monkeys (left) and sooty mangabeys (right) are natural hosts of the HIV-like simian immunodeficiency virus, meaning that they never develop anything resembling AIDS.

Dancing with an escape artist

Sarah DeWeerd describes the intricate relationship between HIV and the host immune system, each influencing the other's next moves.

When people talk about infectious disease, the host is often described as being at war with the pathogen, which is an apt metaphor. Look closely at HIV infection, however, and it also seems like a dance: the virus and the immune system match each other, step for step, in a performance as intricate as it is deadly.

Each person's immune system is unique and dances differently. Although HIV infection usually leads to AIDS, the timing and course of progression varies greatly from person to person.

In the past few years — thanks to methods that detect HIV soon after infection — researchers have learnt that the immune response in the first weeks often sets the long-term course of the disease. They are also learning how the body makes antibodies to HIV, and how this immune response evolves over time — insights that might aid the search for a vaccine.

HIV becomes detectable in the bloodstream about ten days after the initial infection. Over the next two to three weeks, it spreads through

the body, replicating wildly. At the same time, killer T cells — which recognize and destroy infected cells — begin proliferating rapidly.

HIV levels in the bloodstream peak 21 to 28 days after infection, and T cells spike a week or two later. As T cells accumulate, they begin to reduce the amount of virus in the bloodstream.

With some viruses, this is where the dance ends. “There are many infections we have where T-cell immune responses control the virus,” says Sir Andrew McMichael, professor of immunology at Oxford University's Weatherall Institute of Molecular Medicine. In these cases — chicken pox, for example — the virus remains in the body, but usually does not cause long-term problems.

In most people infected with HIV, however, killer T cells become progressively less effective. HIV has an almost limitless capacity to mutate, and in effect is constantly trying on new disguises — turning the dance into a masquerade.

Because T cells respond to HIV, it is tempting to conclude that the virus leads the dance. The immune system also exerts a subtle influence,

however, shaping the evolution of the virus. When virus levels initially peak, HIV is relatively homogeneous, with millions of identical copies of the infecting strain. Last year, researchers showed that the viral sequence later evolves precisely in the parts to which the T cells respond¹.

“The mutants [are] being selected very rapidly, within days, right in front of your eyes,” says McMichael, who led the study. Within weeks, the body is teeming with ‘escape mutants’ that the T cells must learn to recognize.

Handicapped defence

From the beginning, HIV selectively targets and kills CD4 or helper T cells, “the master coordinators of the host immune response,” notes John Moore, professor of immunology and microbiology at Weill Cornell Medical College in New York. “So you have this hole in the immune repertoire.”

Eventually, the depletion of CD4 cells leaves the body susceptible to pathogens, signalling the progression to AIDS. Even early on, however, the CD4 cells are dysfunctional, impairing the immune response to HIV. Without functioning helper T cells, newborn killer T cells fail to mature, or die prematurely, and so the body becomes progressively less able to generate new ones as the virus evolves.

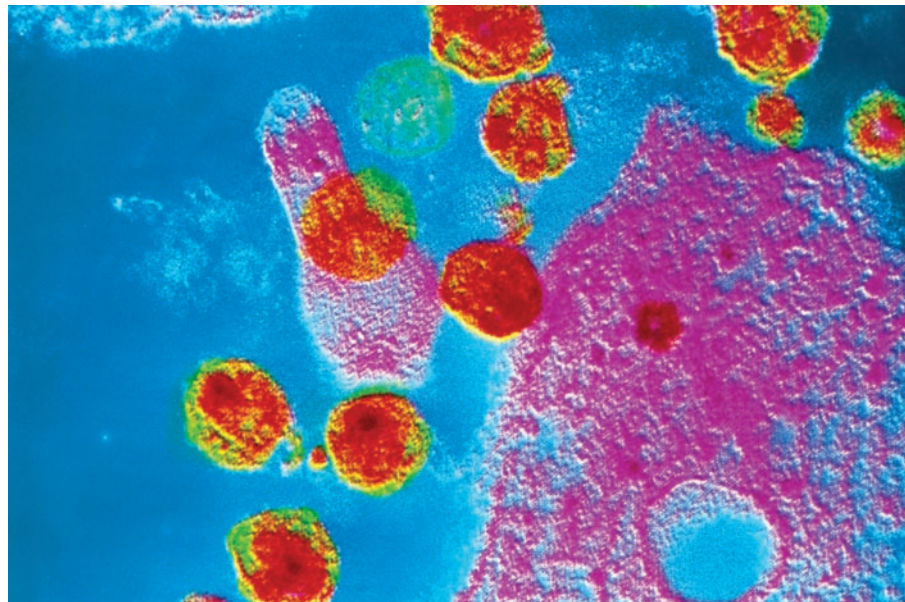
Three to six months after infection, the level of HIV in the bloodstream stabilizes to a ‘set point’. This predicts the course of an individual's disease: a higher set point usually means a quick progression to AIDS, and a lower set point means a slower sequence.

What defines the set point is not fully understood, but “there's some sort of balance between the nature or virulence of the virus and the immune response of the host,” McMichael says. For example, a highly virulent strain of HIV is likely to result in a high set point even if the immune response is strong.

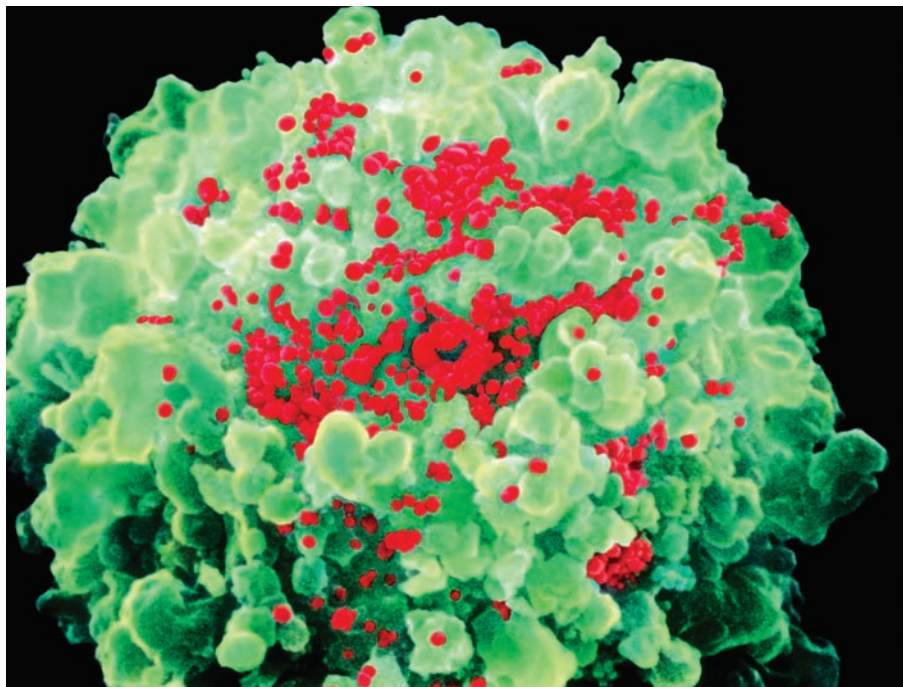
The varying strength of the immune response can be traced (see sidebar) in large part to human leukocyte antigen (HLA) molecules. Three types of HLA class I molecule, HLA-A, HLA-B and HLA-C, are found on the surface of every host cell. These HLA molecules hold on to protein scraps from the virus and signal to the killer T cells that the cell is infected.

The genes for these HLA molecules are the most variable in the genome — each has hundreds of different alleles. This variability, reshuffled each generation, allows the body to recognize a huge variety of pathogens.

The *HLA-B* gene seems to have the strongest overall effect on HIV, says Mary Carrington, head of the HLA Immunogenetics Section at the US National Cancer Institute. Those who carry some *HLA-B* alleles, such as *B27*, *B57*



HIV (red) selectively targets and kills CD4 immune cells (pink), the master coordinators of the immune response.



Immune cells (like the one above, in green) exert a subtle influence on the evolution of the HIV virus (red).

and B58, tend to develop AIDS slowly. Other alleles, such as B07 and B35, are associated with a rapid progression.

Those alleles that offer protection bind to highly conserved HIV proteins that are essential to the virus' function. If the virus mutates in these proteins, it often does so at a cost, becoming less virulent — and contributing to a lower set point. In contrast, less protective alleles bind pieces of proteins that HIV “can mutate away and not have any effect on viral fitness,” Carrington says.

Slow burn

The dance with HIV includes another set of partners: antibodies that recognize and bind to viral proteins. Antibodies appear early in HIV infection, at a similar time to killer T cells. The early antibodies are not effective, however, and do not reduce virus levels in the bloodstream.

Antibodies that neutralize HIV and prevent it from infecting cells are produced about three months after exposure. This is relatively late compared with the response to many viral infections, and might be partly due to viral damage to B cells and the germinal centres where they mature.

These problems have long been known to occur in chronic HIV infection; however, researchers showed last year that they begin in the first weeks or even days of infection². “It was surprising to find that there was such early destruction and damage to germinal centres,” says lead investigator Barton Haynes, director of the National Institutes of Health-funded Center for HIV/AIDS Vaccine Immunology (CHAVI).

When neutralizing antibodies finally are produced, the virus mutates to evade them, the body fashions a new response, the virus escapes again, “and so on, in an endless ‘to-ing and fro-ing’ between the antibody system and the virus,” says Dennis Burton, professor of immunology and microbial science at the Scripps Research Institute in La Jolla, California. “But there isn’t any great impact [of the antibodies] as far as we can see on the overall virus replication.”

“There’s some sort of balance between the nature or virulence of the virus and the immune response of the host.”

Two to three years after infection, about one in every four or five people infected with HIV begins to make broadly neutralizing antibodies, which inactivate many different strains. Ironically, by the time they make the antibodies, these people have too much virus, too many flaws in the immune response and too many reservoirs for the virus to benefit from them. What is more, HIV can mutate to evade these antibodies much faster than the body can develop new ones.

Still, broadly neutralizing antibodies are something of a ‘holy grail’ for vaccine researchers. “They’re the sorts of antibody responses that we’d like to elicit or induce through vaccination,” says Burton.

Two large-scale efforts are under way, one coordinated by CHAVI and the other by the International AIDS Vaccine Initiative, to understand better how these antibodies are made. For each effort, researchers took blood samples from hundreds of people every few months, beginning a few weeks after HIV infection.

The researchers plan to isolate the broadest, most potent antibodies from the participants, then retrace how the virus and antibody

Genetic invitation

For years, scientists have been trying to understand why some people with HIV quickly develop AIDS and others seem more resistant. One method, genome-wide association, involves scanning bits of DNA scattered throughout the genome to look for variations more common in the latter group.

