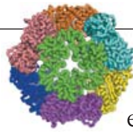
WORLD VIEW What the economic strife in Spain means for science **p.7**

BUILDING SITE The small proteins that combine for greater things **p.8**



LIP SERVICE Monkey mouth moves could explain human speech **p.9**

Back to Earth

The world has a surfeit of pledges, commitments and treaties. What it needs from the second Earth summit in Rio is firm leadership and a viable plan for success.

Twenty years ago, *Nature* proclaimed the first Earth summit in Rio de Janeiro to be a success (see *Nature* **357,** 523–524; 1992). The article offered a sober defence of a political process that had suffered intense criticism, but it also provided an impassioned assessment that pulled no punches. It challenged Rio's detractors as naive optimists operating in a bubble chamber of "utopian rhetoric". And it hit out at many of the political elite, including the summit chairman Maurice Strong, for feeding the delusion with dangerously grandiose proclamations, seemingly suggesting that one good summit on sustainable development could solve all of the world's problems. But *Nature* took heart in humanity's formal acknowledgement of the monumental challenges ahead, and applauded the creation of an evidently incremental process to address those issues.

As the world heads into the second United Nations Conference on Sustainable Development in Rio later this month, the collective failure to fulfil those initial pledges is all too evident. Countries have increased the rhetoric and their political commitments, but there is little to show for 20 years of work, apart from an impressive bureaucratic machine that has been set to indefinite idle. On urgent environmental issues, the world has perfected the art of incremental negotiation and redefined circular motion. Meanwhile, as documented elsewhere in this issue, pressure on the planet continues to build, greenhouse-gas emissions are still rising and species are still disappearing (see page 19).

In short, development continues apace, as it must and should in order to lift the world's poorest out of poverty, but it is hardly sustainable. The goal of stabilizing greenhouse-gas emissions seems just as daunting today as it did two decades ago, and people continue to devour the world's remaining wild habitat at an alarming pace.

TIME TO REASSESS

So what is the purpose of the Rio+20 meeting? It cannot be a celebration. Nor should it be a platform for major new treaties and commitments — the world is awash with both, and to no avail. Instead, the second Earth summit is a chance to take honest stock of the situation and present ways to break political deadlock and hasten progress on the ground, in the air and in the oceans.

Diplomats, politicians, scientists and environmentalists alike must acknowledge where environmental politics has failed — although it is just as important to recognize where there has been progress. Greenhouse-gas emissions may be climbing at a breakneck speed, but governments around the globe have at last begun to take global warming seriously and to prepare their citizens for a changing climate. And although the global picture for biodiversity loss is gloomy, Brazil itself shows what can be achieved, despite an ongoing row about forest policy (see page 13). Deforestation in the Brazilian Amazon is down by a whopping 78% from its recent high in 2004. If Brazil can maintain that progress — and Norway has put a US$1-billion reward on the table as encouragement — it would be the biggest environmental success story in decades, and would set an example to other countries that want to protect their tropical forests.

Global effort remains important to address global problems but, as deforestation in Brazil shows, much of the current progress on the environment occurs at the national and, often, sub-national level. Fed up with the slow pace of international negotiations, state and national governments are moving forwards on their own, experimenting with ideas and policies that may one day spread around the world. This activity is helping to bridge the increasingly arbitrary gap between industrialized and rapidly emerging economies. If countries make any progress on their pledge to sign a new global-warming treaty by 2015, it will be thanks in no small part to the fact that many involved are developing their own independent climate policies.

"The world has perfected the art of incremental negotiation and redefined circular motion."

The international process remains important because it has focused resources and intellectual energy on global warming, biodiversity and sustainable development, and scientists can rightly claim a place at the heart of this political process. Climate diplomats have tied all of their major decisions over the past two decades to reports produced by the Intergovernmental Panel on Climate Change. Governments have invested in Earth and environmental sciences. And scientists have improved their understanding of climate and ecological processes while providing policy-makers with new tools and indicators to assess the threats ahead. Many scientists are frustrated by the lack of political progress, and rightly so, but there is little else to do but keep calm and carry on.

Nevertheless, it is hard to avoid a certain sense of gloom, if not doom. Despite progress on some issues — ozone loss, for example — the disconnect between science and politics seems to be growing, not shrinking. The accumulating evidence screams that the consequences of inaction could be dire. As each day passes, the problems become more expensive to solve and the number of available options decreases. New clean-energy technologies could make all the difference to climate, but many governments in the industrialized world are investing less money in clean energy now than they were just a few years ago.

In 1992, *Nature* warned against thinking that a single summit could eradicate poverty and redistribute wealth while setting specific limits on greenhouse gases. The expectations for Rio+20 are so low that almost any agreement or affirmation would qualify as a success. The fact is that politicians know what needs to be done, and countries committed to doing it 20 years ago; what is missing is political leadership and solutions that are cheap, scalable and politically viable. For the second time, the world has a chance to craft a workable agenda, but the elusive key to success lies in finding a way to overturn the widespread reluctance to make the necessary investments in time, money and intellect to get the job done. ■

EPE/PHOTOSHOT

Turn Spain's budget crisis into an opportunity

Strict funding cuts mean that the country's research system must renew its focus on quality rather than quantity, says science secretary **Carmen Vela**.

Spain's 2012 budget is the most austere in our democratic history. The government has been forced to optimize its limited resources in all areas — and science, technology and innovation cannot be exempted. That is why I have agreed to a significant, although not insurmountable, decrease in resources.

In the part of the budget allocated mainly to the grants and subsidies that are indispensable to research, there has been a €475-million (US$591-million) reduction: a decrease of 22.5%. This comes on top of cuts in previous years, so it cannot be denied that we face a very challenging situation.

We know what needs to be done. My department must prioritize and strive for excellence. One-quarter of the Spanish labour force is unemployed, so although investment in science, technology and innovation is a priority, it must also be realistic.

Look around and don't be fooled. We must now stop talking about the importance of science, and instead commit ourselves to the need for excellence in science.

Research, development and innovation in Spain have unquestionably advanced over the past decade. But this accelerated growth can hamper the effective management of resources, and a number of overlapping institutions and functions have been created. Currently, there is a biotechnology research centre or a science park in almost every Spanish region.

To strengthen the research system in our country we must slim it down, but it is important to cut back on quantity, not quality. This process will be complex and unpopular: after all, nobody likes cuts or readjustments. Under the changes that I announce here, only those scientists who can demonstrate that they are pushing the frontiers of our knowledge will be allotted resources. We want to support only the really competitive projects that are bearing fruit, or that show the potential to do so through recent results, and which aim to improve the daily lives of our citizens. Competitiveness will become part of the process of obtaining state funding.

WHEN IT COMES TO SCIENCE, OUR **NUMBER-ONE PRIORITY** REMAINS SUPPORT FOR SPAIN'S **EXCELLENT RESEARCHERS.**

We must develop and optimize the Science, Technology and Innovation restructuring law that came into force last year: to do this, we will create a government agency to evaluate and fund research and development (R&D). This agency will make the management of public funding more efficient; establish enforceable commitments in management contracts; and give more autonomy and responsibility to the science and technology community.

We also need the private sector to have a role in R&D, and we are looking into possible options, including optimization of the tax framework for R&D and crowd-funding schemes.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/wp3h4j

I encourage our researchers to demonstrate their excellence by competing internationally with the best in Europe. The European Union foresees an investment of more than €80 billion through the Horizon 2020 Framework Programme for 2014–20. Our scientists must seek and win some of this money.

We will encourage this through the recruitment of specialists to propose and manage European projects. We will also attach increased weight to previous European and international experience when making domestic funding decisions.

When it comes to science, our number-one priority remains support for the excellent researchers that Spain already has. In February, we were able to continue the Researcher Staff Training funding scheme for our youngest talent with the same total funds as last year.

We need to change the number of researchers by maintaining and improving the quality of the contracts while reducing the quantity. We would have needed to do this anyway: the Spanish R&D system is not large enough to justify paying as many researchers as it currently does.

We will reduce the number of grants offered each year in the Ramón y Cajal tenure-track programme for young researchers, from 250 in 2010 to around 175 in 2012. However, the quality of each grant will improve, with researchers gaining more independence through higher initial grants and a better distribution of the salary. The programme's total budget will increase from €45 million in 2010 to €54 million in 2012.

There will be similar changes to other significant programmes: the Juan de la Cierva postdoctoral grants, the Torres Quevedo industrial-research grants and employment of technical support staff members in Spanish universities and public-research organizations. We will offer 700–800 grants in total for these programmes this year, in comparison with 960 in 2011.

The situation is not ideal, but continued criticism will not help us to dig our way out. Excellence involves having an attitude based on effort and work, not just on criticism. It is not enough to focus on the present without planning for the future. My job, and that of my team, is to achieve excellence in investment using the available resources.

Albert Einstein, one of the few scientists whom people in Spain were able to name in a survey last month, once said that there is a driving force more powerful than steam, electricity and atomic energy: the will. With will, our slimmed-down R&D system will be able to take advantage of the crisis — and emerge from it stronger than ever. ■

Carmen Vela *is Spain's secretary of state for research, development and innovation.*
e-mail: secretaria.seidi@mineco.es

RESEARCH HIGHLIGHTS

Selections from the scientific literature

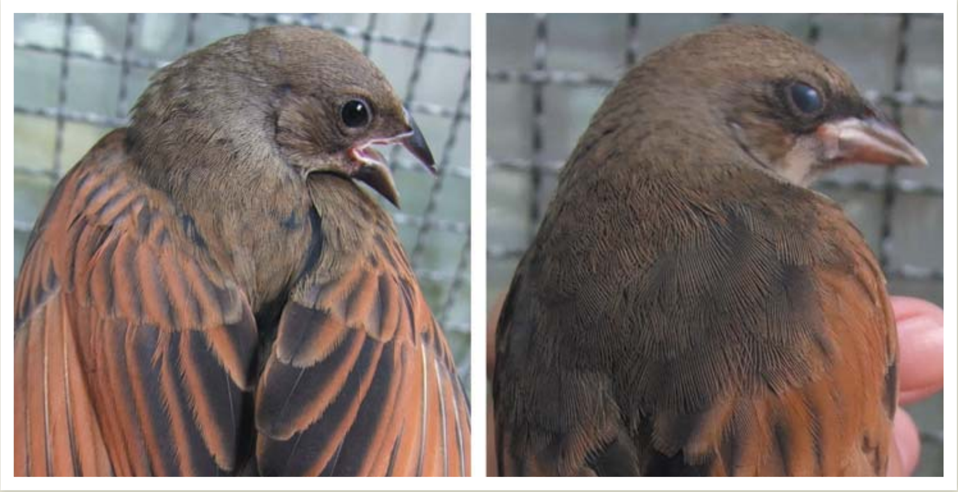
NEUROSCIENCE

Cell transplants for pain

Transplanting embryonic neurons into the mouse spinal cord seems to alleviate neuropathic pain — chronic pain that arises spontaneously or in response to the slightest touch due to peripheral nerve injury.

In this type of injury, signalling by spinal cord neurons that produce a neurotransmitter called GABA is reduced. João Bráz and his colleagues at the University of California, San Francisco, transplanted GABA-producing interneurons from the mouse embryonic forebrain into the spinal cords of adult mice with peripheral nerve injury. The transplanted cells integrated into the hosts' spinal cord circuitry and eliminated touch-induced hypersensitivity in mice with neuropathic pain, but had no effect in animals with inflammatory pain.

Neuron **74, 663–675 (2012)**



C. URSINO

EVOLUTION

Fledglings occupy nests, fool hosts

Birds such as the cowbird that lay their eggs in the nests of other birds have evolved strategies to disguise their chicks and not just their eggs.