A handful of studies have identified several single-nucleotide polymorphisms (SNPs) associated with better control of HIV. Most variants are in human leukocyte antigen (HLA) molecules, which help the immune system to recognize infected cells.

For example, the first genome-wide association study of HIV linked a variant of the *HLA-C* gene to a stronger immune response and slower progression to AIDS³. Variations in *HLA-B* also seem to delay disease course.

Results from an international study of about 2,000 HIV-infected individuals, expected to be published later this year, confirm the importance of the HLA system, says Mary Carrington, head of the HLA Immunogenetics Section at the US National Cancer Institute.

These studies reliably pick up common variations present in at least 4% of the population. Rarer variants might also affect HIV control, notes David Goldstein, director of the Center for Human Genome Variation at Duke University. To identify rare variants, he argues, researchers should compare whole genome sequences rather than looking at markers for differences.

Goldstein is planning to sequence the genomes of at least 50 people who have been infected with HIV for years but have not developed AIDS. He also plans to compare the genomes of people who progress relatively rapidly with those who progress slowly.

Although the high cost and slow speed make whole-genome sequencing difficult, both limitations are rapidly diminishing, Goldstein says. “A lot of this story is going to be told over the next two or three years.” —S.D.

response have evolved. “The exciting thing for me is that more and more people are becoming interested in the host side of this,” says Haynes. “We’re going to have to understand this area of biology before [an HIV] vaccine is going to be made.”

Sarah DeWeerd is a freelance writer in Seattle.

1. Goonetilleke, N. et al. *J. Exp. Med.* **206**, 1253–1272 (2009).
2. Levesque, M. C. et al. *PLoS Med.* **6**, e1000107 (2009).
3. Fellay, J. et al. *Science* **317**, 944–947 (2007).

Tiny steps towards an HIV vaccine

Recent successes are reinvigorating research into a vaccine for HIV, reports **Cassandra Willyard**.

Last September, after more than 20 years of disheartening research, HIV vaccine researchers had their first, albeit small, taste of success.

In a clinical trial of about 16,000 people in Thailand, the candidate vaccine RV144 lowered the rate of HIV infection by about 30%.

RV144 remains far from market, but the trial gave the field a badly needed boost. After Merck's much-hyped V520 failed in 2007, many researchers questioned whether a successful vaccine could be developed.

"We showed for the first time that it's possible," says Jerome Kim, deputy director of science at the US Military HIV Research Program and a lead investigator on the RV144 trial.

No one yet knows why RV144 worked or how to improve on it; however, this glimmer of success, together with recent clues to other ways to boost the immune response to HIV, has researchers feeling optimistic. "This really is a renaissance," says Wayne Koff, chief scientific officer of the International AIDS Vaccine Initiative (IAVI) in New York.

Thai mystery

Vaccines generally rely on two tactics: tricking the immune system into producing antibodies

that can bind to key parts of a virus and prevent it from infecting cells; and activating the immune system's cellular response, which targets and destroys infected cells.

The trouble is that HIV varies widely — even within an infected individual. HIV infection tends to elicit antibodies and immune cells that are specific to one particular strain, and no help against others.

This variability defeated vaccine efforts until RV144. In this combination formula, ALVAC (developed by Sanofi Pasteur) is designed to elicit a cellular immune response, whereas AIDSVAX (made by Global Solutions for Infectious Diseases) is designed to induce production of antibodies¹.

The odds were against this combination from the beginning: ALVAC's efficacy had never been tested alone, and AIDSVAX had flopped in two previous trials. One group of leading researchers was so sure of failure they called on the government to scrap the study. So why did the combination succeed?

Barton Haynes, head of the National Institutes of Health (NIH)-funded Center for HIV/AIDS Vaccine Immunology, is probing this mystery by analysing blood samples from trial participants. "We need to study this trial to try and understand everything we possibly can about what happened and why," Haynes says.

Genetic diversity

Meanwhile researchers are hunting for 'broadly neutralizing antibodies,' which can bind to numerous HIV strains and block them from infecting cells. A vaccine that could trick the body into producing these antibodies might prevent infection.

In 2009, researchers from IAVI and The Scripps Research Institute in La Jolla, California, isolated two broadly neutralizing antibodies from the blood of an HIV-positive individual in Africa². Last October, at a conference in Paris, NIH researchers announced three more antibodies that bind to a new target on the virus. "We've clearly broken through a

wall," says Gary Nabel, head of the NIH's Vaccine Research Center.

Isolating the antibodies is only half the challenge. How can the vaccine prompt the body to produce them? "That's the \$64,000 question," says Dennis Burton, director of IAVI's Neutralizing Antibody Consortium, and professor of immunology and microbial science at Scripps.

Burton and others are trying 'reverse engineering' techniques — working backwards from the antibody structure — to build molecules that stimulate the body to produce them.

In March, at a meeting in Banff, Canada, Merck researchers announced that they have designed a molecule that elicits a neutralizing antibody in guinea pigs and rabbits. It is the first time that "this kind of reverse engineering process has been validated on the biologic side," says Koff, "so it's a real advance."

Some groups are exploring strategies that bypass the immune system. For example, a team at the Children's Hospital of Philadelphia

has devised a way to insert the genes for these antibodies into cells using an engineered virus. These recombinant vectors produce enough antibodies to protect monkeys from infection with an HIV-like virus³. The team plans to test the concept in people.

A good vaccine would also boost the cellular immune response. To overcome HIV's diversity, researchers are testing artificial genetic sequences that mimic parts of its genome but cover the widest possible array of strains.

"These sequences are not [ones] that are found in their entirety in any naturally occurring virus," says Dan Barouch, chief of vaccine research at Beth Israel Deaconess Medical Center in Boston, Massachusetts.

Injected into rhesus monkeys, these 'mosaic' sequences incite an immune response broader and deeper than with the natural virus^{4,5}. A clinical safety trial is planned, Haynes says.

Although these studies seem promising, continuing obstacles make HIV researchers loath to predict when a vaccine might be ready.

"What you can say is that it would be a major miracle if we had one in less than ten years," Nabel says. "On the other hand, we're doing everything we can to surprise ourselves." ■

Cassandra Willyard is a freelance writer in New York.

"We're doing everything we can to surprise ourselves."



Although no one yet knows why, a candidate vaccine called RV144 seems to prevent HIV infections.

1. Rerks-Ngarm, S. *et al. N. Engl. J. Med.* **361**, 2209–2220 (2009).
2. Walker, L. M. *et al. Science* **326**, 285–289 (2009).
3. Johnson, P. R. *et al. Nature Med.* **15**, 901–906 (2009).
4. Barouch, D. H. *et al. Nature Med.* **16**, 319–323 (2010).
5. Santraand, S. *et al. Nature Med.* **16**, 324–328 (2010).

Joining forces

No single strategy alone is likely to thwart HIV's spread. Researchers are turning to 'prevention packages' of two or more approaches, **Cassandra Willyard** reports.

Developing a vaccine against HIV has proven infinitely complex, but designing alternative methods of prevention has not been any easier. Of the handful of strategies tested in recent years, only male circumcision has been an unequivocal success. Older strategies, such as condoms and needle exchange, all sounded more promising than they turned out to be.

After decades of disappointment, researchers are turning to antiretroviral therapy — the one tool proven to stop the virus in its tracks. Trials are under way to test whether oral antiretroviral pills or gels laced with the drugs can prevent infections. Researchers are also testing whether treating infected individuals early can make them less infectious (see sidebar on page S12).

"We are in a new antiretroviral-based prevention generation," says Mitchell Warren, director of the AIDS Vaccine Advocacy Coalition, a New York-based non-profit organization aimed at accelerating HIV-prevention research.

Scientists are also assessing combinations of two or more of the strategies — for example condoms, needle exchange, male circumcision and antiretroviral drugs. "Because no new prevention intervention will be a silver bullet, we must understand the optimal combinations of interventions," says Stefano Bertozzi, HIV director for the Bill & Melinda Gates Foundation's Global Health Program.

The ideal prevention strategy is cheap, effective and easy to use. Although condoms meet these requirements, many men are reluctant to use them. In the early 1990s, researchers began thinking about strategies that women could control. They estimated that a microbicide that is 60% effective at preventing HIV infection could prevent 2.5 million infections in just three years¹.

Unfortunately, the microbicide field has been plagued by failures. Scientists have developed dozens of candidates, and led at least six into clinical trials. Despite promising preclinical data, however, each has failed to prevent infection. Alarming, a couple even

appeared to increase women's risk of contracting the virus.

In December, researchers reported the most recent failure: a gel called PRO2000, which had shown promise in a small study, but had no effect on HIV transmission in a trial of 9,385 African women.

Preventive pills

These disappointing results were not a major setback, says Sharon Hillier, professor of obstetrics, gynaecology and reproductive sciences at the University of Pittsburgh in Pennsylvania. "We had already moved on to the next generation of products, which are much more potent."

The new generation of microbicides, two of which are in clinical trials, are laced with potent antiretroviral drugs. Results from the first trial, a 900-person test of a vaginal gel containing tenofovir, are expected to be announced in July at the International AIDS Conference in Vienna.

The preventive power of antiretroviral pills has been harnessed before, to block transmission of the virus from mother to unborn child. For example, a single dose of the antiretroviral nevirapine, given to the mother during labour and to the infant after birth, cuts HIV transmission by more than 40%.

Scientists reason that the pills might also prevent the virus from taking hold in uninfected adults. This pre-exposure prophylaxis (PrEP) strategy is being tested in at least five independent trials. The first results are expected later this year. "We are going to get our first glimpse at whether these new approaches with antiretrovirals work," Warren says.

For such methods to be effective, they must be used regularly — daily or before every risky sexual encounter. That might be feasible in the controlled setting of a clinical trial, but impractical in the real world, notes Robin Shattock, professor of cellular and molecular infection at St George's, University of London.

Some researchers are devising ways to make these methods easier to use. For example, a



This Chinese device, a ShangRing, holds the foreskin in place, allowing it to be snipped away quickly with no stitches required.

vaginal ring — much like the ones used for contraception — could deliver low doses of the drug for a month or more. Researchers at Weill Cornell Medical College in New York have designed a ring that delivers both antiretroviral drugs and contraceptives, although it has not been tested outside the laboratory².

Similarly, antiretroviral injections that last for several months, like the contraceptive injection Depo-Provera, might be preferable to a daily pill. "The less you have to think about it, the more likely it will be adopted," Shattock says.