María De Mársico and her colleagues at the University of Buenos Aires observed eggs or hatchlings from the screaming cowbird (*Molothrus rufoaxillaris;* pictured left) or the shiny cowbird (*Molothrus bonariensis*) that were placed or laid in the nests of the baywing (*Agelaioides badius;* right), and measured the fledglings' survival rates. The researchers found that the baywing hosts rejected 83% of the shiny cowbird fledglings, but none of the screaming cowbirds. Analysis of the birds' plumage revealed that the differences in colour between screaming cowbird and baywing fledglings are likely to be indistinguishable to the avian eye. The begging calls of these species are also very similar, whereas the shiny cowbirds differ in their calls and appearance.

Proc. R. Soc. B **http://dx.doi.org/10.1098/rspb.2012.0612 (2012)**

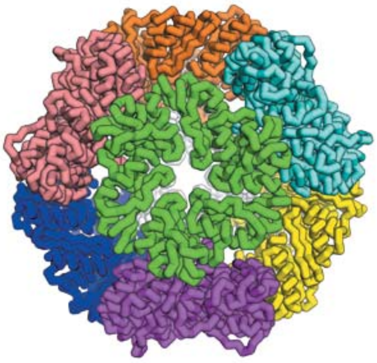
BIOCHEMISTRY

Proteins designed to self-assemble

Large protein complexes can be designed and created using smaller protein building blocks that self-assemble.

David Baker at the University of Washington, Seattle, and his team report a method for producing such proteins. The authors simulate the docking of protein building blocks in desired architectures and then design amino-acid sequences for these blocks that result in low-energy interfaces between the blocks, driving self-assembly. The researchers incorporate the genes that encode the designed blocks into the bacterium *Escherichia coli*, which produces the proteins that then spontaneously self-assemble into the target architectures. The team created two cage-like proteins: one consisting of 24 building blocks in a 13-nanometre-wide complex with octahedral symmetry (**pictured**) and another comprising 12 subunits with an 11-nanometre-wide tetrahedral symmetry.

Science **336, 1171–1174 (2012)**



AAAS

ANTHROPOLOGY

Rich milk for poor girls

Poor mothers in northern Kenya produce fattier milk for their daughters than for their sons, whereas those who are better off financially favour their sons over their daughters. The findings support a 1973 hypothesis that predicts that poor mothers will invest more resources in daughters, who stand a greater chance of increasing their status through marriage than do poor males. Conversely, mothers from wealthier families give more to their sons, who can pair with multiple females.

Masako Fujita at Michigan State University in East Lansing and her team assessed the fat content of milk from 83 mothers living in villages in which men can have multiple wives. The authors found that, when they controlled for factors such as age and dietary fat intake, mothers with less land and fewer livestock provided richer milk to their daughters than to their sons.

Am. J. Phys. Anthropol. **http://dx.doi.org/10.1002/ajpa.22092 (2012)**

NEUROSCIENCE

Antidepressants' cellular target

Certain antidepressant drugs seem to work by acting on a tiny population of brain cells in the cortex. Identifying the specific cells targeted by these drugs, which are known as serotonin-specific reuptake inhibitors (SSRIs), could aid the development of more selective medicines.

Nathaniel Heintz and his team at the Rockefeller University in New York homed in on a group of cortical neurons that express a protein called p11 — levels of which are decreased in depression. This protein regulates the signalling of the transmitter serotonin in the brain.

Mice that were treated long-term with the SSRI fluoxetine showed alterations in the expression of many genes in p11-producing cortical neurons, but not in neurons that lacked p11. When the researchers deleted the gene that encodes p11 from cortical neurons, the mice no longer responded to the SSRI in behavioural tests that model aspects of depression.

Cell **149, 1152–1163 (2012)**

CANCER

Forcing cells to divide

Aggressive breast cancers dubbed 'triple negative' could one day be treated by inhibiting a protein that helps to control the initiation of mitosis — the segregation of copies of DNA strands that precedes cell division.

The growth of these tumours depends on the protein WEE1, which prevents cells from entering mitosis too early. Nicholas Turner at the Institute of Cancer Research in London and his team treated cultured breast cancer cells with an experimental drug that inhibits WEE1, and an approved therapy, gemcitabine, that prevents new DNA from being made. The combined treatment sent cells into mitosis before they had finished copying their DNA, eventually triggering programmed cell death.

The treatment also induced early mitosis in human colon cancer cells that had been implanted into mice, and caused the tumours to grow more slowly than did treatment with either drug alone.

Cancer Discov. **http://dx.doi.org/10.1158/2159-8290.CD-11-0320 (2012)**

GLACIOLOGY

Greenland glacier map

Satellite observations made during the International Polar Year 2008–09 have yielded a near-complete map of ice motion in Greenland.

Eric Rignot and Jeremie Mouginot of the University of California, Irvine, combined radar data from three satellites to map the velocities of the island's largest glaciers at high resolution. The speed at which Greenland's glaciers are moving towards the coast ranges from just a few centimetres per year to 13 kilometres per year for the fastest-moving ice stream, Jakobshavn Isbræ. Glaciers in areas with high annual precipitation are generally moving faster than those in Greenland's dry and cold north.

The map provides a new constraint for ice-sheet models, the authors say.

Geophys. Res. Lett. **http://dx.doi.org/10.1029/2012GL051634 (2012)**

VASCULAR BIOLOGY

Stem cells from blood vessels

The development of certain vascular diseases involves the division and migration of blood vessel cells. These seem to arise from stem cells in the vessel wall, not from smooth muscle cells as previously thought.

Song Li at the University of California, Berkeley, and his team isolated the stem cells from rat, mouse and human arteries and showed that the cells can specialize into other cell types including smooth muscle and fat cells. To identify the origin of the dividing vascular cells, the authors fluorescently tagged smooth muscle cells in mice; they found, however, that the dividing vascular cells did not fluoresce but instead expressed stem-cell markers. Vascular cells isolated from an injured mouse carotid artery consisted mainly of these stem cells, suggesting that it is stem cells that divide upon injury, not smooth muscle cells.

Nature Commun. **http://dx.doi.org/10.1038/ncomms1867 (2012)**

COMMUNITY CHOICE

The most viewed papers in science

IMMUNOLOGY

With stress comes inflammation

✪ HIGHLY READ
on www.pnas.org
in April

Long-term stress boosts inflammation, making people more susceptible to colds.

Sheldon Cohen at Carnegie Mellon University in Pittsburgh, Pennsylvania, and his team quarantined 276 volunteers and then exposed them to a rhinovirus that causes the common cold. People who had recently been experiencing a threatening stressful event were more likely to develop a cold than those who had not experienced stress. In the stressed individuals, white blood cell numbers did not correlate with levels of the hormone cortisol — indicating that the cells were insensitive to cortisol's anti-inflammatory effect.

In another study of 79 volunteers exposed to a rhinovirus, those with cortisol-insensitive white blood cells had higher nasal levels of immune-signalling molecules that promote inflammation. The results suggest that stress induces cellular resistance to cortisol, reducing the body's ability to regulate inflammation — which is at the root of many diseases, including heart disease.

Proc. Natl Acad. Sci. USA 109, **5995–5999 (2012)**

EVOLUTION

Monkey lips smack of speech

Human speech could have evolved from monkey lip-smacking, an affectionate gesture that monkeys make towards each other.

Asif Ghazanfar at Princeton University in New Jersey, Tecumseh Fitch at the University of Vienna and their colleagues made X-ray movies of macaques during episodes of lip-smacking. The monkeys moved their lips five times per second — the same frequency as occurs in human speech, and much faster than when the monkeys chewed. Moreover, the monkeys' lip movements were independent of their throat movements during lip-smacking, much like human speech.

Curr. Biol. **http://dx.doi.org/10.1016/j.cub.2012.04.055 (2012)**

C. RUOSO/MINDEN PICTURES/FLPA

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

NASA/SDO

Venus crosses the Sun

Venus made its transit across the Sun on 5–6 June, watched eagerly by professional and amateur astronomers, as well as by millions of people following it on live webcasts. The rare event occurs twice in close succession roughly every 120 years; the last transit was in 2004 and the next one will not be until 2117. Scientists hoped to get a once-in-a-lifetime snapshot of the climate of the entire planet (rather than the time- and space-limited atmospheric snapshots taken by the Venus Express probe), and to use the transit to check the way we measure exoplanets circling distant stars. See go.nature.com/yhpiqu for more.

POLICY

European tension

European science ministers have agreed a general structure for the region's enormous 2014–20 research funding programme. But at negotiations in Brussels on 31 May, the European Parliament and member states in eastern Europe all insisted that they want to keep the 'closed club' of western European research communities from winning the bulk of the funding. Mechanisms might include twinning wealthy western universities with eastern partners. Final decisions on the programme — the budget of which is yet to be confirmed, and wavers around €80 billion to €88 billion (US$99 billion to $109 billion) — will not be completed until 2013. See go.nature.com/pjueyp for more.

Integrity rebuke

At least 14 US federal agencies have scientific-integrity policies that do not comply with requirements laid down by the administration of President Barack Obama, according to the Project on Government Oversight (POGO), a non-profit watchdog based in Washington DC. In a letter to the Office of Science and Technology Policy on 31 May, POGO said the problem is that some agencies don't extend their policies to contractors or grant recipients. The watchdog singles out the Department of Energy for particular concern, because much of its research comes from non-governmental parties.

Golden age of gas

Natural gas (methane) could overtake coal as the world's second-largest energy source (after oil) by 2035 — but only if the industry respects environmental and social concerns about how to extract it, according to an International Energy Agency report released on 29 May. The report, *Golden Rules for a Golden Age of Gas*, recommends full transparency and engagement with local communities when, for example, extracting recently discovered reserves of shale gas. It adds that although a switch to gas would lower carbon dioxide emissions, this would not be enough to limit the world's long-term predicted temperature increase to 2 °C above pre-industrial levels.

NOAA funding row

The US National Oceanic and Atmospheric Administration (NOAA) has clashed with Congress over attempts to finance its National Weather Service (NWS), after an investigation revealed on 24 May that the service had misappropriated funds. It had redirected millions of dollars to regional weather offices without Congress's permission. On 25 May, John Hayes, director of the NWS and assistant administrator of NOAA, resigned — although officials denied any link with the funding revelations. NOAA is now asking Congress to allow it to redirect US$36 million to the NWS from other departments in fiscal year 2012 — but some senators say they will not approve this without more details on why funds were originally misappropriated.

FUNDING

Texas cancer review

The Cancer Prevention and Research Institute of Texas (CPRIT) in Austin is to re-review a controversial US$18-million technology grant that it speedily awarded without scientific review in March to the University of Texas MD Anderson Cancer Center in Houston. The controversy at the state-funded institute began when its top scientist, Alfred Gilman, announced on 8 May that he would resign, citing concerns about the grant as among his reasons. But the re-review will be carried out by a commercialization review council, not a scientific body, CPRIT said on 30 May. See go.nature.com/7v59sq for more.

PEOPLE

Winning millions

The three Shaw prizes for 2012, worth US$1 million apiece, were announced on 29 May. David Jewitt of the University of California, Los Angeles, and Jane Luu, at the Massachusetts Institute of Technology's Lincoln Laboratory in Lexington, shared the astronomy prize for their work on trans-Neptunian objects. (The pair and a collaborator were also awarded the Kavli prize; see below.) The life-sciences and medicine prize went to Franz-Ulrich Hartl of the Max Planck Institute for Biochemistry in Munich, Germany, and Arthur Horwich of Yale University School of Medicine in New

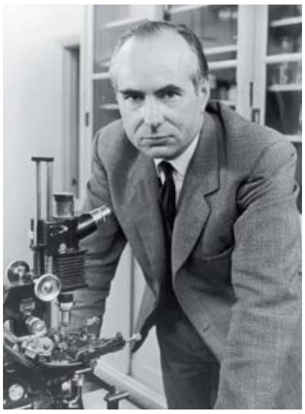
Haven, Connecticut, for research on protein folding. Maxim Kontsevich, at France's Institute of Advanced Scientific Studies outside Paris, won the mathematics prize for his work on algebra and geometry with applications in quantum physics. See go.nature.com/62ttvb for more.