Aggressive strategies

Scientists have known for more than a decade that people who have low viral loads are less likely to pass on the virus. Some studies have shown that treating HIV-infected individuals early decreases their risk of infecting their partners. One research team is testing this approach — dubbed 'treatment as prevention' — in a clinical trial of 1,750 couples.

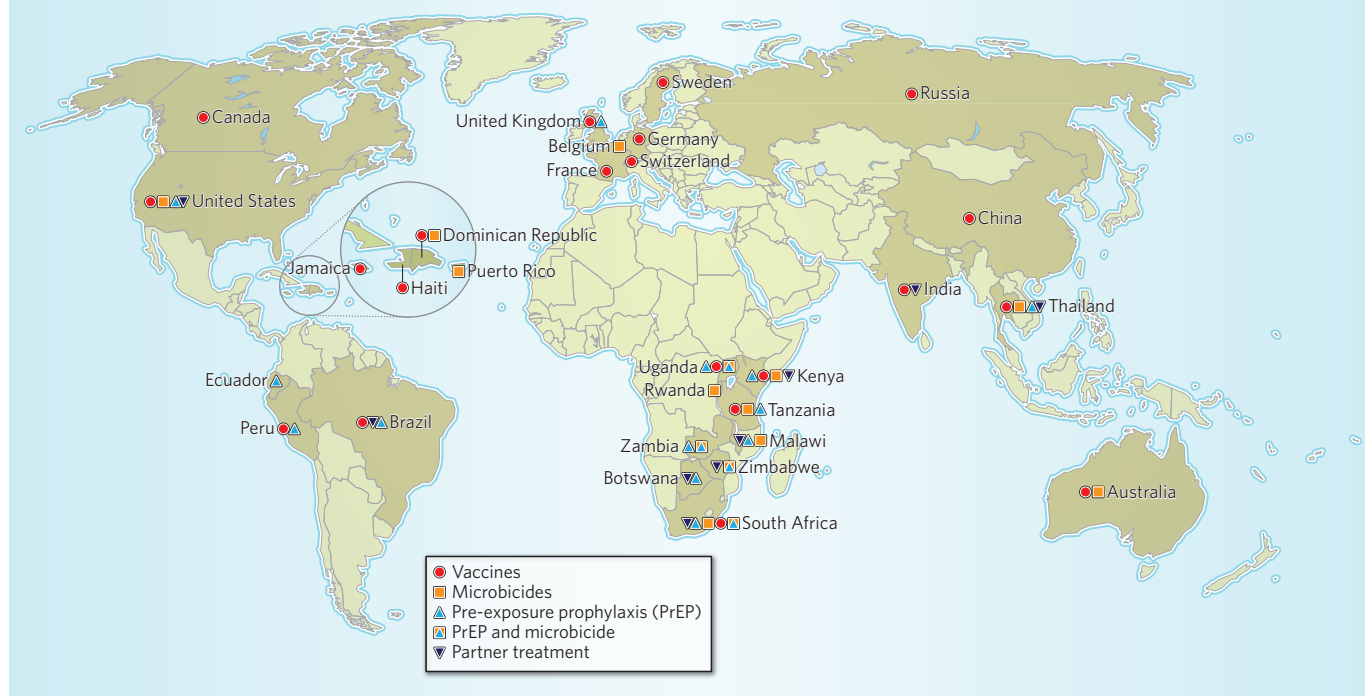
Last year, the World Health Organization (WHO) proposed that an aggressive global strategy of yearly universal HIV tests and immediate antiretroviral treatment for those infected could dramatically slow the spread of HIV within a decade, and reduce prevalence to 1% within 50 years³.

Many experts argue, however, that this 'test and treat' approach will be difficult to implement. "I think it will be a challenge even in the US, let alone in places with a much higher HIV burden," says Connie Celum, director of the University of Washington's International Clinical Research Center in Seattle. "How do you think about test and treat when it's such a long way to go to even treat those who will die in a few years if they don't get [antiretrovirals]?"

The US National Institutes of Health (NIH) is planning to test a less-regimented strategy in Washington DC and New York City's Bronx neighbourhood. The three-year study, set to

"Because no new prevention intervention will be a silver bullet, we must understand the optimal combinations of interventions."

ONGOING TRIALS OF NEW PREVENTION OPTIONS WORLDWIDE



Scientists are testing every preventive approach available, including microbicides, vaccines and antiretroviral pills, alone and in combination.

begin in June, will rely on local health centres. The researchers plan to encourage HIV testing through media campaigns, and to help those who test positive to find doctors to prescribe treatment. “We really want real-world data,” says David Burns, branch chief for HIV prevention research at the US National Institute of Allergy and Infectious Diseases.

Some interventions take longer to gain traction than others. In the case of male circumcision, for example, lack of political will, limited resources and even cultural mores have deterred some communities from adopting the procedure. “Nothing gets to the heart of social cultural issues as much as a man’s penis and his foreskin,” Warren says.

Three independent clinical trials showed, in 2005 and 2006, that circumcision cuts men’s risk of contracting HIV by about one-half. Consequently, the WHO, in 2007, began recommending male circumcision for HIV prevention, and the strategy is slowly gaining ground.

At least 8 of the 13 African nations identified as priority countries in 2007 by the WHO and the United Nations Joint Programme on HIV/AIDS (UNAIDS) are scaling up male circumcision services. Kenya, which launched its campaign in 2008, has made the most progress. As of February 2010, doctors there had circumcised 90,000 men. The government aims to reach a million men — nearly all those uncircumcised in the country — by 2013. South Africa, in April, launched free male circumcision services as part of its national HIV campaign. The country’s ministry of health aims to help 2.5 million men by 2015.

Cutting edge

In many countries, there is a dearth of trained professionals who can perform the surgical procedure. Some research teams have set up training camps, and others are investigating ways to make the procedure easier and safer to perform.

One clinical trial, organized by researchers at Weill Medical College of Cornell University in New York, is testing a Chinese device called ShangRing, which consists of two plastic rings that sandwich the foreskin, allowing it to be snipped away quickly. A typical circumcision takes about 20 to 30 minutes, but ShangRing cuts that down to about 5 minutes — with no stitches required. However, the device must be left in place for ten days to allow the wound to heal.

Even if millions of men sign up, circumcision neither eliminates their risk of infection nor directly benefits women. “It can’t be considered as the sole method of prevention,” says Maria Wawer, professor of population, family and reproductive health at Johns Hopkins University.

In fact, no single strategy — not even a vaccine — is likely to be 100% effective. In some cases, “two interventions applied together can be greater than the sum of their parts,” says Timothy Hallett, a research fellow in the department of infectious disease epidemiology at Imperial College London. There are few hard data, however, on which combinations might be best for a given population.

Last year, the NIH funded six groups to design “prevention packages” for specific populations — including injecting drug

users in Eastern Europe, men who have sex with men in the Americas, and households in Uganda. The institute plans to fund several more this year.

Celum, who is designing a package for Uganda, is investigating the feasibility of sending health workers door-to-door to test for HIV. Based on the results, they could develop an on-the-spot “prevention prescription,” she says.

Real-world tests of these combinations are not always feasible. Clinical trials would cost hundreds of thousands of US dollars and take years to yield results. What is more, says Burns, “it quickly becomes a very complex research design.” The NIH projects will be assessed using small pilot studies, with successful strategies tested in larger clinical trials.

Mathematical models offer a quick and inexpensive way to skip some cumbersome steps. “You can simulate the impact of an intervention,” says Hallett, who is working with Celum on the Uganda project.

These new approaches require unprecedented levels of collaboration between immunologists, virologists, modellers, and behavioural and social scientists.

“People are moving beyond the [attitude that], ‘There’s only my approach that’s going to work,’” Celum says. “The move towards combination prevention is breaking down some of those divides.”

Cassandra Willyard is a freelance writer in New York City.

1. Watts, C. & Vickerman, P. *AIDS* **15**, S43–S44 (2001).
2. Saxena, B. B. et al. *AIDS* **23**, 917–922 (2009).
3. Granich, R. M. et al. *Lancet* **373**, 48–57 (2009).

The outlook for a cure

There is a formidable arsenal of drugs available to treat HIV. **Virginia Hughes** finds that, for the first time in years, there is also renewed hope of a cure.

Not long ago, you would have been hard-pressed to find an HIV researcher who would utter the word 'cure'.

HIV has a remarkable ability to resist anti-viral drugs and hide in the body, so the idea of eradicating the virus seemed impossible. Suggesting otherwise, researchers feared, could create false hope and complacency.

In the past few years, however, there are increasingly loud whispers about a cure for HIV. The year 2007 saw the clinical debut of integrase inhibitors, which prevent HIV from inserting into the host genome. The following year, a bone-marrow transplant eliminated the virus from the body of an infected German man (see sidebar, *Interfering with genes*). Last year, breakthroughs in cell-culture techniques allowed researchers to screen for drugs that can lure HIV from its hiding places.

"I was very pessimistic five years ago, but we had to try," says Warner Greene, director of the Gladstone Institute of Virology and Immunology in San Francisco. "And as we've tried, I've become much more optimistic that we might be able to achieve a drug-free remission."

His optimism is understandably tinged with caution, however, as scientists promising eradication were proven wrong once before.

In the summer of 1996, researchers attending the eleventh international AIDS meeting in Vancouver trumpeted data showing that certain combinations of antiretroviral drugs can suppress HIV to undetectable levels in the blood.

The buzz grew over the course of the following year. In May 1997, David Ho's group at the Aaron Diamond AIDS Research Institute in New York reported in *Nature* that HIV levels in eight people dropped by two orders of magnitude within ten days of receiving a particular three-drug combination and were undetectable within eight weeks¹.

Using a mathematical model, the researchers estimated that, barring any complications, the drugs could eradicate the virus from an infected person in less than three years.

If that sounded too good to be true, it was. In the same issue of the journal, Bob Siliciano's group from Johns Hopkins University inspected the small number of dormant immune cells that harbour HIV. Because these cells do not replicate, they are impervious to treatment. Once activated, however, his team found that these cells can start pumping out the virus into the blood and lymph nodes².

Two other groups described these so-called latent reservoirs in 1997, and two years later Siliciano's team estimated that it would take about 60 years for drugs to flush out HIV from these stores³. "It was a real blow, I think, to people who were interested in eradication," Siliciano says.

The field instead shifted focus to preventing resistance by using combinations of three or more drugs — dubbed highly active antiretroviral therapy (HAART) — and to decreasing the side effects of treatment.