Kavli prizes

One day after winning the Shaw prize (above), David Jewitt and Jane Luu learned that they had won a biennial Kavli astrophysics prize, together with Michael Brown of the California Institute of Technology in Pasadena. The Kavli nanoscience prize was awarded to Mildred Dresselhaus of the Massachusetts Institute of Technology (MIT), and the neuroscience prize was shared between Cornelia Bargmann of the Rockefeller University in New York, Winfried Denk of the Max Planck Institute for Medical Research in Heidelberg, Germany, and Ann Graybiel of MIT. Each prize is worth US$1 million. See go.nature.com/m53tkz for more.

Nobel laureate dies

Physiologist and biophysicist Andrew Huxley, who shared the 1963 Nobel Prize in Physiology or Medicine, died on 30 May, aged 94.



KEYSTONE/GETTY

He won his Nobel for work done with Alan Hodgkin in the late 1930s and after the Second World War at Britain's Plymouth Marine Laboratory. The two identified — from experiments on the axon of the giant squid — how electrical impulses travel along nerve cells. Knighted for his science in 1974, Huxley (**pictured**) was also president of the Royal Society between 1980 and 1985, and master of Trinity College, Cambridge, UK, from 1984 to 1990.

RESEARCH

Spyware analysis

A massive computer virus, dubbed Flame or sKyWIper, that seems to be targeting computers in the Middle East, and Iran in particular, may have been active for the past five years, according to a report by a research team involved in analysing the sophisticated and highly complex computer code. The conclusion, in a document released on 28 May by the Laboratory of Cryptography and System Security at the Budapest University of Technology and Economics in Hungary, is based on file names first spotted in Europe in 2007. Unlike the Stuxnext malware discovered in 2010 and credited with damaging Iran's nuclear centrifuges, the Flame virus seems to be a form of spyware. See go.nature.com/ucxbev for more.

BUSINESS

Genetics patents

The consumer genetic-testing company 23andMe announced its first gene-related patent on 28 May, and said that it intended to file more, causing concern among some customers over whether the patent would impede access to genetic data. The company, in Mountain View, California, uses data gathered from consenting customers to find genetic variants that are associated with disease and other traits. Its first patent is on a test for a version of a gene that may confer susceptibility to Parkinson's disease. In response to queries about how it would enforce patents, the company added that it did not think they should be used to obstruct research. See go.nature.com/6ajqox for more.

COMING UP

10–14 JUNE
The American Astronomical Society meeting in Anchorage, Alaska, includes discussions of concepts for telescopes on the Moon.
go.nature.com/bzonto

13 JUNE
NASA's NuSTAR telescope, which will examine high-energy X-rays produced at the thresholds of black holes (see *Nature* **483,** 255; 2012), has its earliest scheduled launch date.
go.nature.com/1tribi

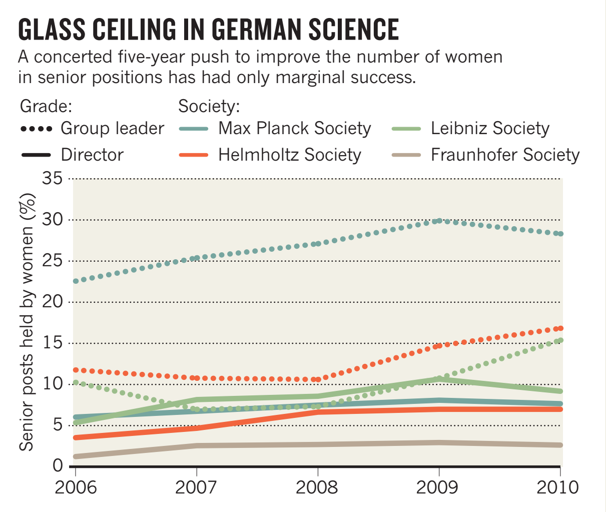
14–15 JUNE
Reports on the future of the US biomedical workforce — including recommendations to improve diversity — are delivered to a National Institutes of Health advisory committee.
go.nature.com/rfowsx

Dragon returns

The first flight to the International Space Station (ISS) by a private, commercially owned vehicle ended in success on 31 May, when the Dragon capsule — launched by SpaceX of Hawthorne, California — splashed down in the Pacific Ocean. SpaceX can now begin in earnest on 12 ISS resupply missions, under a US$1.6-billion commercial cargo contract with NASA. At more than $100 million per launch, that is not a cheap ride — but far cheaper than the space shuttle, which cost around $1.5 billion per launch. See go.nature.com/fxlcbv for more.

TREND WATCH

In late 2006, leading German science organizations and politicians signed up to a five-year effort to improve the representation of women in senior research positions. But an analysis published by the German Science Council on 29 May found only a modest improvement in equality. The chasm between the career success of male and female scientists is "less drastically remarkable" than 5 years ago, the report says. It proposes that organizations revisit pledges and set target quotas for top positions.



SOURCE: GERMAN SCIENCE COUNCIL

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

NEWS IN FOCUS

N. DUPLAIX/NATIONAL GEOGRAPHIC/GETTY

Revisions to Brazil's forest code could open the door to more deforestation along the nation's rivers.

ENVIRONMENT

President prunes forest reforms

Rousseff rejects elements of Brazil's revised forest code.

BY JEFF TOLLEFSON

Brazil's vast forests lost some legal protections last week, but less than environmentalists had feared. On 28 May, President Dilma Rousseff vetoed a dozen sections of the revamped forest code passed a month earlier by the lower house of Brazil's National Congress (see *Nature* http://dx.doi.org/10.1038/nature.2011.9584; 2011).

Although Rousseff denied environmentalists' push for a full veto, she removed many of the bill's contentious provisions, including one that would have effectively granted an amnesty for any illegal deforestation conducted before July 2008. She also issued an executive order to fill in the gaps created by her veto. Rousseff and her ministers defended their decision as a realistic compromise that promotes agriculture but also protects the environment. Many expect further legislative wrangling as Congress reviews the new language in the coming months.

The revised code still requires that landowners maintain a proportion of their land as forest, ranging from 20% for those in coastal regions to 80% in the Amazon. Rousseff restored obligations for landowners to restore forests that were cut down illegally, although she created exemptions that could relieve numerous small properties of this obligation.

Rules about riparian areas are another sore point. Whereas the old forest code required landowners to maintain corridors of riverbank forest 30–500 metres wide, depending on the size of the waterway, the revised law and presidential order reduce those requirements to just 5–100 metres. They also eliminate protections for steep slopes and allow landowners to meet some of their obligations to restore forest with permanent plantations of exotic trees, such as eucalyptus and oil palm.

RETURN TO RIO
For Earth Summit news:
www.nature.com/rio20

Brazil's Congress has until September to overturn Rousseff's vetoes with a simple majority of both houses, and the president's executive order will expire in late July unless approved by Congress. The two sides are bracing for another battle as the reviews move forward.

The politicized dispute is precisely what Brazil's leading scientific institutions were trying to avoid when they pushed for a temporary truce between conservation groups and rural landowners last year. Government enforcement has helped to reduce the rate of illegal deforestation dramatically since 2004, but it has also sparked a push-back from politicians in Congress representing agricultural interests. The Brazilian Academy of Sciences and the Brazilian Society for the Advancement of Science wanted time to identify a compromise that might garner support from both sides, but that idea foundered as positions hardened.

"We proposed a moratorium of two years to study the issue," says Luiz Martinelli, an ecologist at the University of São Paulo in Brazil who helped to organize an October 2011 position paper on forest-code reform. "Now we are in the middle of this mess, and I don't think it is going to end soon."

Environmentalists have generally criticized Rousseff's decision, but Dan Nepstad, a California-based ecologist with the Brazilian Amazon Environmental Research Institute, says that the revised code does not necessarily spell doom for the Amazon. Various agriculture and forestry initiatives at the state, national and international levels could help to maintain progress in reducing deforestation. The question, Nepstad says, is whether the new code will clarify rules for landowners and enforcement agencies or whether it will sow doubts and ambiguities that may undercut compliance.

For many, there is a sense that politics has got in the way of real reform. "We should have had legislation that gives economic incentives to farmers to recover their forests and manage their land in a sustainable way," says Adriana Ramos, executive secretary of the non-profit Socioenvironmental Institute in Brasilia. "We lost an opportunity." ■

SCIENCE AND SOCIETY

South Korea surrenders to creationist demands

Publishers set to remove examples of evolution from high-school textbooks.

BY SOO BIN PARK IN SEOUL

Mention creationism, and many scientists think of the United States, where efforts to limit the teaching of evolution have made headway in a couple of states[1]. But the successes are modest compared with those in South Korea, where the anti-evolution sentiment seems to be winning its battle with mainstream science.

A petition to remove references to evolution from high-school textbooks claimed victory last month after the Ministry of Education, Science and Technology (MEST) revealed that many of the publishers would produce revised editions that exclude examples of the evolution of the horse or of avian ancestor *Archaeopteryx*. The move has alarmed biologists, who say that they were not consulted. "The ministry just sent the petition out to the publishing companies and let them judge," says Dayk Jang, an evolutionary scientist at Seoul National University.

The campaign was led by the Society for Textbook Revise (STR), which aims to delete the "error" of evolution from textbooks to "correct" students' views of the world, according to the society's website. The society says that its members include professors of biology and high-school science teachers.

The STR is also campaigning to remove content about "the evolution of humans" and "the adaptation of finch beaks based on habitat and mode of sustenance", a reference to one of the most famous observations in Charles Darwin's *On the Origin of Species*. To back its campaign, the group highlights recent discoveries that *Archaeopteryx* is one of many feathered dinosaurs, and not necessarily an ancestor of all birds[2]. Exploiting such debates over the lineage of species "is a typical strategy of creation scientists to attack the teaching of evolution itself", says Joonghwan Jeon, an evolutionary psychologist at Kyung Hee University in Yongin.

The STR is an independent offshoot of the Korea Association for Creation Research (KACR), according to KACR spokesman Jungyeol Han. Thanks in part to the KACR's efforts, creation science — which seeks to provide evidence in support of the creation myth described in the Book of Genesis — has had a growing influence in South Korea, although the STR itself has distanced itself from such doctrines. In early 2008, the KACR scored a hit with a successful exhibition at Seoul Land, one of the country's leading amusement parks. According to the group, the exhibition attracted more than 116,000 visitors in three months, and the park is now in talks to create a year-long exhibition.

"The ministry just sent the petition out to the publishing companies and let them judge."

Even the nation's leading science institute — the Korea Advanced Institute of Science and Technology — has a creation science display on campus. "The exhibition was set up by scientists who believed in creation science back in 1993," says Gab-duk Jang, a pastor of the campus church. The institute also has a thriving Research Association for Creation Science, run by professors and students, he adds.

ANTIPATHY TO EVOLUTION

In a 2009 survey conducted for the South Korean documentary *The Era of God and Darwin*, almost one-third of the respondents didn't believe in evolution. Of those, 41% said that there was insufficient scientific evidence to support it; 39% said that it contradicted their religious beliefs; and 17% did not understand the theory. The numbers approach those in the United States, where a survey by the research firm Gallup has shown that around 40% of Americans do not believe that humans evolved from less advanced forms of life.