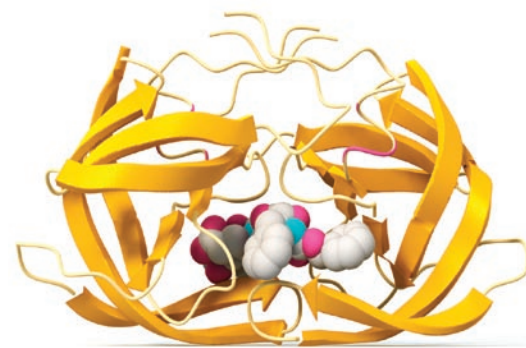
These efforts were hugely successful: there are 32 approved HIV drugs, and at least a dozen more in the pipeline (Table 1), which, together, can suppress the virus for decades.

"Now that HAART works so well," Siliciano says, "we're turning to the next step: can we actually cure anybody?"

In the pipeline

Like most viruses, HIV hijacks its host's cellular machinery to replicate. HIV is a member of a particularly nefarious family of viruses, however, which insert their DNA into the host genome. When the host cell replicates, its daughters make more virus.

Antiretroviral drugs target different stages of this replication cycle, limiting the viral load and allowing immune cells a chance to clear out infected cells. The most widely used drugs block the enzymes that HIV needs to infect new cells. The standard regimen includes two drugs that block reverse transcriptase, which HIV needs to convert its RNA genome into DNA, and one protease inhibitor that prevents viral particles from maturing.



Protease inhibitor drugs (spheres) bind to a viral enzyme (yellow) and prevent HIV particles from maturing.

In 2007, Merck released raltegravir. This was the first drug to target HIV's integrase enzyme, which stitches the viral DNA into the host genome.

Clinicians welcomed drug cocktails spiked with raltegravir after large clinical trials showed they trounce drug-resistant strains of HIV, and suppress virus levels significantly faster than do standard combinations. Several other integrase inhibitors are in the pipeline, and early data suggest that they are better than raltegravir.

"The number of papers on HIV integrase inhibitors has just boomed over the last two years," notes Yves Pommier, chief of the Laboratory of Molecular Pharmacology at the US National Cancer Institute. "It's marvellous that now we have very effective inhibitors for all the three HIV enzymes — it will be very hard for the virus to escape them all."

Other HIV proteins might also make good targets. In 2002, Michael Malim and colleagues fired up the field with the discovery that one of these proteins, Vif, degrades a human enzyme, APOBEC3G, which evolved eons ago to damage viral DNA⁴. Without Vif, APOBEC3G would block HIV replication.

That basic research is starting to pay off, and many scientists are hunting for drugs that block Vif. In 2008, Greene's group at the Gladstone Institute launched a collaboration with Gilead Sciences to find Vif inhibitors.

Another tactic is to target the host cell, rather than the virus. HIV primarily attacks CD4 T cells, which normally help the immune system fend off invaders. A handful of new compounds block CCR5, which is a receptor on the surface of CD4 cells that HIV must bind in order to penetrate the cell.

In 2007, Pfizer debuted the first CCR5 blocker, maraviroc, which is effective at keeping

"It's marvellous that now we have very effective inhibitors for all the three HIV enzymes — it will be very hard for the virus to escape them all."

Interfering with genes

In 2008, doctors in Germany made an astounding announcement: they had wiped out HIV from a middle-aged man when they gave him a bone-marrow transplant to treat his leukaemia.

The doctors had replaced the bone marrow with that of a donor who carried a genetic variation that disrupts the function of the CCR5 receptor in T cells. When this molecular doorway is broken, HIV cannot enter host cells. To this day, despite having stopped all antiretroviral therapy, the man has no detectable HIV in his blood.

Bone-marrow transplants are not a practical option for treating HIV infection: upwards of one in three people who receive them die, and they cost at least US\$150,000. Still, the case offers hope for scientists using gene therapy to tinker with CCR5, and perhaps other genes involved in HIV replication.

For example, RNA interference, which is a method of silencing genes, has been shown to prevent the virus from replicating and mutating in cultured T cells. Researchers have also launched several clinical trials using methods to modify the CCR5 gene in host cells.

Carl June and his colleagues at the University of Pennsylvania are harvesting T cells from HIV-infected individuals and disrupting the CCR5 gene using zinc fingers — protein components that recognize specific DNA sequences and turn genes on or off. Preliminary results from the 18-person trial are expected in March 2011.

At a conference in January this year, June reported that the gene therapy has allowed one participant who is off antiretroviral treatment to maintain undetectable levels of virus for two weeks longer than is typical.

Last year, the first randomized and placebo-controlled test of a gene therapy for HIV showed that a technique that uses ribozymes — RNAs that cut other kinds of RNA — to disrupt HIV genes is safe, but does not reduce viral load.

Although these methods have not yet been proven effective, the potential benefits have encouraged researchers.

"Drugs have to be taken every day, but gene therapy could do the same thing in a one-time treatment," notes Ben Berkhout, professor of virology at the University of Amsterdam. —V.H.

Table 1 | New HIV drugs

Drug	Type	Manufacturer	Status
Rilpivirine	NNRTI	Tibotec	Phase III
UK-453061	NNRTI	ViiV Healthcare	Phase II
IDX889	NNRTI	ViiV Healthcare	Phase II
Apricitabine	NRTI	Avexa	Phase III
Amdoxovir	NRTI	RFS Pharma	Phase II
Maraviroc	CCR5 antagonist	ViiV Healthcare	Approved in 2007; marketed as Selzentry/Celsentri
Vicriviroc	CCR5 antagonist	Schering-Plough	Phase III
PF-232798	CCR5 antagonist	ViiV Healthcare	Phase II
PRO 140	CCR5 inhibitor (monoclonal antibody)	Progenics Pharmaceuticals	Phase II
Raltegravir	INI	Merck & Co., Inc.	Approved in 2007; marketed as Isentress
Elvitegravir	INI	Gilead Sciences	Phase III
GSK1348572	INI	ViiV Healthcare/Shionogi	Phase II
GSK1265744	INI	ViiV Healthcare/Shionogi	Phase II
Bevirimat	Maturation inhibitor	Myriad Pharmaceuticals	Phase II
Ibalizumab	Entry inhibitor (monoclonal antibody)	TaiMed Biologics	Phase II
Unknown	Vpu inhibitors	Biotron	Preclinical stage
Unknown	Vif inhibitors	Gilead Sciences/Gladstone Institute	Preclinical stage
Truvada	Combination pill (two reverse transcriptase inhibitors)	Gilead Sciences	Approved in 2004
Atripla	Combination pill (Truvada + reverse transcriptase inhibitor)	Gilead Sciences/Bristol-Myers Squibb	Approved in 2006
'Quad' pill	Combination pill (Truvada + elvitegravir + boosting agent)	Gilead Sciences	Phase III

INI, integrase inhibitor; NNRTI, non-nucleoside reverse transcriptase analogue; NRTI, nucleoside reverse transcriptase analogue.

viral levels in check. However, people who have been on HAART for long periods tend to carry HIV strains that use the CXCR4 receptor instead of CCR5. Drugs that block CXCR4 are also being developed, but animal studies suggest the drugs are toxic.

Daily dose

No matter how effective the available drugs, the harsh reality is that they must be taken for life — although when to start them is an equivocal point (see sidebar, Perfect timing). If treatment is interrupted, HIV rapidly springs back into action, rising to detectable levels within weeks.

Partly to make taking the medications more palatable, and partly to keep the HIV drug arsenal stocked with new compounds, many companies have invested in making pills that combine multiple drugs.

"Taking one pill a day is much easier than taking two or three," says Michael Mullen, acting chief of infectious diseases at Mount Sinai Medical Center in New York. "Pharma is realizing that the most important challenge in antiretroviral therapy is improving adherence by simplifying regimens."

So far, only the commercially available pill Atripla — one of the most widely prescribed

HIV drugs in the United States — combines drugs from different classes. More are on the way, however.

Last year, GlaxoSmithKline and Pfizer together launched ViiV Healthcare, a company that will focus on creating new antiretroviral medicines, particularly combination pills. In April, Gilead Sciences began phase III testing of its 'Quad' pill, which combines two reverse transcriptase inhibitors and one integrase inhibitor.

These combinations might make taking the pills less cumbersome, but they will not be able to alleviate the harsh side effects. Depending on when therapy is begun, the lifespan for those on HAART is 10 to 30 years shorter than average. They also suffer from a host of conditions including heart, kidney, liver and bone disease, cancers and serious cognitive problems.

Some studies suggest that these complications are more the result of a sustained inflammatory response to the virus than of drug toxicity. In either case, the chronic problems are a natural consequence of long-term infection.

The ideal HIV drugs would not just suppress the virus, but would eliminate it from the body — something that is just beginning to look feasible.



Although antiretroviral drugs are effective, they must be taken for life, and can cause harsh side effects.

Elusive target

To eradicate the virus from the body, scientists must first pinpoint its hiding places. HIV is particularly good at staying invisible. In 1997, three teams independently discovered one HIV reservoir: resting memory T cells. These cells stay quiet for decades, and are activated only when the immune system encounters an invader that it has seen before.

An HIV-infected individual carries about one million infected resting memory cells. However, while in this state, these cells are invisible to the immune system. The only way to destroy the reservoir, scientists reason, is to stimulate the cells to begin making virus, thereby rendering them vulnerable to antiretroviral drugs.

In the late 1990s, researchers used agents such as interleukin-2 growth factor, which activates all T cells, to prod the cells out of their resting state. Provoking a global immune response is dangerous, however, and can trigger massive leakage of fluids into tissues.

One alternative is to tickle the latent cells so that they express HIV proteins without replicating themselves. For example, inhibitors of histone deacetylases — enzymes that suppress HIV transcription — could stimulate latent cells to produce HIV.

Screening for such drugs is technically challenging because resting T cells tend to die in culture. In the past year, however, several groups have genetically or chemically engineered the cells so that they can survive longer in culture, allowing scientists to screen for drugs that target the viral reservoirs.

"It's actually fairly easy to find compounds that turn on latent HIV without causing global T-cell activation," says Siliciano. "We've already

found several." Although the compounds identified so far are probably too toxic for use in people, he notes, it is encouraging that the technique works on the small scale.

There might also be many other reservoirs in which HIV replicates at low levels. These could be blood stem cells, other immune cells such as macrophages, or inaccessible caves such as the brain or the gastrointestinal tract.

"If we're going to come up with an eradication strategy, it's not going to be just as simple as purging the virus from T cells," says Mario Stevenson, professor of molecular medicine at the University of Massachusetts.

Last year, several leading researchers called for a large collaboration involving academia, industry and the government to identify HIV's hiding places and investigate ways to drag the virus out of them⁵ (see page S21).

The ultimate test for any treatment that attempts to clear out the reservoirs is to take patients off therapy and see whether their viral loads stay in check — which some deem unethical.

Even the most enthusiastic researchers admit that the field is far from understanding latency, and is at least a decade away from producing compounds that could clear out the viral reservoirs. If history is any guide, however, they will not stop trying.

"Scientists are stubborn," Stevenson says. "That persistence on the part of the scientific community hopefully exceeds the persistent qualities of the virus."

Virginia Hughes is a freelance writer in New York City.

1. Perelson, A. S. *et al. Nature* **387**, 188–191 (1997).
2. Chun, T. K. *et al. Nature* **387**, 183–188 (1997).
3. Finzi, D. *et al. Nature Med.* **5**, 512–517 (1999).
4. Sheehy, A. M. *et al. Nature* **418**, 646–650 (2002).
5. Richman, D. D. *et al. Science* **323**, 1304–1307 (2009).

Perfect timing

How early should someone infected with HIV begin treatment? Doctors in the developed world recommend drugs when the number of an infected individual's CD4 immune cells falls to less than 350 per cubic millimetre of blood. Many doctors in the United States say treatment should begin as soon as HIV infection is diagnosed, even if CD4 counts are normal.

"This very aggressive approach is new, and it's based on non-definitive data," says Steven Deeks, professor of medicine at the University of California, San Francisco.

Still, Deeks supports early intervention because research has shown that lower CD4 levels pre-therapy lead to more age-related problems. "The longer you wait, the more inflammation, the more immunologic dysfunction and perhaps more of these diseases will occur," he says.

"The longer you wait, the more inflammation, the more immunologic dysfunction and perhaps more of these diseases will occur."

In April 2009, a report in the *New England Journal of Medicine* found that beginning therapy when CD4 levels are less than 350 cells per cubic millimetre of blood carries a 69% higher risk of death compared with levels of 351 to 500 cells per cubic millimetre.

This was an observational study, however, so researchers could not determine whether certain characteristics of people who get early treatment — such as better drug-adherence rates or less recreational drug use — explain why they fare better.

To resolve the debate, an international research team last year launched a randomized, prospective clinical trial, called the Strategic Timing of Antiretroviral Treatment, which is projected to run until 2015.

Even if early treatment is beneficial, cautions Gregg Gonslaves, an AIDS activist who has been on antiretroviral therapy for 15 years, "actual clinical practice may play out entirely differently, in terms of people's ability to stay adherent to medications and deal with side effects." —V.H.



Forgotten lessons

For many people in the developed world HIV is no longer the death sentence it once was. **Paroma Basu** explores the consequences of complacency.

Discussions about the global AIDS epidemic invariably focus on sub-Saharan Africa and other impoverished regions. That is no surprise: sub-Saharan countries alone are home to two-thirds of all HIV-positive people in the world.

It is easy to forget that the epidemic is thriving in more privileged areas, where it silently takes a steadily mounting toll. In 2007, as many as 54,000 adults and children died from AIDS in North America, according to the World Health Organization (WHO).

Since the 1980s, outreach and prevention programmes have helped avert thousands, if not millions, of HIV infections in high-income countries. Readily available antiretroviral drugs have helped turn it into a chronic, manageable condition.

Public-health officials worry that these successes have, however, spawned an epidemic of blasé attitudes. People are less attuned to messages of prevention, and the importance of routine HIV tests and adherence to treatment. Risky behaviours are beginning to re-emerge, even as numbers of HIV infections are on the upswing.

Between 2000 and 2007, for example, the estimated rate of newly reported HIV cases in Europe rose from 39 to 75 per million people, according to the Joint United Nations Programme on HIV/AIDS (UNAIDS).

"We are seeing a wave of resurgence of new infections [in developed countries]," says Paul de Lay, UNAIDS deputy executive director of programmes.

Dangerous attitudes

In the 1980s and early 1990s, when AIDS was associated with an inevitable and harrowing death, messages about HIV prevention carried a sense of urgency.

"No adequate therapies were available [then]," recalls Anthony Fauci, director of the US National Institute of Allergy and Infectious Diseases (NIAID), who cared for some of the earliest HIV patients. "Messages of prevention and being sexually careful were very well heard because there was a big fear factor in society."

In the mid-1990s, pharmaceutical companies developed a slew of protease inhibitors — which block HIV from making copies of itself — and other drugs that cut viral loads. Combinations of these drugs dramatically decreased the number of people who became ill from opportunistic infections or died from AIDS.

More than 30 antiretroviral medications are now available, and dozens more are being developed (see table on page S12).

The newer treatments are also easier to take. In the early days, some HIV-infected people had to take up to 22 pills a day on a strict schedule, some with food, others requiring refrigeration and so on. New combination pills allow people to take as few as four a day and the number is set to decrease even further. For example, Gilead Sciences is developing a 'Quad' pill that packs a full day's anti-HIV regimen into one tablet.

The WHO estimates that at least 695,500 people in high-income countries were on antiretroviral therapy at the end of 2008, and are living longer, more productive lives.

A 2008 analysis in *The Lancet* found that in high-income countries, an HIV-positive 20-year old given proper antiretroviral treatment is likely to live to at least 69 years of age. In the earliest days of the epidemic, patients died, on average, within 26 weeks of infection, says Fauci.

As a result, the pall cast by HIV in these countries has faded — and given way to public complacency.

One survey of 2,554 adults in the United States, released last year by the Henry J. Kaiser Family Foundation, reported that just 6% of Americans say HIV is the most urgent health problem in the country, compared with 44% in 1995.

Those attitudes, combined with the flagging economy, have meant that "funds [for HIV/AIDS prevention] have been reduced and programmes are being cut back," says de Lay.

Many young people do not have a real sense of HIV's dangers, he adds. "Young, vulnerable people are moving into high-risk behaviours, whether heterosexual or homosexual [activity], or injecting drugs."

"We are seeing a wave of resurgence of new infections."



Targeted prevention

The epidemic in developed countries has always been concentrated in high-risk groups, such as men who have sex with men (MSM), sex workers, prison inmates, injecting drug users and immigrants. Social shifts within these populations have created new pockets of infection.

For example, MSM still account for more than one-half of all new infections in the United States, but a higher proportion of the cases are now in homosexual men of African-American and Hispanic descent, who are infected at a much younger age.

Rates of infection in African-American women are 15 times higher than in white women. In Canada, aboriginal women are increasingly testing positive for HIV, and in Europe, migrant populations — predominantly from the African subcontinent — are at higher risk for heterosexually acquired infections.

"Given that HIV is a dynamic epidemic that is continually evolving, we have to think of prevention in a dynamic way, directly reaching out to and engaging communities in a culturally competent manner," says Kevin Fenton, director of the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention at the Centers for Disease Control and Prevention (CDC).

To reach African-Americans, the CDC, in April this year, launched the 'Act Against AIDS



Demonstrators outside the White House in Washington DC protest against declining funds for HIV/AIDS prevention and cure.

Leadership Initiative, a five-year programme that recruits established institutions to the fight against HIV. In February this year, the CDC also initiated the 'I Know' campaign, which encourages young people to talk about HIV on social networking sites.

Prevention campaigns should also be tailored to specific high-risk groups, such as sex workers at truck stops, homosexual men who use crystal methamphetamine or older women married to injecting drug users, experts say.

Health authorities have avoided publicly singling out marginalized populations, notes Bruno Spire, research director of the Marseille branch of the Institut National de la Santé et de la Recherche Médicale. "The solution is to empower these groups, so it's not about prevention for these groups but prevention with these groups," says Spire, who is HIV-positive.

Peer education

Spire and colleagues are experimenting with peer-to-peer outreach programmes, in which high-risk community members educate others. In one preliminary project showing early success, MSM who are not medical professionals were trained to administer diagnostic tests to others in four French cities.

In 2006, the CDC recommended making HIV tests a part of routine medical care.

The following year, UNAIDS and the WHO proposed similar guidelines. A complex web of social, psychological, legal and financial factors prevent many people at risk from getting tested, however.

For example, lack of health insurance, cultural misconceptions and worries over immigration status deter many economic migrants from learning their status. Laws that prosecute individuals who knowingly transmit HIV also prevent high-risk groups from seeking care.

Many perceive themselves — often wrongly — as being at low risk, says Christoforos Malouris, director of programmes at the Amsterdam branch of the Global Network of People Living with HIV. For instance, few prevention or testing schemes target the elderly.

Even in 2010, the stigma associated with HIV cannot be underestimated, adds Malouris, who is HIV-positive.

"[In the MSM community] I hear a lot of rejection because of HIV status and see a lot of 'sero-sorting' of partners according to status, so you see a lot of positive-positive and negative-negative couples," he says. "This is an indication of the ignorance of what is really risk. Instead of learning about it, people try to deal with it by not dealing with it at all."

According to UNAIDS, an estimated 21% of HIV-infected individuals in the United States

and 27% in Canada are unaware of their HIV status. In Europe, up to 38% of infected people are diagnosed after the virus has caused irreparable damage to their immune systems.

"The problem is that when these patients finally do enter care, most are at a very late stage of infection," says Jens Lundgren, director of the Copenhagen HIV Programme and professor of viral diseases at the University of Copenhagen. "This late percentage situation has not improved at all, and remains an absolutely unresolved problem in the developed world."

Early diagnosis is important for several reasons. A growing body of research suggests that early and easy access to treatment and support services can limit the virus' spread.

For example, a widely discussed 2008 analysis by the Swiss National AIDS Commission found that an HIV-positive person with an undetectable viral load after six months of antiretroviral therapy who has no other sexually transmitted infections carries a negligible risk of passing on the virus.

Prevention fatigue

As HIV/AIDS has matured into a chronic disease, developed nations face a host of new concerns, including the care of long-term survivors and the best time to initiate antiretroviral treatment (see sidebar on page S13).

Long-term survivors can grow tired of practicing safe sex, adhering to the demanding drug regimens and informing new partners of their HIV status — what experts call 'prevention fatigue' — notes Koen Block, executive director of the Belgium-based European AIDS Treatment Group.

They also struggle with depression, social isolation and loneliness. One 2006 study of 914 people in New York City found that two-thirds showed symptoms of depression. "We also see very fragile and inadequate social networks for a lot of these folks," says Mark Brennan, senior research scientist at the New York-based AIDS Community Research Initiative of America, which conducted the study.

Medical problems associated with taking antiretroviral drugs for many years are a sobering reminder that, despite all that scientists have learnt, there is a long way to go before infected people can live trouble-free lives.

"Antiretroviral therapy is really one of the great advances of modern medicine — people are getting decades of productive life out of these drugs," notes Steven Deeks, professor of medicine at the University of California, San Francisco. "They're just not yet getting a complete life."

Paroma Basu is a freelance writer in Lausanne, Switzerland.