The roots of the South Korean antipathy to evolution are unclear, although Jeon suggests that they are partly "due to strong Christianity in the country". About half of South Korea's citizens practice a religion, mostly split between Christianity and Buddhism.

However, a survey of trainee teachers in the country concluded that religious belief was not a strong determinant of their acceptance of evolution[3]. It also found that 40% of biology teachers agreed with the statement that "much of the scientific community doubts if evolution occurs"; and half disagreed that "modern humans are the product of evolutionary processes".

Until now, says Dayk Jang, the scientific community has done little to combat the anti-evolution sentiment. "The biggest problem is that there are only 5–10 evolutionary scientists in the country who teach the theory of evolution in undergraduate and graduate schools," he says. Having seen the fierce debates over evolution in the United States, he adds, some scientists also worry that engaging with creationists might give creationist views more credibility among the public.

Silence is not the answer, says Dayk Jang. He is now organizing a group of experts, including evolutionary scientists and theologians who believe in evolution, to counter the SRT's campaign by working to improve the teaching of evolution in the classroom, and in broader public life. ■

1. Thompson, H. *Nature* http://dx.doi.org/10.1038/nature.2012.10423 (2012).
2. Xu, X., You, H., Du, K. & Han, F. *Nature* **475,** 465–470 (2011).
3. Kim, S. Y. & Nehm, R. H. *Int. J. Sci. Edu.* **33,** 197–227 (2011).

MORE ONLINE

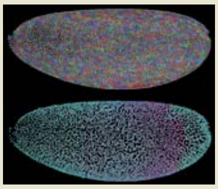
TOP STORY

Andromeda on collision course with the Milky Way go.nature.com/djnfcz

MORE NEWS

- Slow progress on Indonesian deforestation ban go.nature.com/whmkp8
- Mysterious radiation burst recorded in tree rings go.nature.com/zn56y6
- US agencies tackle dearth of evidence on painkillers go.nature.com/rladj2

VIDEO



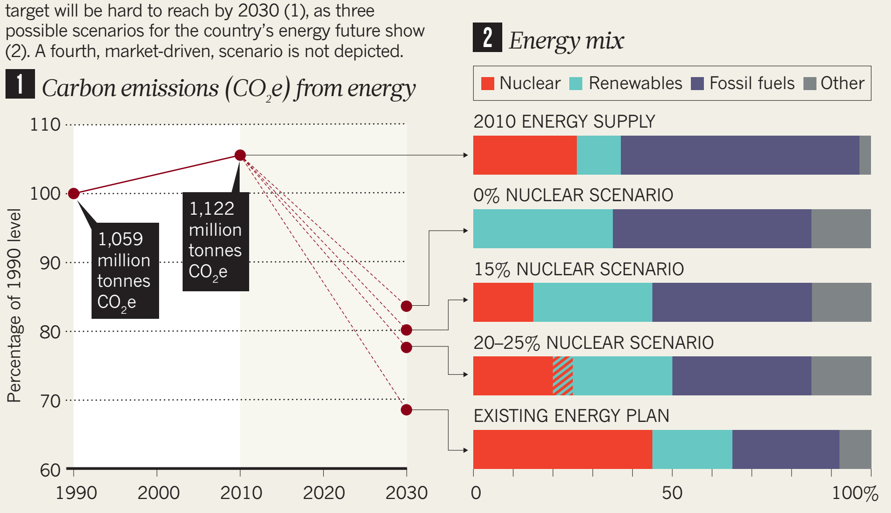
Watching an embryo as it develops cell-by-cell go.nature.com/bywonh

NASA, ESA, Z. LEVAY, R. VAN DER MAREL, T. HALLAS, A. MELLINGER

KELLER LAB (JANELIA FARM HHMI)

ENERGY SEESAW

If nuclear power is reduced, Japan's existing carbon target will be hard to reach by 2030 (1), as three possible scenarios for the country's energy future show (2). A fourth, market-driven, scenario is not depicted.



SOURCE: MINISTRY OF ECONOMY TRADE AND INDUSTRY

POLICY

Japan considers nuclear-free future

Options require big boost for renewable energy sources.

DAVID CYRANOSKI, TOKYO

It's official: nuclear power will have a much smaller role in Japan's energy future than was once thought. Since the meltdowns and gas explosions at the Fukushima Daiichi nuclear power station in March 2011, all of Japan's remaining reactors have been shut down for inspections and maintenance. Last week the government offered a glimpse of their future, and that of the country's nuclear power in general, when it published an outline of four ways to satisfy Japan's future energy demands. One scenario recommends using a market mechanism to determine the nuclear contribution. Under the other three, nuclear power would supply at most one-quarter of Japan's energy by 2030 — and in one case, none at all.

The scenarios come from a 25-person advisory committee to the industry ministry. The committee has been meeting since last October to discuss revisions to the 2010 Basic Energy Plan, which had proposed that nuclear energy would generate 45% of the country's electricity by 2030. The sharp reductions in that proportion mean that Japan will struggle to reach the 31% reduction in carbon dioxide emissions that it had planned by 2030; three of the new scenarios post more modest targets of 16%, 20% or 23% (see 'Energy seesaw'). A fifth plan included a heavier dependence on nuclear power (35%), enabling greenhouse-gas reductions of 28%, but committee head Akio Mimura, president of Nippon Steel based in Tokyo, said that he had made the "heartbreaking" decision to discard that option because of popular opposition to nuclear energy.

Before Fukushima, Japan's energy plans depended on an ambitious expansion of its nuclear capacity, which accounted for 26% of the country's electricity in 2010. That plan faced opposition before the disaster (see *Nature* **464,** 661; 2010) and is now all but dead, with local governments blocking efforts to restart existing plants. Some prefectures, including Fukushima, hope to use renewable energy sources to become self-sufficient without reliance on nuclear energy. Toru Hashimoto, mayor of Osaka, has used his opposition to restarting the reactors to become, according to polls, the country's most popular politician. Some tip him to be the next prime minister.

Given that opposition, Tetsunari Iida, director of the Institute for Sustainable Energy Policies in Tokyo, says anyone with sense would forget nuclear power. "Advocates of nuclear have no idea of reality check. They just keep repeating, 'we can do it,'" says Iida, who advises Hashimoto on energy.

➲ WWW.NATURE.COM/JAPANQUAKE

The nuclear-free plan calls for renewable energy sources including wind and solar to provide 35% of Japan's electricity in 2030, up from 11% now. Iida, a member of the committee and a supporter of the plan, says that the goal is achievable. Solar capacity has been growing quickly since 2009, with a feed-in tariff allowing people with solar panels on their homes to sell energy back to the grid at a premium price. A similar subsidy for industrial solar plants will come into effect on 1 July, and companies including Kyocera, a Kyoto-based solar-technology manufacturer, are developing 'megasolar' farms. Iida expects to see an extra 3–5 gigawatts of solar capacity over the next year.

But many consider an over-reliance on renewables to be expensive and unrealistic. The no-nuclear scenario could reduce the country's gross domestic product by up to ¥31 trillion (US$396 billion) per year, according to the committee. Some local government officials, including Hashimoto, have begun to soften their position on nuclear energy in a bid to protect jobs, and two reactors in Fukui prefecture look set to be restarted.

Tatsujiro Suzuki, vice-chairman of the cabinet-level Japan Atomic Energy Commission, admits that it will be "very difficult" to sell the country on nuclear energy again. He says the government must first prove that Fukushima is safe and that contamination from the disaster poses no threat, and then put in place new regulatory bodies. The government has decided to replace the industry ministry's tainted Nuclear and Industrial Safety Agency, which has been criticized as too close to industry, with an independent nuclear regulator. "Trust has been damaged. We have to rebuild it," says Suzuki.

But he recognizes that no new reactors are likely to be built in the next two decades. Without them, Japan would need to use some of the existing reactors past their intended 40-year life spans to reach the scenario in which 20–25% of power is nuclear. That would not be popular in a country skittish about nuclear safety.

The most likely option now seems to be the 15%-nuclear scenario. Under current estimates of future demand, that is roughly what nuclear energy would produce in 2030 if current reactors are restored to service, then retired as they reach 40 years.

The cabinet-level Energy and Environment Council will consider the four scenarios over the summer. It will also weigh advice from the Japan Atomic Energy Commission on whether to consign nuclear waste to permanent storage or persist with another pre-Fukushima goal: to develop reprocessing facilities that recycle nuclear waste into fresh reactor fuel.

Iida is happy that energy efficiency and renewable energy sources play a big part in all of the scenarios, but he is not convinced that the popular opposition to nuclear power will be reflected in the policy that emerges. "This country is not so democratic," says Iida. ■

MEDICINE

Antibody alarm call rouses immune response to cancer

Trial drug outperforms earlier efforts to marshall the body's defences to combat tumours.

BY ERIKA CHECK HAYDEN

Researchers working in the burgeoning field of cancer immunotherapy last week announced that a way of arming the body's natural defences to fight tumour cells has proved effective against three different types of cancer.

An antibody-based treatment developed by Suzanne Topalian, an oncologist at the Johns Hopkins University in Baltimore, Maryland, and her colleagues either eliminated or shrank tumours in 49 of 236 patients with certain types of advanced skin, kidney and lung cancer. Previous cancer immunotherapies have worked in smaller percentages of patients. The results of the phase I clinical trial were published on 2 June (S. L. Topalian *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1200690; 2012).

"I think it really changes the field, because the response rates are much higher," says Antoni Ribas, a cancer researcher at the Jonsson Comprehensive Cancer Center of the University of California, Los Angeles, who is testing a similar treatment in clinical trials.

The latest therapy works by reactivating T cells, which identify 'foreign' cells and either bind to and destroy them or recruit other immune cells to make the attack. Proteins on the surface of some tumour cells can lull T cells into quiescence by binding to a receptor known as 'programmed cell death 1' (PD-1), effectively disarming the body's immune response. The immunotherapy blocks tumour cells from binding to PD-1, thereby reactivating T cells and allowing them to orchestrate an attack on the cancer (see 'Waking up the body's defences').

The antibody, made by researchers at Bristol-Myers Squibb in Princeton, New Jersey, also had long-lasting effects for some patients in the latest trial. Twenty people responded to the therapy for a year or longer — an unusually durable response for patients with advanced cancers.

This type of immunotherapy has been posited as a potential cancer treatment for a century, but has only recently begun to bear fruit. In 2010, the US Food and Drug Administration (FDA) approved sipuleucel-T, which is marketed under the trade name Provenge by Dendreon of Seattle, Washington. The treatment is a vaccine made by incubating a patient's own immune cells with molecules that stimulate the cells to attack prostate tumours.

In 2011, the FDA approved another drug, ipilimumab, to treat melanoma. Sold by Bristol-Myers Squibb as Yervoy, the drug targets cytotoxic T-lymphocyte antigen 4 (CTLA-4), which, like PD-1, is exploited by tumour cells to stop T cells from attacking. However, CTLA-4 binding can occur between T cells and other cells in the body, whereas the PD-1 binding partner is expressed only by tumour cells.

Sipuleucel-T and ipilimumab work in relatively few patients. In the clinical trial that led to approval of ipilimumab, for instance, the drug was effective in 11% of patients with melanoma. Sipuleucel-T is estimated to have response rates lower than 10%. By contrast, anti-PD-1 therapy produced responses in 18% of patients with lung cancer, 28% of patients with melanoma and 27% of patients with kidney cancer. "If we break the 10% ceiling, it's becoming even more important and clinically relevant," Ribas says.

That the drug also works in lung cancer — which kills more people each year than any other cancer — is a breakthrough, says Jedd Wolchok, an oncologist at the Memorial Sloan-Kettering Cancer Center in New York, who is testing therapies aimed at blocking PD-1 in clinical trials. "The most important lesson that the practising oncologist is going to take away is that immunotherapy is now relevant to some of the very common diseases that they see," Wolchok says.

Because stimulating the immune system can induce deadly autoimmune responses, a major concern with any immunotherapy is safety. Three patients in the anti-PD-1 trial — 1% of those treated — died from lung inflammation caused by the trial drug. Overall, 11% of the patients had serious side effects.

"If we can get away from the significant adverse events that we've seen and help larger numbers of patients, then immunotherapy can be a game changer," says Ira Mellman, vice-president of research oncology at Genentech in South San Francisco, California.

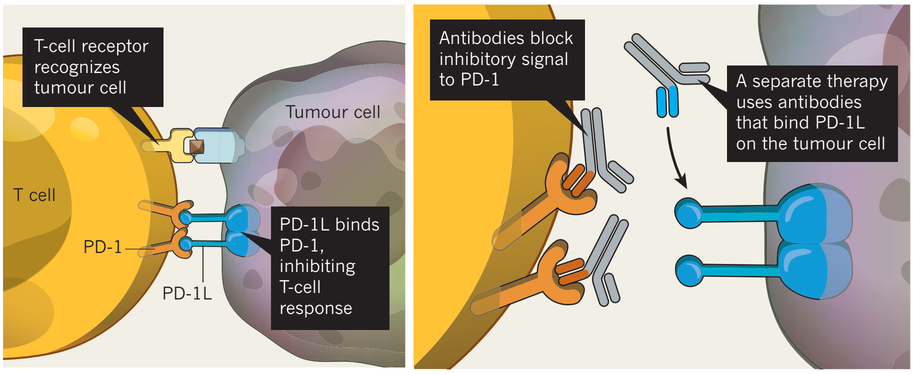
Researchers at Genentech and elsewhere are testing a similar therapy that targets the PD-1 ligand (PD-1L), the PD-1 binding partner expressed by tumour cells. Mellman predicts that targeting PD-1L might reduce side effects by taking aim at the tumour cells while leaving T cells free to bind to molecules that prevent autoimmune reactions.

It is not clear why certain cancers seem to respond better than others to drugs that block PD-1. Although the anti-PD-1 therapy had no effect on colon cancer in the most recent trial, the treatment did work for one patient with colon cancer in an earlier trial. That patient is still well four years later, Topalian says. She adds that identifying a biomarker to predict who will respond to the therapy is a priority. "We really think we're only scratching the surface of the potential of this drug," she says. ■

➲ NATURE.COM
Read more about cancer immunotherapy.
go.nature.com/vdannw

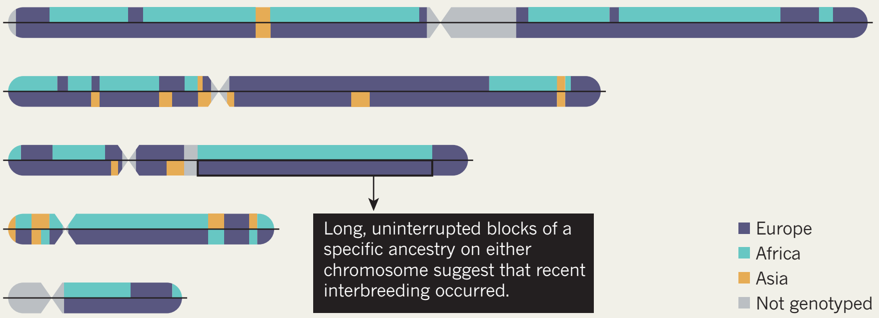
WAKING UP THE BODY'S DEFENCES

Tumour cells can inhibit the body's immune response by binding to proteins, such as PD-1, on the surface of T cells. Antibody therapies that block this binding reactivate the immune response.


T-cell receptor recognizes tumour cell
Tumour cell
T cell
PD-1
PD-1L
PD-1L binds PD-1, inhibiting T-cell response
Antibodies block inhibitory signal to PD-1
A separate therapy uses antibodies that bind PD-1L on the tumour cell

PAINTING YOUR ANCESTRY

Chromosome painting compares DNA in various stretches of a person's genome to the same regions in reference populations worldwide. This helps to determine a person's origins in detail. Shown are 5 of the 23 pairs of chromosomes from an African American who has mixed African, European and Asian ancestry.



Long, uninterrupted blocks of a specific ancestry on either chromosome suggest that recent interbreeding occurred.

SOURCE: 23ANDME

GENOMICS

Ancestry testing goes for pinpoint accuracy

Companies use whole genomes to trace geographical origins.

BY EWEN CALLAWAY

Condoleezza Rice, former US Secretary of State and national security adviser, ought to be a tough woman to surprise. Yet when Henry Louis Gates Jr, host of a US television series called *Finding Your Roots*, revealed that nearly half of her genetic ancestry could be traced to Europe, Rice, an African American, told Gates, "I'm stunned."

Although it is no secret that many African Americans have some European ancestry — a legacy of the transatlantic slave trade — advances in DNA analysis are beginning to provide more detailed insight for individuals. Commercial ancestry testing, once the province of limited information of dubious accuracy, is taking advantage of whole-genome scans, sophisticated analyses and ever-deeper databases of human genetic diversity to help people to answer a simple question: where am I from?

Until a few years ago, most ancestry tests for individuals relied on short stretches of DNA in cell-powering organelles called mitochondria, which are inherited through the mother, or on the Y chromosome, which a father passes down to his sons. As humans fanned out from Africa some 40,000 to 80,000 years ago and populated the world, mitochondria and Y chromosomes developed specific changes that were tied to different populations. Yet these 'uniparental markers', which chart an unbroken chain back through either the maternal or paternal line, are rarely unique to a population.

For example, a set of Y-chromosome markers called Haplogroup R1b is common among Western European men, but a small fraction of North Africans have it, too. Similarly, "if men have a Y chromosome that is more common in Scandinavia than England, they're convinced they're a Viking", says Mark Jobling, a geneticist at the University of Leicester, UK. But that is not necessarily the case. Such nuances are not always conveyed by the companies that offer such services, notes Jobling. What's more, Y-chromosome or mitochondrial markers trace only one strand in a person's ancestry.

A more complete picture lurks in the full genome. It is trickier to tease out ancestry information from our 22 pairs of non-sex chromosomes, because whole stretches of DNA in these are mixed up in every generation. But in recent years, researchers have made strides mapping the ancestry of certain populations — including Europeans (J. Novembre *et al. Nature* **456**, 98–101; 2008) and Indians (D. Reich *et al. Nature* **461,** 489–494; 2009) — by simultaneously analysing hundreds of thousands of single DNA letter changes across the genome. These surveys of human genetic diversity, including the International HapMap Project and the 1000 Genomes Project, have provided the data needed for more sophisticated ancestry testing.

"If men have a Y chromosome that is more common in Scandinavia, they're convinced they're a Viking."

Genetics company 23andMe, based in Mountain View, California, was one of the first to offer its customers ancestry analysis based on whole-genome scans, at a cost of US$300–400. The company worked closely with the television network PBS on the latest series of *Finding Your Roots*, which uses whole-genome analysis to investigate the genealogy and deeper ancestry of figures such as Rice, actor Kevin Bacon and celebrity entrepreneur Martha Stewart.

The discovery that a significant amount of Rice's genetic ancestry can be traced to Europe came from an analysis called chromosome painting (see 'Painting your ancestry'), in which genetic variations in small chunks of each chromosome are compared with matching regions in different populations around the world (European, West African and East Asian in the case of 23andMe, although the company is soon due to extend the analysis to 20 populations, including Native American and South Asian).

Carlos Bustamante, a population geneticist at Stanford University in California, advises 23andMe and Ancestry.com, which runs a testing service called AncestryDNA from its base in Provo, Utah. He says that he works with scientists at the companies to make sure that the information they offer customers stands up to scrutiny. For instance, Ancestry.com's predictions — which include possible connections to customers who might be distant relatives, on the basis of shared stretches of chromosomes — come with confidence estimates. "The important thing is to say 'here are the limits of what we know,'" says Bustamante.

As ancestry testing trawls deeper into the genome, such cautionary notes will become more important. It may be possible, for instance, to determine more precisely when someone's ancestors from different populations interbred, says Sarah Tishkoff, a geneticist at the University of Pennsylvania in Philadelphia. But such calculations make assumptions about past population sizes and other demographic factors. Conveying such caveats to personal-genomics customers could prove difficult.

Individuals may soon be able to trace the geographic origins of their ancestors more precisely. An academic project called People of the British Isles has distinguished the genetic signatures of people from neighbouring UK counties. This level of precision was attained by analysing the genomes of people from rural regions, whose ancestors tended to live in the same place, says project leader Walter Bodmer, a geneticist at the University of Oxford, UK. Genome information from these individuals could form a database that others could use to track their distant ancestors.

Tishkoff says that deeper surveys of human genetic diversity are needed elsewhere, particularly in Africa. Angola, for example, where many slaves originated, is not represented in existing surveys, making it impossible for African Americans to trace ancestors who once lived there. "You can't tell someone they can trace ancestry to a certain region if that region has never been studied," she says. ■

SECOND CHANCE FOR THE PLANET

TWENTY YEARS AGO, when the world's leaders pledged to protect Earth's climate and biodiversity at the Rio Earth Summit, they knew it would not be easy. But few could have guessed how much worse the situation would get. In 1992, the atmosphere held fewer than 360 parts per million (p.p.m.) of carbon dioxide; the concentration is now nearing 400 p.p.m. and surging upwards. At the same time, species are disappearing at an accelerating rate.

On the eve of the second Rio Earth Summit, *Nature* explores the causes and consequence of those changes, as well as the efforts that are being made to avert the worst outcomes. Our assessment shows how little progress nations have made towards honouring the commitments they made in 1992 (see pages 5 and 20).

There are some success stories. Brazil has made remarkable advances in curbing deforestation and setting aside protected swathes of land. But with its more recent policies raising concern, the country must do even more to safeguard its environment, argue representatives of Conservation International (see page 25). And although international environmental efforts have stalled, a bottom-up approach that changes how corporations operate could turn a profit for both the planet and for business, says Pavan Sukhdev (see page 27).

RETURN TO RIO

For Earth Summit news:
www.nature.com/rio20

In sharp contrast to the political stalemate over the past two decades, scientists have developed a more sophisticated understanding of the roots and effects of the current environmental crisis. Articles and reviews elsewhere in this issue explore how shrinking biodiversity is affecting ecosystems and how much the voracious consumption patterns of the developed world are to blame — see Bradley Cardinale *et al.* (page 59), David Hooper *et al.* (page 105) and Manfred Lenzen *et al.* (page 109).

Anthony Barnosky and his colleagues (page 52) argue that the global ecosystem could eventually pass a tipping point and shift into a new state, the likes of which are hard for science to predict. But there are ways to avoid that fate, say Paul Ehrlich and his colleagues (page 68), who suggest techniques to make societies more sustainable and to head off many of the world's chronic environmental problems.

Earth and its inhabitants have a second chance in Rio. They may not get many more. ■

JULENE HARRISON (PHOTO: ILKER CANIKLIGIL/SHUTTERSTOCK)

RIO REPORT CARD

The world has failed to deliver on many of the promises it made 20 years ago at the Earth summit in Brazil.