HIV worldwide

Developing solutions

There is more to combating HIV in the developing world than providing affordable drugs. **T. V. Padma** looks at the innovative new strategies being employed.

"Mambo?"

The single Kiswahili word for "How are you?" arrives in a weekly text message from the AIDS clinic in Nairobi.

From Kajiado, 200 miles away, the clinic's patients, mostly members of pastoral Maa-sai communities, respond with "Sawa" (OK) or "Shida" (problem). If, after two days, the patient does not respond, a nurse follows up with a telephone call. This simple system confirms that patients remember to take their drugs and are feeling well.

The scheme, which began in May 2007, takes advantage of the fact that, even though the roads between Kajiado and Nairobi are poor, the mobile telephone service is inexpensive and reliable.

The texting scheme is the brainchild of researcher Richard Lester, who noticed that about one-half of his patients in Nairobi owned mobile telephones and about 90% had access to a shared telephone.

With funds from the US President's Emergency Plan for AIDS Relief (PEPFAR), Lester and the University of Nairobi in 2007 launched a randomized trial to test whether mobile telephones can help improve follow-up and outcome in patients in remote rural areas. Preliminary analysis of the data suggests that those who participate have lower viral loads, making

them less likely to develop drug resistance or transmit the infection.

"This has huge implications," says Lester, assistant clinical professor at the University of British Columbia. Hospitals and clinics in Kenya and Ethiopia are planning to adopt the scheme, he says.

Mobile telephones are just the latest strategy in the developing world's fight against HIV/AIDS. In the past three decades, these countries have launched awareness campaigns, built testing and counselling facilities, expanded their research capacity, collaborated in international trials and negotiated for better access to drugs. With no vaccine or microbicide in sight, governments are devising innovative approaches, including door-to-door testing and social networking sites.

Innovative approaches

In April this year, South Africa, until recently the 'poster child' for AIDS denialism, launched the country's biggest HIV testing programme. President Jacob Zuma publicly disclosed his HIV status — negative — to lessen the stigma. The government also announced free male-circumcision services as part of its prevention programme.

To those who have been on the front lines of the disease from the beginning, the landscape is dramatically different.

"When as a very young doctor I saw my first HIV patient with Kaposi's sarcoma, I thought it was something unique. I did not anticipate the devastation that would unfold before me," says Salim Abdool Karim, director of the Centre for the AIDS Programme of Research in South Africa (CAPRISA). Yet, "[if] I have seen the devastation, I have also seen the miracle of hope offered by antiretrovirals," Karim says.

Despite undeniable gains, however, the course of the epidemic remains largely unchanged. In sub-Saharan Africa, for every person who gains access to antiretroviral drugs, two get infected with HIV, Karim says.

New complications arise all the time. HIV has revived tuberculosis (TB; see page S18),

making room for deadlier, drug-resistant versions. As in richer nations (see page S14), doctors in some developing countries are seeing early heart attacks and kidney failure in HIV-positive individuals.

"It's like we are running backwards on the treadmill," Karim says. "We are not stemming the tide of the epidemic."

Broken promises

Prevention programmes in most developing countries rely largely on international funds, which are vulnerable to 'donor fatigue' and the global economic downturn.

Despite lofty promises, many donor agencies have not delivered. For example, the Global Fund to Fight AIDS, Tuberculosis and Malaria pledged US\$10 billion a year when it was set up in 2001, but has delivered only US\$3 billion a year so far.

In 2009, funds from US-based charities — except from the Bill & Melinda Gates Foundation — had decreased by 3% since 2007/2008, and funds from

European charities had decreased by 7% since 2006, according to an April 2010 report from the International Treatment Preparedness Coalition of HIV-infected people and their supporters.

"Abandoning the AIDS response now will inevitably lead to a return to headlines about people dying of AIDS that we read at the beginning of the decade," the report warns.

PEPFAR is one of the few schemes to have maintained funding. Launched in 2003 by then-President George Bush, it was extended for a further five years in 2008. This legislation authorizes up to US\$48 billion to combat global HIV/AIDS, TB and malaria.

"In a time of tightening budgets and economic constraints, this request for the eighth year of PEPFAR is the largest request to date in a president's budget," says Eric Goosby, PEPFAR's US global AIDS coordinator.

PEPFAR programmes continue to scale up prevention, treatment and care for HIV/AIDS. According to a September 2009 analysis, PEPFAR has directly supported antiretroviral therapy for more than 2.4 million infected individuals.

As welcome as donor aid is, however, it can skew national priorities. As an example, Karim points to South Africa, where HIV research infrastructure expanded impressively, with international studies on both basic and clinical research.

"But much of the research agenda meets international rather than domestic priorities," Karim notes. For example, South Africa is researching HIV vaccines, but not the high



Text messages sent to mobile telephones are helping clinics in Nairobi follow up on patients in remote villages.



M. LAKSHMAN/AP PHOTO

Despite undeniable gains against the epidemic, in countries like India (pictured above) and Pakistan, the epidemic is spreading to heterosexual couples.

levels of infection among its young women or why a substantial increase in condom distribution — from 8 million in 1994 to 376 million in 2006 — has not reduced its rate of new infections in high-risk groups.

National governments are no better at funding their AIDS programmes. At an April 2001 summit in Abuja, Nigeria, 52 African countries pledged to allocate at least 15% of their national budgets for health. In 2007, only three (Botswana, Djibouti and Rwanda) were on track, and three others (Burkina Faso, Liberia and Malawi) had reached some targets.

These slow and bureaucratic governments are no match for the rapid shifts in the epidemic.

HIV/AIDS in Central and Southeast Asia has spread from injecting drug users to their sexual partners. In countries such as India and Pakistan, the epidemic once spread mainly through commercial sex work and drug use, but is increasingly affecting heterosexual couples.

Affordable drugs

Developing countries also need to sustain and expand treatment programmes that depend on cheap HIV drugs, the demand for which continues to grow.

Over past years, international charities have taken the lead in providing cheap medicines to the poor. For example, the Clinton Foundation has negotiated lower prices with 8 firms on 40 drug formulations and with 12 suppliers for testing kits. This has translated into cheaper drugs for two million people, nearly one-half of the infected population in developing countries.

Much credit also goes to India and Brazil, which thumbed their noses at drug companies and encouraged other developing nations to do the same.

Before 2005, Indian laws recognized patents only for the process used to make a drug, not

for the drug itself. Indian companies used this loophole to produce cheap generic versions of expensive antiretroviral drugs.

These days, 92% of those receiving treatment in low- and middle-income countries take generic drugs made in India — the ‘pharmacy of the developing world’.

In 2000, Brazil threatened to issue a compulsory license — a clause in international patent law that allows countries to waive patents during national health emergencies — arguing that its growing AIDS epidemic was a national emergency.

It carried out its threat in 2007, issuing a compulsory license to import efavirenz — which prevents HIV from replicating — from an Indian firm. In 2008, Brazil’s patent office also rejected a patent for tenofovir.

Following in Brazil’s defiant footsteps, about a dozen developing countries have issued compulsory licenses. “Brazil revolutionized global AIDS treatment and shaped global AIDS treatment policy,” says Amy Nunn, assistant professor of medicine at Brown University in Rhode Island.

Things could go wrong again, however. In 2005, India agreed to recognize international patents, meaning its companies can produce only generic drugs that are already on the market. In five or ten years, this is likely to create a serious shortfall in affordable drugs.

“It is a very complex issue,” says Mauro Schechter, professor of infectious diseases at the University of Rio de Janeiro. “We need more innovation but at the same time, people should have access to treatment,” Schechter says. “How do you do both at the same time? We have not found the answer.”

One potential solution is UNITAID, a not-for-profit patent pool set up in March 2010, in which drug companies forgo their

patent rights in selected countries, and allow local firms to make medicines with mutually agreeable licence fees.

Networks of hope

Developing countries are also coming up with innovative solutions to their own, and others’, problems in different arenas.

For example, in the late 1990s, the Boston-based non-profit group Partners in Health founded the HIV Equity Initiative to provide treatment and care to infected people in Haiti. This small charity has grown into a network of nine health centres that serve 1.2 million people under a national programme supported by Haiti’s health ministry.

In 2007, a doctor from Haiti set up a similar rural clinic with volunteers in the mountains of Lesotho, an African country with no medical school and about 80 doctors to attend to its two million people.

India, Brazil and South Africa are also collaborating on research projects, including HIV vaccines, combating HIV–TB and creating maps of viral diversity.

Brazil provides locally made HIV drugs to almost a dozen countries in Central and South America, and in Africa. In March this year, several Portuguese-speaking countries together set up a network on HIV and sexually transmitted diseases.

At an individual level, too, the urge to help and support each other is obvious. In Kajiado, for example, those who receive the weekly “Mambo?” text forward it to relatives who are not part of the Kenyan trial.

“As people become more connected, they become more hopeful,” says Lester. “Hope removes stigma.”

T. V. Padma is a freelance writer in New Delhi and South Asia editor of SciDev.net

Fighting the monster

Co-infection with HIV and tuberculosis is a potent combination. **Amy Maxmen** investigates the impact of this deadly duo.

Sayoki Mfinanga calls it “the monster” for the way it wreaks havoc on a body. Co-infection with HIV and tuberculosis (TB) can be devastating, triggering rapid weight loss, severe pneumonia and, often, a quick death.

Treating both diseases simultaneously does not improve matters, and in fact can cause fatal kidney and brain damage. “We don’t understand what happens when you combine TB and AIDS drugs,” says Mfinanga, director of the Muhimbili Medical Center at the National Institute for Medical Research in Dar es Salaam, Tanzania.

With weakened immunity, HIV-positive patients are highly susceptible to TB, and TB makes HIV disease progress quicker. According to a 2009 World Health Organization (WHO) report, there were more than 1.4 million cases of TB–HIV co-infection worldwide, resulting in about 0.5 million deaths in 2008. In sub-Saharan Africa, home to nearly 80% of those infected with both diseases, TB is the leading cause of death for HIV-infected individuals.



Many tuberculosis patients on the mend relapse after taking antiretroviral drugs, developing swollen neck glands and high fevers.

TB might be an important reason why HIV patients in the developing world suffer higher rates of mortality after beginning antiretroviral therapy, says Steve Lawn, associate professor of infectious diseases and HIV medicine at the Desmond Tutu HIV Centre in Cape Town. Although the probability of death after the first 12 months of antiretroviral therapy is about 1.8% in developed countries¹, Lawn and colleagues found that the rate is as high as 26% in sub-Saharan Africa².