BY JEFF TOLLEFSON & NATASHA GILBERT

The tropical air was charged with hope and despair as the world's leaders descended on Rio de Janeiro for the United Nations' Earth summit in May 1992. Countries were buoyed by a string of successful environmental treaties in the 1970s and 1980s, capped by a landmark deal to save the ozone layer in 1987. Yet the Earth summit in Rio, which drew 178 nations and around 100 heads of state, was also rife with frustration and distrust. Diplomats had spent the previous two years drafting a pair of treaties intended to safeguard Earth's biodiversity and climate, but the talks had recently faltered as rich and poor countries split over who should pay for protecting the planet.

In the end, the leaders decided that they could not go home empty handed. They signed off on both the Convention on Biological Diversity and the Framework Convention on Climate Change, making broad pledges to solve some of the most complex problems facing humanity. Countries also agreed to a laundry list of goals spelled out in a document known as Agenda 21, which eventually spawned the Convention to Combat Desertification. Although the agreements lacked teeth, they created formal international processes that engaged almost the entire world and eventually led to more targeted accords (see 'Global awakening').

At the end of the summit, Richard Benedick, who had negotiated the ozone accord for the United States, told *The New York Times* that "the history books will refer back to this day as a landmark in a process that will save the planet from deterioration". But he and others warned that progress would not come quickly.

The pace turned out to be far slower than anticipated, however. Although nations have made some marginal advances, the three conventions have failed to achieve even a fraction of the promises that world leaders trumpeted two decades ago. Dismal grades dominate *Nature*'s report cards on the Rio treaties, although the assessment also highlights some progress and offers pointers for the future. As diplomats and leaders prepare to converge on Rio this month for the UN Conference on Sustainable Development, or Rio+20, they will be looking back to consider how to do better.

RETURN TO RIO
For Earth Summit news:
www.nature.com/rio20

CLIMATE OF INACTION

The climate numbers are downright discouraging. The world pumped 22.7 billion tonnes of carbon dioxide into the atmosphere in 1990, the baseline year under the UN Framework Convention on Climate Change. By 2010 that amount had increased roughly 45% to 33 billion tonnes. Carbon dioxide emissions skyrocketed by more than 5% in 2010 alone, marking the fastest growth in more than two decades as the global economy recovered from its slump. And despite constant deliberations under the convention, the overall growth rate of global emissions hasn't changed much since 1970 (see 'Report card: UN Framework Convention on Climate Change').

"Plausibly we are a little better off than if we didn't have all of this diplomacy," says David Victor, director of the Laboratory on International Law and Regulation at the University of California, San Diego. "But the evidence is hard to find."

Ratified by 194 countries plus the European Commission, the treaty sought to stabilize emissions at a level that would "prevent dangerous anthropogenic interference with the climate system". Although there were no specific targets, wealthy countries agreed to take the lead and help poor countries with monetary and technological aid. In 1997, negotiators followed up with the Kyoto Protocol, which entered into force in 2005 and committed industrialized countries to reduce their collective emissions of all greenhouse gases by 5.2% (compared with 1990) by 2012.

Overall, industrialized countries are on track to surpass the Kyoto goal with a reduction of some 7%, but this is largely due to the demise of the Soviet Union and its inefficient factories, as well as to the industrial slump caused by the recent economic crisis, which is starting to reverse. The United States, the developed world's largest greenhouse-gas producer, never ratified the protocol and increased its greenhouse-gas output by 11% between 1990 and 2010. In the meantime, developing countries more than doubled their emissions, increasing their share of the global total from 29% to 54%.

In spite of the failure to rein in emissions, the climate treaty has performed better on many lesser goals. The international process it spawned encouraged investment in climate science and provided a venue for scientists and policy experts to showcase their work. Periodic scientific reports by the Intergovernmental Panel on Climate Change (IPCC) underpinned each major round of treaty talks. The negotiations also helped to raise awareness of climate change across the globe. Governments began working on climate adaptation, sustainable agriculture and reducing tropical deforestation, and Kyoto sparked experimentation with carbon markets and new ways of transferring money and technology to poor countries.

But on the core challenge of overhauling the global energy industry and reducing emissions, the questions remain the same 20 years later: who must do what and who pays?

The original treaty introduced the notion of "common but differentiated responsibilities", with a heavier burden on wealthier countries, historically responsible for the largest share of greenhouse-gas emissions. That concept was put into practice through the Kyoto Protocol, when industrialized countries agreed to reduce emissions and provide aid to developing countries, which took on no formal obligations. But as the world changed and the proportion of emissions increasingly shifted towards developing countries, the treaty remained static.

JULENE HARRISON (PHOTO: CELSO DINIZ/SHUTTERSTOCK)

The result has been a prolonged stand-off, with poor nations demanding that their wealthy neighbours do more and industrialized countries increasingly concerned about skyrocketing emissions among the rapidly emerging economies. In particular, the United States baulked at the idea of moving forward without China, which is now the world's largest emitter, whereas China cited its lower per-capita emissions in questioning whether the United States is doing enough. Negotiators wrestled with those issues at the 2009 climate summit in Copenhagen, where China, Brazil, South Africa and other major developing countries promised for the first time to reduce emissions. Last December in Durban, South Africa, countries agreed to negotiate a new global climate treaty by 2015 that would include formal commitments from both developed and developing countries.

REPORT CARD

UN FRAMEWORK CONVENTION ON CLIMATE CHANGE

MAIN ASSIGNMENT

STABILIZE GREENHOUSE-GAS EMISSIONS	F

OTHER ASSIGNMENTS

TRACK GREENHOUSE-GAS EMISSIONS AND SINKS *The climate convention has helped to create national inventories of greenhouse-gas emissions, land-use trends and carbon uptake by forests.*	A
PROMOTE AND DISPERSE CLIMATE-FRIENDLY TECHNOLOGIES *The Clean Development Mechanism allows industrialized countries to offset their emissions by paying for clean energy and other projects in developing countries, but the programme has been limited in both reach and effectiveness.*	D
PROMOTE SUSTAINABLE LAND MANAGEMENT *The climate talks have encouraged efforts to advance sustainable agriculture and reduce tropical deforestation.*	C
PREPARE FOR THE IMPACTS OF CLIMATE CHANGE *Many of the 194 countries that are party to the convention have only recently begun formulating plans to prepare for a warmer world.*	C
ADVANCE CLIMATE RESEARCH AND POLICY ANALYSIS *The UN process has encouraged investments in climate science, energy technologies and social sciences.*	A
ESTABLISH A DIPLOMATIC PROCESS *The annual 'Conference of the Parties' to the climate convention, or COP, has become an international roadshow for professional climate diplomats.*	A

Climate negotiators will gather in Doha, Qatar, in November to begin the process of designing that new treaty, but scepticism remains. "The only way we are going to achieve significant emissions reductions is through technology," says Barry Brook, director of climate-change research at the University of Adelaide's Environment Institute in Australia. The fundamental barrier is the cost of clean-energy alternatives, he says. "A lot of this can't be driven by an international process."

Some argue that the climate talks might be more fruitful if the focus were on securing agreement within groups of major economic powers such as the G20, which is responsible for more than 80% of global emissions. But even if the cacophony of voices in the UN negotiations makes progress difficult, many believe that the process has helped to inspire countries, local governments and even corporations to tackle the issue of climate change in a more serious way.

"What we have today is nowhere near what the science says we need," says Manish Bapna, acting president of the World Resources Institute in Washington DC. "But is it closer than we would have been in the absence of climate negotiations? I would say the answer is an unequivocal yes."

BIODIVERSITY ON THE SIDELINES

"Let us have the courage to look in the eyes of our children and admit that we have failed." That stark message came from Ahmed Djoghlaf in October 2010, when he addressed the 193 parties to the Convention on Biological Diversity (CBD) at a summit in Nagoya, Japan. As executive secretary of the CBD at the time, Djoghlaf lamented that countries were nowhere near to meeting the treaty's chief goal of "significantly" cutting species loss by 2010 (see 'Report card: Convention on Biological Diversity'). Instead, he said, "we continue to lose biodiversity at an unprecedented rate".

Some 30% of amphibians, 21% of birds and 25% of mammal species are at risk of extinction, according to the International Union for Conservation of Nature (IUCN) based in Gland, Switzerland. The CBD has failed to slow the problem, say biodiversity scientists, because it did not set concrete and focused targets, and it provided no means to measure progress towards protecting wildlife and ecosystems.

At the Nagoya meeting, countries agreed on a set of 20 goals — the Aichi targets — which include halving the rate of loss of natural habitats, one of the biggest threats to biodiversity, by 2020. Another target seeks to protect 17% of the world's land area in nature reserves by 2020. In addition, the CBD parties put money towards developing better indicators for measuring progress.

The 20 Aichi targets are a step in the right direction but they still miss the mark, warn scientists and conservationists. "The Aichi targets are still not very focused and they add no obligations on countries to comply with them. There is an unwillingness among countries to accept obligations," says Stuart Harrop, a wildlife-management lawyer and director of the Durrell Institute of Conservation and Ecology at the University of Kent in Canterbury, UK.

Another long-standing problem with the CBD has been that it lacked a dedicated body, similar to the IPCC, that would provide scientific advice and help it to define quantifiable targets. The CBD gained an equivalent scientific arm only two months ago, when the Intergovernmental Platform on Biodiversity and Ecosystem Services was launched. "It has not been a science-based convention," says Anne Larigauderie, a plant ecologist and executive director of DIVERSITAS, an international biodiversity research programme headquartered in Paris.

In addition, countries lack the observational infrastructure to track the state of their national biodiversity. The CBD currently relies on data compiled by conservation groups, including the IUCN's Red List of threatened species. Poor investment in observation systems means that there are still large gaps in the data on local and global biodiversity, says Larigauderie.

Lack of funding for biodiversity conservation has also constrained progress, says Cyriaque Sendashonga, a zoologist and director of global policy at the IUCN. In Nagoya, countries agreed to report on their biodiversity spending at the CBD summit this October in Hyderabad, India. They will also discuss ways to boost spending, including redirecting current subsidies that are harmful to the environment towards conservation actions. This could generate US$50 billion annually for biodiversity, says Sendashonga.

In the end, progress on preserving global biodiversity has stalled

REPORT CARD

CONVENTION ON BIOLOGICAL DIVERSITY

MAIN ASSIGNMENT

REDUCE THE RATE OF BIODIVERSITY LOSS	F

OTHER ASSIGNMENTS

DEVELOP BIODIVERSITY TARGETS *Nations have only just started to establish focused targets for biodiversity and ways to assess it.*	D
PROTECT ECOSYSTEMS *At least 10% of the world's ecologically valuable regions on land was protected by 2010, but only about 1% of those in the oceans.*	C
SHARE GENE WINDFALL *The Nagoya Protocol on the sharing of commercial benefits derived from the collection and use of genetic material has been signed by 92 countries, but is not yet in force. Only a few companies so far have shared such benefits with the source country.*	E
RECOGNIZE INDIGENOUS RIGHTS *Nations are very variable in honouring the rights of indigenous people, especially in creating protected areas within their territory.*	D
PROVIDE FUNDING *Countries have made many commitments but honoured few of them.*	F
REGULATE GENETICALLY MODIFIED ORGANISMS *The Cartagena Protocol, signed by 103 countries, is designed to help regulate the movement of genetically modified organisms between countries, and came into force in 2003.*	A

GLOBAL AWAKENING

The treaties that emerged from the 1992 Rio summit followed several major environmental agreements and spawned a series of subsequent accords.

1972 The United Nations Conference on the Human Environment in Stockholm is the first major international meeting devoted to environmental problems.

1987 The Montreal Protocol on Substances that Deplete the Ozone Layer requires nations to eliminate chemicals that harm stratospheric ozone.

1992 RIO SUMMIT

UN FRAMEWORK CONVENTION ON CLIMATE CHANGE (UNFCCC)

1992 Adoption of general climate treaty without specific targets.

1997 The Kyoto Protocol limits greenhouse-gas emissions from industrialized countries.

CONVENTION ON BIOLOGICAL DIVERSITY (CBD)

1992 Agreement on convention to conserve biological diversity.

2000 The Cartagena Protocol on Biosafety regulates the transport of genetically modified organisms.

UN CONVENTION TO COMBAT DESERTIFICATION (UNCCD)

1994 Treaty signed to prevent and reverse land degradation.

U. MARCELINO/REUTERS

Indigenous people of the Amazon protesting against construction of the Belo Monte dam, which they fear will damage the Xingu River.

Lead by example

As host nation of Rio+20, Brazil should choose the right course for its own development, say **Fabio Scarano**, **André Guimarães** and **José Maria da Silva**.

The United Nations Conference on Sustainable Development, which returns to Rio de Janeiro this month 20 years after the Rio Earth Summit of 1992, will be held under a cloud. The cities of Copenhagen, Nagoya, Cancún, Changwon and Durban have all recently played host to meetings of the three major conventions that were established at the first Rio Summit — on biodiversity (CBD), climate (UNFCCC) and desertification (UNCCD). All are now bleak reminders of humankind's inability to deliver on sustainable development goals. So does Rio stand a chance of being more than just a collective moan about past failures?

The answer depends in large part on the actions of the host country, which can set the tone for such meetings. Brazil offers cause for optimism — it has progressively led negotiations to set ambitious sustainable development targets for the planet in recent years[1], and some innovative projects are under way at the state level. Yet the federal government has made decisions on home turf that go against the same global policies that it advocates.

RETURN TO RIO
For Earth Summit news:
www.nature.com/rio20

Brazil is at a crucial juncture, and needs to decide whether to develop sustainably, or in traditional ways that endanger natural capital. As one of 17 nations that together contain 70% of the planet's biodiversity, Brazil is 'megadiverse'; it holds 12% of the world's fresh water, and is the largest terrestrial carbon sink. It is also thriving financially: the country survived the economic crisis, becoming the world's sixth largest economy. Yet Brazil ranks 84th on the United Nations Development Programme's human development index, owing to problems with social inequity and poverty. This makes it the perfect venue for the Rio+20 meeting, which will focus precisely on how to increase ▶

▶ human well-being while maintaining or enlarging natural assets. If it is to lead by example, the nation must choose the right course for further development now.

Biodiversity conservation is one area in which the nation is facing a critical decision point. In 2003–08, Brazil was responsible for 70% of new land protection on the planet: about 50% of the Brazilian Amazon is now inside protected areas and indigenous territories, which has substantially reduced deforestation rates. But 2011, the first full year of office for President Dilma Rousseff, saw an embarrassing mark on its track record: for the first time in more than 15 years, the federal government did not create any new protected areas and, worse, it reduced the area covered by some of them.

DAM DAMAGE

The government has allowed the creation of new hydropower plants on undisturbed Amazonian rivers at the expense of indigenous locals and the environment — allegedly because of the nation's growing energy needs. It accelerated the construction of the Santo Antônio and Jirau dams on the Madeira River in 2009, and the Belo Monte dam on the Xingu River in 2011.

In January 2012, it decided to reduce the size and move the boundaries of eight protected areas in the Tapajós region in central Amazon to allow construction of yet more dams. That move was challenged in Brazil's Supreme Court in February by the federal public ministry, which said it was unconstitutional. But in May, Brazil's Congress approved the government's decision. With the construction of the Tapajós dams yet to begin, there is hope that this decision might be reversed. There are alternatives: the country should instead consolidate power generation on the rivers that already provide 80% of its energy, increase efficiency in energy transmission and invest seriously in research on alternative energy sources.

In 2010, Brazil's Congress launched an innovative policy to reduce the carbon footprint of agriculture, by providing incentives for farmers to use sustainable practices that mitigate and reduce greenhouse-gas emissions. Yet in late April 2012, the same Congress approved changes in Brazil's Forest Code that forgive past acts of illegal deforestation, thus reducing the requirement for rural landowners to conserve or restore natural land cover on their properties. The government's Institute for Applied Economic Research estimates that, as a result, nearly 47 million hectares of natural ecosystems could be lost in years to come[2].

This seriously undermines Brazil's commitment to reduce Amazon deforestation by 80% by 2020, made by former president Luiz Inácio Lula da Silva at the Copenhagen climate conference in 2009. The last hope for a reversal of this ugly scenario lies with a partial veto from President Rousseff, who rejected some of the proposed changes on 25 May — but this must still be approved by the Congress.

Brazil does not need more deforested land to increase its agricultural production[3]. The country has some 60 million hectares of fertile soils that are currently being used for unproductive cattle-raising at an average of one head of cattle or fewer per hectare. By comparison, 62 million hectares are being used for highly productive, modern agribusiness. By making cattle-raising more intensive and expanding agriculture into the freed space, Brazil could arguably double its production of food, fibres, fuel and commodities without cutting down a single tree. Such land reform, however, is a politically sensitive issue.

The Brazilian Congress will face another controversy in July 2012. It will vote on a bill that would allow mining activities inside indigenous reserves by paying royalties to indigenous peoples.

"Some intriguing green economy initiatives have emerged at the municipal and state level."

With regard to marine conservation, Brazil negotiated in favour of a 10% marine protection target by 2020 at the last CBD conference in Nagoya, Japan, in 2010. Yet only 1.5% of its exclusive economic zone is protected, and an estimated 80% of Brazilian marine fisheries are overexploited. More marine protected areas are obviously needed. Yet some estimates indicate that nearly 9% of priority areas for marine conservation have already been conceded to oil companies in Brazil for offshore exploration[4].

LOCAL LEADERSHIP

These examples clearly show that the government often acts as if development and environmental conservation were opposing forces. Yet some intriguing green economy initiatives have emerged at the municipal and state level.

In the Amazonian state of Acre, for example, a community-run, sustainable forest-management system that was launched in 2000 resulted, on average, in a two- to threefold increase in farmers' incomes by 2001, and a 12-fold increase in the value of rural property by 2012, compared with non-participating farms[5,6]. A few years after Amapá state initiated a conservation network to protect 72% of its territory in 2003, the state showed some of the highest annual growth rates for human development in Brazil. And in 2007, the state of Amazonas launched the Bolsa Floresta programme, an initiative that provides financial compensation and health assistance to locals in exchange for zero deforestation of primary forests[7].

In another example, the state of Espírito Santo has launched a project to restore 200,000 hectares of altered landscape by 2025. The intention is to create natural corridors between remnants of native vegetation, to protect water resources and to provide alternative job and business opportunities in a state where oil, mining and forestry are expanding rapidly.

The federal government needs to follow these examples and do more to turn Brazil into a green superpower. In particular, we would like to see Brazil use Rio+20 to launch a $3-billion green development fund. The money could come from environmental compensation agreements with energy and mining industries; for example, Norte Energia, the company building the Belo Monte dam, is meant to pay the government some 3.3 billion reais (US$1.6 billion) alone, and the compensation for new offshore oil development is still under debate. This fund could be used for more initiatives at local and state level that promote human well-being while maintaining or enhancing natural capital. Perhaps 20% of the fund could be reserved to help other nations in South America and Africa to follow the same track. Such national commitments are not unprecedented: in 2008, Norway donated $1 billion to the Amazon Fund, which acts against deforestation.

The basic ingredients for Brazilian leadership are in place: political and economic stability, growing institutional capacity, a strong private sector, globally competitive academia and abundant natural capital. The country has a moral obligation to help Rio+20 to succeed. We hope that it takes such steps. The planet cannot afford to wait until Rio+30 for action. ■

Fabio Scarano, **André Guimarães** *and* **José Maria da Silva** *work with the non-profit environmental organization Conservation International in Rio de Janeiro, Brazil, and in Arlington, Virginia, USA.*
e-mail: f.scarano@conservacao.org

1. Mittermeier R. *et al. Natureza e Conservação* **8,** 197–200 (2010).
2. Instituto de Pesquisas Economicas Aplicadas *Código Florestal: Implicações do PL 1876/99 nas Areas de Reserva Legal* (2011); available at http://go.nature.com/rvutbc.
3. Martinelli, L. A. *et al. Biota Neotropica* **10,** 323–330 (2010).
4. Greenpeace *Mar, petróleo e biodiversidade: a geografia do conflito* (2010); available at http://go.nature.com/g5xwcw
5. Franco, C. A. & Esteves, L. T. in *XLVI Congresso da Sociedade Brasileira de Economia, Administração e Sociologia Rural* (2008); available at http://go.nature.com/tt1wgf.
6. Rocha, C. O. D. in *Simpósio de Manejo Florestal da Amazonia Brasileira* (2010); available at http://go.nature.com/5ib1ep.
7. Viana, V. M. *Estudos Avançados* **22,** 143–153 (2008).

M. TANTUSSI/BLOOMBERG/GETTY

The shoe firm Puma has taken innovative steps towards disclosing its environmental impact.

The corporate climate overhaul

The rules of business must be changed if the planet is to be saved, says **Pavan Sukhdev**.

There is an understandable tendency to put problems of the 'global commons' — climate change or biodiversity loss, for instance — in the hands of intergovernmental institutions. However, thus far these bodies have failed miserably. The United Nations Framework Convention on Climate Change (UNFCCC), born at the Earth Summit in Rio de Janeiro, Brazil, 20 years ago this month, has been unable to get its signatory governments to arrest greenhouse-gas emissions. And the Convention on Biological Diversity (the UNFCCC's non-identical twin, also born at that 1992 meeting) and its signatory governments have collectively failed to slow the loss of biodiversity.

These failures point to the need to recognize the key role of the private sector in determining economic direction and resource use globally. For effective climate-change or biodiversity solutions, members of the corporate world need to be brought to the table as ethical stewards of shared planetary resources, and not, as they have been so far, as self-interested exploiters of common wealth.

The private sector produces almost everything we consume, generating more than 60% of global gross domestic product (GDP) and employment. A study by the United Nations Environment Programme Finance Initiative[1] estimated that in 2008, annual corporate 'externalities' — the costs to society of the leading 3,000 public companies globally, in the form of emissions, freshwater use, pollution, waste and land-use change — added up to US$2.15 trillion, or 3.6% of global GDP. Two-thirds of this is due to the forecasted costs of climate-changing emissions. Such large costs to society of 'business as usual' show that today's corporation is, on average, an economic agent lacking in social purpose and focused on financial return to shareholders. Not surprisingly, it produces today's 'brown' economy, delivering private gains at the expense of public losses by increasing environmental risks and ecological scarcities.

Consumerism is often blamed, and consumers can indeed make crucial choices on the basis of how much material and energy is used for making, packaging and transporting goods. But on this road of economic choices, it is corporations, not consumers, in the driver's seat, and they are driving us in the wrong direction. Corporate advertising converts our insecurities into a chain of wants, needs and excessive demands, which have made our ecological footprint exceed the planet's ability to produce resources and absorb emissions — by more than 50% (ref. 2).

Corporate lobbying often influences national policies to create advantage for particular industries or companies, to the detriment of the public good. Around $1 trillion of harmful subsidies now support the brown economy, including more than $650 billion in price and production subsidies for fossil fuels, and more than $300 billion for mostly unsustainable agriculture and fisheries[3]. We are now consuming nature's capital, not its interest. And yet we have enshrined this corporate model in business law and practice, and, indeed, celebrated it as a crowning success of our times.

The rules of the game need to be changed, so that corporations can compete on the basis of innovation, resource conservation and satisfaction of multiple stakeholder demands — rather than on the basis of who is the most effective at influencing government regulation, avoiding taxes and obtaining subsidies for harmful activities in order to optimize shareholder returns. And those rules need to be changed quickly. The pace at which we are approaching the safe operating limits of our planetary systems[4] suggests that a new corporate model must be ready to dominate economies by 2020. Targeting 2050 or 2100 might be too late and, therefore, solely an academic exercise.