In May 2010, the Center for Global Health Policy drew together leaders in US science policy and advocacy to discuss the alarming increase in TB fuelled by HIV in the developing world. “In 2008, TB killed more people than anytime in recorded history,” said Peter Cegielski, team leader for drug-resistant TB at the US Centers for Disease Control and Prevention. “In sub-Saharan Africa, that’s purely because of HIV.”

The deadly fallout of co-infection is forcing officials to take notice — and revise their priorities. “We consider co-morbidity of HIV and TB so prominent that it makes sense to use HIV money for TB research,” says Anthony Fauci, director of the US National Institute of Allergy and Infectious Diseases. “In fact,” Fauci says, “what I’m planning to do is to utilize some of our clinical trial networks that were only for HIV to study TB alone and TB with HIV.”

Exacting its toll

In the past few years, access to HIV medicines in the developing world has improved. As more individuals have begun taking the drugs, however, they have faced unexpected complications. Interactions between drugs for HIV and TB can cause nausea, allergic reactions and joint pain. Rifampin, which is the mainstay of TB treatments, weakens the efficacy of antiretroviral drugs such as protease inhibitors and the non-nucleoside inhibitors efavirenz and nevirapine. The *Bacillus*

Calmette-Guérin (BCG) TB vaccine can be lethal in HIV-infected infants.

Many TB patients who were previously improving with drugs relapse after starting a course of immune-boosting antiretroviral therapy, with a paradoxical reaction called immune reconstitution inflammatory syndrome (IRIS). “It may be that as antiretroviral drugs improve immunity, an inflammatory reaction is directed at the mycobacterium causing TB, and that reaction gives rise to the IRIS symptoms,” says Graeme Meintjes, senior clinical researcher at the University of Cape Town.

IRIS typically occurs two weeks after TB patients begin taking antiretroviral drugs. The lymph nodes in the neck swell, high fevers hit

and dry coughing begins. Some patients develop abscesses, stomach pain or kidney damage. About one person in every ten who develop IRIS has meningitis or inflammation of the brain, and many of these people die. Because definitions of IRIS vary and there are no diagnostic

tests, reports of its frequency range from 8% to 43% of TB patients taking AIDS drugs³.

IRIS was first described in the mid-1990s, but has become more common since. “There’s no question that the large-scale roll-out of antiretroviral therapy has helped millions of lives, but to be blind to the consequences of massive interventions, which are lifelong and delivered to a large proportion of your population, is not helpful,” says Robert Wilkinson, professor of infectious diseases at the University of Cape Town.

In some cases, the immune-suppressing steroid prednisone can mitigate IRIS. Because of the risk of side effects such as diabetes, high blood pressure and osteoporosis, however, prednisone should not be given for more than a few months. Worse, if prednisone is given to people who have drug-resistant TB or another infection that could be mistaken for IRIS, it might exacerbate the condition.

Eye of the monster

One way to prevent IRIS might be in the timing of antiretroviral therapy. Some clinicians try to avoid IRIS by delaying antiretroviral drugs until after the patient has completed TB treatment, but a February 2010 report of a controlled clinical trial in South Africa cast doubt on this strategy. The researchers reported that twice as many people died in the group that delayed taking antiretrovirals as in the group that treated both diseases at the same time⁴.

This has stirred some controversy: in June 2010, other researchers challenged the study

“We consider co-morbidity of HIV and TB so prominent that it makes sense to use HIV money for TB research.”



JORGEN SCHYTTE/STILL PICTURES

Co-infection with TB-HIV (as in the woman being examined, above) can be devastating, triggering rapid weight loss, severe pneumonia and, often, a quick death.

design⁵, and questioned the influence of drug-resistant TB⁶ and other factors that contribute to IRIS⁷. Three ongoing large trials are exploring the ideal time for those on TB treatment to begin antiretroviral therapy.

Other researchers are testing various combinations of antiretroviral and TB drugs. For instance, Wilkinson's team is assessing the safety of combining HIV treatments with the TB drug isoniazid. The WHO recommends isoniazid over rifampin for HIV-positive individuals in areas of high TB prevalence who do not show signs of TB, although it has not been rigorously tested in this setting.

Frustrated doctors say they hope that as each new HIV or TB drug enters the market, they will not need to scramble to learn how co-infected patients react.

"We quite often sit at meetings where pharmacies try to convince us to use their [HIV] drugs for South Africa, but unless the drugs are compatible with pregnancy and TB, unless they are compatible with rifampin, they aren't very useful for us," says Linda-Gail Bekker, deputy director of the Desmond Tutu HIV Center. "They need to be tested in pregnant women and TB patients — but that rarely happens, and often that's an afterthought."

Some drug developers, however, are finally realizing the importance of developing TB solutions that work for HIV-positive patients. "You can't use a TB vaccine worldwide on a massive scale and screen everyone in advance for HIV — that's impractical," says Gordon Douglas, executive chairman of the Aeras Global TB Vaccine Foundation, a non-profit organization that partners with companies to develop TB vaccines.

The clinical trial for one of Aeras' lead candidates, AERAS-402, includes HIV-infected

adults in South Africa. Animal data suggest that it will be safe for infants with HIV, Douglas says. Thus far, there have been no serious adverse reactions in adults. David McMurray, an immunologist at Texas A&M College of Medicine who is not affiliated with Aeras, says the vaccine "ought to be perfectly safe in HIV-positive individuals." Results are expected before 2013.

Clinical trials for HIV vaccines have not yet been designed with TB patients in mind, however. "A fundamental issue is that if someone has an episode of TB and develops immune dysfunction, will that affect the ability of that individual to mount a response when they're given an HIV vaccine?" asks Clive Gray, department head of HIV immunology at the National Institute for Communicable Diseases in Johannesburg. "The way to answer that is to do an HIV vaccine trial in people with immune memory to TB, but it's not really on the agenda now."

The spectre of drug resistance

Another looming threat is drug-resistant TB, particularly in crowded hospital wards where susceptible HIV-positive patients provide a reservoir for resistant strains. This became apparent in 2006 at a hospital in Tugela Ferry, South Africa, where 52 of 53 HIV-positive individuals infected with extensively drug-resistant TB died within a few months — many before they got the results from diagnostic tests.

Resistant strains of TB continue to sweep through HIV clinics. "We used to think that HIV was in Africa and [multidrug-resistant] TB was in Russia — well, not anymore," says Gail Cassell, vice president for scientific affairs at Eli Lilly in Indiana. "In 2008, we did big surveys in the Ukraine and in Latvia and

found that HIV and [multidrug-resistant] TB rates are going up, and the two things show a very clear association."

Drug-resistant TB takes longer to treat and requires treatments that are 50 to 200 times more expensive. According to a 2010 WHO report, the estimated cost of treatment between now and 2015 will reach US\$16.2 billion. Although new drugs to tackle resistant strains are being developed, it will take many years for them to be affordable for the majority. "In five years, there will not be new anti-TB drugs for use in developing countries because the compounds in the pipeline won't be ready until 2015 — if we're lucky," says Paul Nunn, coordinator of the WHO's Stop TB Partnership.

Perhaps answers will flow from recent investments by large organizations and institutes in rich countries. The Howard Hughes Medical Institute, for example, plans to invest US\$60 million in its new research centre on TB-HIV in South Africa, slated to open by late 2012.

Those with fewer resources do what they can. Mfinanga's group is trying to determine the ideal timing for antiretroviral therapy in people infected with TB, knowing that it is small incremental advances in knowledge — and not a knight in shining armour — that will help slay the monster.

Amy Maxmen is a freelance writer in New York City.

1. Braitstein, P. *et al.* *Lancet* **367**, 817–824 (2006).
2. Lawn, S. D., Harries, A. D. & Wood, R. *Curr. Opin. HIV AIDS* **5**, 18–26 (2010).
3. Meintjes, G. *et al.* *Lancet Infect. Dis.* **8**, 516–523 (2008).
4. Karim, S. S. A. *et al.* *N. Engl. J. Med.* **362**, 697–706 (2010).
5. Wilson, D. & Meintjes, G. *N. Engl. J. Med.* **362**, 2137 (2010).
6. Garcia-Vidal, C., Salvado, M. & Salavert, M. *N. Engl. J. Med.* **362**, 2137–2138 (2010).
7. Kadiravann, T. *N. Engl. J. Med.* **362**, 2138 (2010).



HIV worldwide

Grassroots initiatives

A Cambodian group has developed a pioneering community-based approach to HIV and TB care and research. **Amy Maxmen** describes how this powerful model is being expanded to other war-torn countries.

For 30 years, the paediatric ward of the Khmer Soviet Friendship Hospital stood crumbling at the southern end of the Cambodian capital Phnom Penh.

In 2008, supported largely by private donations, the Cambodian Health Committee, a non-governmental organization (NGO), transformed the ward into the Sullivan Out-patient Center for children with HIV and tuberculosis (TB). Later that year, Médecins Sans Frontières (MSF) entrusted the care of 205 HIV-positive children to the centre.

Now called the Global Health Committee (GHC), the Cambodian group was born in 1994 at a refugee camp on the Cambodia–Vietnam border. The genocidal Khmer Rouge regime targeted health-care professionals, so, when its reign ended in 1979, those who survived were left with no medical infrastructure.

Since its inception, the GHC has cared for more than 5,000 HIV patients and cured 25,000 people of TB, says Anne Goldfeld, the group's co-founder and professor of medicine at Harvard University. With donors ranging from the drug company Eli Lilly to the United Nations World Food Programme and the US National Institutes of Health, the GHC provides medicines, meals and care for an estimated 15,000 people. The group manages five centres for HIV and TB care and research in Cambodia, stretching from rural provinces to Phnom Penh.

Community involvement

In 2008, the group expanded its model to Ethiopia. Staff members have also visited refugee camps in Pakistan, and are discussing plans to move into Swaziland and Vietnam. Their approach differs from other missions that deliver care after war, says Paul Hamilton, project director at an aid organization in Uganda, who is not affiliated with the GHC. "I think their commitment is longer and they really work on incorporating the community. And the second thing that is really different is that they do a lot of basic research."

Many international medical NGOs are in the process of leaving Cambodia. At times like these, "the [GHC] becomes more important," says Didier Laureillard, a physician specializing in HIV and TB at the European Georges Pompidou Hospital in Paris, who volunteered with MSF in Cambodia.

The group owes its success in Cambodia in part to sustainability brought about by training and working with the community. Its staff members visit patients at home and enlist local health-care workers to monitor the nutritional quality of meals eaten by those receiving antiretroviral drugs from its clinics.

Goldfeld also initiated basic research in the mid-1990s. Her team is conducting a clinical trial on the ideal time for those co-infected with TB and HIV to begin taking AIDS drugs.

"They really work on incorporating the community."



The Cambodian group has cared for many children infected with HIV and tuberculosis, including Ye How (at age six in 2006, top; and after successful treatment a year later, bottom).

"There were a lot of people who doubted our ability to do it because Cambodia had no infrastructure," she says. "Not only did we end up recruiting 661 patients but, what's more, we've enhanced care all over Cambodia for TB and AIDS."

Fighting the brain drain

Goldfeld has mentored local clinicians in writing proposals, designing studies and publishing results. That training is bound to have a broader impact, says Kristian Olson, a clinician educator at Massachusetts General Hospital and health advisor for the GHC.

"If you create this cadre of inquisitive physician scientists that are driven to answer compelling questions, you'll do a lot for staff retention and reverse the brain drain," Olson says.

Olson volunteers with GHC's outpost in Addis Ababa, set up with the Ethiopian Ministry of Health's approval, to combat HIV and TB. According to a 2008 World Health Organization report on Ethiopia, 40% of TB patients tested

there were HIV-positive.

The Cambodian team has repeatedly visited Addis Ababa to train local doctors. Last year, with funding from the Jolie–Pitt Foundation, it opened the Zahara Children's Center for HIV-positive kids modelled after the group's similar Maddox Chivan Children's Center in Cambodia.

With some tweaks, the GHC's model for building medical infrastructure will work in any impoverished, war-torn country, says Sok Thim, a former Cambodian refugee and the group's co-founder. "We have the model and the means to deliver it," Thim says, "We want to help wherever there is great need for TB care."

Extending models to other countries is harder than it sounds, cautions Hamilton. "I've seen good models work in one country but not transfer to another." For example, he says, the Ethiopian government is wary of NGOs. Yet those affiliated with the GHC say it has succeeded in Ethiopia partly because it works in partnership with the Ministry of Health.

As a survivor of the killing fields, poverty and TB, Thim relates to the frustrations of Ethiopian health-care workers, Olson says.

"Care providers in Addis Ababa see Thim as a medical expert who easily could have been the poor refugee," says Olson. After training with Thim, Olson says, "I heard them going from no confidence to feeling like they will be [part of] a centre of excellence".

Amy Maxmen is a freelance writer in New York City.

A call for collaboration

Specialization has its place, but truly innovative advances in HIV research usually come from interdisciplinary efforts, reports **Unmesh Kher**.

It is probably a good thing Rowena Johnston got her doctoral degree in neuroscience.

When Johnston joined the Foundation for AIDS Research (amfAR) nearly nine years ago, most people working on HIV would have felt foolish even mentioning a cure. As an outsider, however, she could look at the proposition with fresh eyes.

"I wasn't loaded down with all the baggage that was preventing people from explicitly pursuing a cure," recalls Johnston, who became amfAR's director of research in 2004. Her unabashed ambition has now found expression: in March 2010, amfAR requested proposals for a consortium dedicated to exploring "the barriers to and potential for HIV eradication." In May 2010, it issued grants totalling US\$1.2 million to launch the initiative.

The rest of the field is eyeing the same goal. A paper in *Science* last year sketched out the organizational structure of a 'collaboratory' for a cure. The US National Institutes of Allergy and Infectious Diseases (NIAID) is preparing a request for applications explicitly aimed at exploring how HIV hides out and persists in the body, and for developing safe strategies to eliminate the latent virus.

These schemes have more in common than just their aims. They also share a view of how best to attain them: through an interdisciplinary team of scientists with diverse backgrounds.

amfAR has awarded the bulk of its funding to three collaborative teams that will each address distinct aspects of the problem. These groups are expected to talk frequently and, where helpful, participate in each other's projects. NIAID's programme would also seek to support a collaborative effort that includes teams from academia, the private sector and the US National Institutes of Health — much as was envisioned in the *Science* paper.

A glance at HIV/AIDS research suggests that interdisciplinary groups are increasingly the order of the day. Such groups have emerged

from a conviction shared by many in the field that their profession has become at once too big and too internally ghettoized.

"There are more [HIV] investigators than there are nucleotides in the HIV genome," says Douglas Richman, a professor of pathology and medicine at the University of California, San Diego, and lead author on the *Science* paper. "It shows you just how clever the virus is."

Isolated experts

Excessive specialization has also hampered the flow of information between researchers. For example, HIV immunologists rarely interact with immunologists in other fields, notes Bruce Walker, professor of medicine at Harvard Medical School, and director of the US\$100 million Ragon Institute, launched last year in Massachusetts with an explicitly interdisciplinary mission.

Because there is no mouse model for AIDS, Walker adds, those who develop HIV vaccines do not interact with immunologists who work with mice, and miss out on valuable insights. "I think there are basic universal truths that can be generated in mouse models and applied to humans," he says.

If history is any guide, bridging some of these gaps would be wise. Most major breakthroughs in HIV research have relied on the integrated contributions of multiple disciplines. The discovery of HIV depended on the combined efforts, on two continents, of clinicians treating inexplicably immune-compromised patients, epidemiologists tracking the emerging disease and virologists racing to isolate its causative agent.

Likewise, when the HIV genome was sequenced in 1984, the possibility of attacking it with drugs less toxic and more effective than the first chemotherapeutic agent AZT (zidovudine) drew a host of new scientists into the fray, not least structural biologists. The subsequent determination of HIV protease structures led to the development of inhibitors of the enzyme,



The discovery of HIV by France's Luc Montagnier (left) and American Robert Gallo (right) depended on the combined efforts of clinicians, epidemiologists and virologists.

which was a landmark in rational drug design and primer to a cascade of antiretroviral drug discovery.

"Perhaps the biggest breakthrough of all," says Anthony Fauci, director of NIAID, "has been the successful development of now more than 30 drugs that, in combination, have totally transformed the lives of HIV-positive individuals."

As the field evolved, the number of contributing disciplines multiplied. The development of sensitive assays to quantify HIV in the bloodstream — the typically unsung contribution of analytical biochemistry — touched off an avalanche of discovery in the 1990s.

"The mere development of these assays allowed the field to understand the correlation between viral load and disease progression," recalls David Ho, chief executive officer of the Aaron Diamond AIDS Research Center in New York City.

Ho's work captured the interplay of the virus and its target cells, and linked those dynamics to disease progression. Mathematics and computer modelling played an integral role in those studies, which Ho notes was unusual in virology at the time. His work led to the standard use of drug cocktails to treat AIDS.

Fresh blood

Today, HIV researchers generally agree that the two areas most in need of interdisciplinary efforts are the quest for an AIDS vaccine and the search for a cure. "Those of us who've been in the field haven't solved the problem," says Walker. "What we need is for people from other fields to work on the problem."

Innately interdisciplinary approaches such as systems biology could shed light on some of the toughest problems facing the field, such as how activating certain receptors on cells of the innate immune system — the generic first line of defence — influences the adaptive

"What we need is for people from other fields to work on the problem."



The US\$100 million Ragon Institute, launched last year, employs engineers, clinicians, basic immunologists, computational biologists and geneticists.

immune response to HIV infection, which is dominated by antibody and T-cell responses against specific pathogens or antigens.

"There has been a failure to appreciate what kinds of immune responses are needed to protect against HIV," notes Bali Pulendran, professor of pathology and laboratory medicine at Emory University. "It's a larger issue than just HIV research because the way people have analysed immune responses in the context of any disease, or any antigen, is through the use of standard immune assays."

In 2008, Pulendran applied a systems-biology approach to predict — with high accuracy — the precise immune responses elicited by the powerful yellow fever vaccine. Using a DNA microarray, Pulendran and colleagues measured changes in gene expression in response to the vaccine and then used that expression profile to estimate levels of the subsequent T-cell and B-cell responses.

Similar approaches relying on microarray analysis, proteomics, microRNA profiling and other techniques generate thousands of variables for computational analysis, and could be applied together to solve seemingly intractable problems, says Pulendran. "In the past four or five years or so," he says, "people have begun to understand that to fully capture the molecular and cellular mechanisms of immunity, you need a much more global approach."

Vaccine wars

Most agree that interdisciplinary problem-solving is a good thing. More controversial are the consortia for HIV vaccine research established by two of the field's major funders.

In 2005, the US National Institutes of Health (NIH) established a US\$315 million, seven-year venture named the Center for HIV/AIDS Vaccine Immunology (CHAVI). It was followed a year later by the Bill & Melinda Gates Foundation's US\$287 million constellation of consortia named the Coalition for AIDS Vaccine Discovery. Many complain that these initiatives have concentrated scarce funding on a small group of researchers.

"It has a way of homogenizing the field," says Ho, who first raised his concerns when CHAVI was just a concept. "These people are forced to meet — it's almost dictated by the grant funding structure. They talk about the same things over and over, and pretty soon I think everybody thinks the same way, when what we need in the field is more ideas, new ideas."

Instead, say Ho and others, collaborative groups should form naturally. Julie Overbaugh, a virologist at the Fred Hutchinson Cancer Research Center in Seattle, adds that the AIDS vaccine consortia might also be premature. "We're still early in the game, in terms of potential vaccine candidates that might ever make it to the real world," she says. "The 'groupthink' can be a bit of a problem at this point. Dominant personalities can win in those kinds of settings."

Fauci bristles at such criticism, pointing out that CHAVI's director, Barton Haynes of Duke University, has a reputation for open-mindedness, and routinely takes advice from many investigators in the field.

"CHAVI implements the ideas of a lot of investigators in a synergistic and collaborative way," Fauci says. He points out that the

bulk of funding for HIV vaccines still goes to individual grants. Of the US\$561 million the NIH set aside for HIV vaccine research in fiscal year 2009, CHAVI's cut was US\$49.8 million. Besides, Fauci says, many of the people funded through CHAVI also have independent investigator grants — underscoring the fact that its members are not group thinkers.

Some say a CHAVI-like consortium is just what the field needs to find a cure for HIV. In the *Science* paper, Richman and several other leading scientists argued that the NIH should provide "funding, regulatory oversight and coordination" to a multidisciplinary, multi-sector collaboration — one that depends as much on the contributions of big pharma as those of academia. Richman says he does not believe amfAR's initiative has the resources to fund a true consortium.

Johnston counters that the organization has successfully fostered other underfunded areas of HIV research, including studies on microbicides and the preclinical development of several antiretroviral drug classes. In many cases, she says, amfAR's initial investments were followed by support from other institutions, including the NIH.

The organization hopes to catalyse a collaborative movement towards a cure, says Johnston. "It's probably fair to say that most AIDS researchers are still sceptical that there can ever be a cure," she says. "But our view is that if you don't even try, you can guarantee that there won't ever be one."

Unmesh Kher is a freelance writer in New York City and a consultant for the International AIDS Vaccine Initiative.