EVOLUTIONARY CHANGE

This new type of corporation, which I call Corporation 2020 (ref. 5 and www.corp2020.com), can still be profitable while contributing to a 'green economy': one that increases human well-being and social equity, and decreases environmental risks and ecological scarcities. Some examples show how this can work.

The German sportswear company Puma is leading the way in transparency and disclosure of its external costs to society. It measures, evaluates and publishes data on its carbon emissions, freshwater usage, pollution and waste. The unique aspect of this exercise is that Puma has measured and monetized these impacts, calculating them along its entire supply chain. It has effectively created the world's first environmental profit-and-loss statement. Although Puma disclosed an estimated €145 million (US$182 million) in such externalities for 2010, the revelation was far from the public-relations disaster that some had predicted. The firm is now using what it learned to engage its raw materials and manufacturing supply chain (which is where almost 95% of these externalities arise) to ▶

improve its environmental performance.

A true Corporation 2020 must not only measure and minimize negative externalities, but also actively create positive externalities. A good example of a future-thinking company in this regard is Infosys, an information-technology giant based in India, which is a veritable factory of 'human capital'. Infosys's campus in Mysore has the astonishing capacity to provide the equivalent of a bachelor's degree in computer science to 14,000 employees at a time. It is no coincidence that the firm receives more than one million job applications a year, has an exceptionally talented cadre of professionals and has compounded sales and earnings growth that averages 70%.

In addition to building human capital, a Corporation 2020 must build 'social capital'. Traditional types of community — villages and neighbourhoods — are no longer able to fulfil the human need for social ties. Competition, labour mobility and a culture of long working hours have taken a toll on people's ability to build strong connections where they live. Corporations can be part of the solution to rebuild these networks. An iconic example in the United States is the corporate headquarters of Google in Mountain View, California. By encouraging a sense of community and providing benefits such as gourmet food, laundry services and recreational activities, Google has created a culture in which employees feel proud and energized to come to work.

Forward-looking behaviour tends to be found mostly in companies that have foreseen that providing a benefit to society can also benefit themselves. However, not every desired behaviour of a Corporation 2020 is justifiable purely on the basis of self-interest. Good companies have to swim against the current of perverse taxes and subsidies, competition that cuts corners by hiding costs, and a legal history that supports slavish devotion to maximizing shareholder value. Significant changes in the operating environment for business will be necessary if the corporate world is to move to a green economy.

FOUR SOLUTIONS

Just as a biological species evolves in response to the natural environment and in turn influences its environment, today's corporation can evolve into Corporation 2020 if it has the right regulatory environment. But four vital planks of change must be put in place with urgency.

The first is disclosure of externalities. This will provide the missing information needed by corporate managers, governments, civil society, consumers and investors to differentiate their responses to different corporations. There are many financial reporting agencies around the globe, including the International Accounting Standards Board, the US Financial Accounting Standards Board and the UK Institute of Chartered Accountants of England and Wales, the rules of which require that corporations submit annual financial reports. These bodies should commission research and develop methodologies for measuring the most material corporate environmental externalities, as well as those from human and social capital. These externalities should also be disclosed in statutory annual reports.

Infosys, headquartered in Mysore, India, offers the equivalent of a bachelor's degree to thousands of employees simultaneously.

N. BHOJANI/BLOOMBERG/GETTY

The good news is that the UK institute has already formed a global coalition — the TEEB (The Economics of Ecosystems and Biodiversity) for Business Coalition — with the support of the UK and Singapore governments and the Gordon and Betty Moore Foundation in Palo Alto, California, to develop and promote the use of such methodologies and standards worldwide. Thus far the coalition has embarked on a preliminary exploration of the 100 largest externalities across various industries, locations and ecosystems. Over the next five years, these will be studied in depth to evolve standards for estimating and disclosing such externalities.

The second plank is accountable advertising. Misleading information, unfair persuasion and 'greenwashing' have been common practices recently. Advertising associations, encouraged by consumer-protection agencies and non-governmental organizations, will have to push for more 'information' value in their advertising, as opposed to 'selling' value. Social media are already changing the character of advertising: the power of consumers is growing. Advertising is now much more a conversation between consumers and companies than it used to be. This trend must be accelerated by both institutional support and industry leadership.

The third plank is to limit leverage — the proportion of borrowed funds versus owners' capital[6]. Excessive leverage has been the key driver of most major global recessions — the Latin American debt crisis of the 1980s, the US savings and loans crisis of the 1980s and 1990s, the 1997 Asian debt crisis and the recent mortgages crisis. Recessions, in turn, increase poverty and reduce social equity, acting against a crucial objective of a green economy. Corporations that are 'too big to fail', which effectively have recourse to public funds in times of crisis, now include not just banks but also insurers, mortgage lenders, car-makers, airlines and even hedge funds.

This cannot be a recipe for economic sustainability. G20 governments and central banks should monitor and limit the leverage of major corporations.

The fourth plank is resource taxation. The tax authorities of G20 governments should implement changes in the life-cycle incidence of taxation: more tax should be charged at the point of resource extraction (such as the mining of fossil fuels and minerals) rather than at the point of sale or the point of profit reporting. This will encourage efficient use of materials rather than more mining, more product and packaging and more waste. Sadly, a serious effort of this kind in Australia — an attempt to increase resource taxation in August 2010 — led to an avalanche of lobbying and advertising by mining companies that was so successful that it unseated Prime Minister Kevin Rudd, the main architect of the tax. Some may remember Rudd as the leader behind Australia's signature to the Kyoto Protocol. Although the tax will still be introduced, it has been greatly watered down.

The problems caused by today's corporate structure are complex. They require complex solutions, which can only be delivered through collaboration on a global scale, across a diversity of sectors and through at least these four major planks of change. Humanity has so far been neither willing nor able to engage in this manner. Yet engage we must, and we must begin now. The reward is unsurpassed: a sustainable economy, on a habitable planet. ■

Pavan Sukhdev *is the chief executive of the environmental consulting firm GIST Advisory, and served as head of the United Nations Environment Programme's Green Economy initiative from 2008 to 2011.*
e-mail: pavan@corp2020.com

[1] Principles for Responsible Investment and UNEP Finance Initiative *Universal Ownership: Why Environmental Externalities Matter to Institutional Investors* (Trucost, 2010).
[2] WWF *Living Planet Report 2012* (WWF, 2012).
[3] UNEP *Towards a Green Economy: Pathways to Sustainable Development and Poverty Eradication* (United Nations Environment Programme, 2011).
[4] Rockström, J. *et al. Ecol. Soc.* **14,** 32 (2009).
[5] Sukhdev, P. *Corporation 2020: Transforming Business for Tomorrow's World* (Island, in the press).
[6] Geanakoplos, J. *Nature* **457,** 963 (2009).

FOODCOLLECTION/SUPERSTOCK

Local produce is widely praised, but some argue that a reliance on it endangers food security.

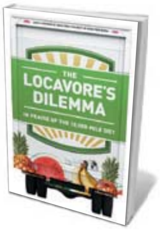
FOOD SECURITY

Eating globally

Tom MacMillan gets a taste of the argument against consuming only locally grown food.

For all the fanfare about local food, you might think that we eat a lot of it. Yet in the United Kingdom and North America, almost everything people eat comes from far away, shipped from distribution centres and delivered by truck. Only a tiny fraction takes a short cut. So, although about one-third of UK shoppers say that they buy local food, the market share is nearer 2–3%.

In *The Locavore's Dilemma*, geographer Pierre Desrochers and economist Hiroko Shimizu suggest that even that is too much. They say that it is ignorant to want shorter supply chains and dangerous to achieve them, whether in the developed or developing worlds. "The road to agricultural, economic and environmental hell," they argue, is "paved with allegedly fresher and more nutritious local meals". With this spirited polemic they want to nip the 'locavore' trend in the bud.



The Locavore's Dilemma: In Praise of the 10,000-Mile Diet
PIERRE DESROCHERS AND HIROKO SHIMIZU
Public Affairs: 2012. 304 pp. $26.99, £18.99

Desrochers and Shimizu argue that encouraging localized supply, and thus diversified farming, strikes at the essence of agricultural development and socioeconomic progress. Hefting food over long distances allows regions to play to their strengths, unlocking productive efficiencies that release people from farm work. This has brought social benefits by letting people engage in other activities, such as medicine and the arts. Against this backdrop locavore logic looks, the authors say, too foodie, protectionist and romantic.

The foodie fallacy is to assume that the answers to food-related problems must lie in the system. Farmers' markets and small grocery shops may enliven our gastronomic lives but, Desrochers and Shimizu remind us, food businesses don't have a monopoly on social capital. Spending less money and time on shopping and cooking leaves more for things such as community volunteering.

Local protectionism is a misguided way to achieve food security, they argue. The monocultures that make up the modern food system distribute risk across regions, and the associated division of labour has delivered financial means of risk-management, such as insurance and futures markets. By contrast, attempts at national self-sufficiency or autarky have fuelled imperialist expansion, whether in ancient Athens or twentieth-century Japan, as rulers have had to push their borders outwards to realize their ambitions.

To Desrochers and Shimizu, locavores are romantics who pine for a fictional yesteryear of 'natural' food and rustic idylls, whereas in fact, they say, shortening supply chains can push up costs, increase poverty and harm the environment. "If our agricultural past was so great," they ask, "why were modern animal and plant breeds, long distance trade in food, and modern production and processing technologies developed in the first place?"

The book's strength lies in the cheerful ruthlessness with which the authors challenge sloppy thinking, special pleading and the lazy logic that assumes that 'local' must be 'best'. Many of its weak points are symptomatic of the genre: its critical gaze points one way only, so the authors indulge in their own share of caricature, selective evidence and overstatement.

The biggest failure is that the argument hinges on an economic history that gives the free market credit for

➲ NATURE.COM
For an interview with locavore chef Alice Waters, see:
go.nature.com/anrdri

every success but blames all problems on political meddling. Given that state intervention has produced notable successes, such as social programmes to reduce hunger, this is simplistic.

The effect is that the authors have little constructive to say about the role of politics in a world in which it inevitably mixes with markets. They fail to ask key questions. For instance, how much has public investment in transport infrastructure and agricultural research and development shaped the marketplace? And what if, rather than being ignorant of the thinking that an ever more specialized division of labour will yield ever greater health, wealth and happiness, locavores are actually challenging it?

For example, Desrochers and Shimizu celebrate the specialization in the food industry that has given us artificial sweeteners to fight type 2 diabetes. But that specialization has also given us abundant empty calories and poverty-wage work, which contribute to the incidence of diet-related diseases. Local food won't solve public-health problems, true, but the authors' critique leaves us no wiser or fitter. If, as they say, "the essence of progress is to create less significant problems than those that existed before", should we just be thankful that we're fat rather than hungry?

The authors' confidence that the system works sits oddly against evidence that above a certain point, growth in gross domestic product is not correlated with improved well-being. At the core of progressive locavore thinking are efforts to address this by questioning the association between material consumption and prosperity, pushing use of renewable resources and reducing economic inequalities.

By hanging their argument on the advantages that we enjoy over our ancestors, Desrochers and Shimizu give us little more than an entertaining defence of business as usual. The UK government's unlocavorish Foresight unit, which advises on how to future-proof policy decisions, found last year that "nothing less is required than a redesign of the whole food system to bring sustainability to the fore". Desrochers and Shimizu's prescription not to mess with the market seems a missed opportunity to say something altogether more imaginative and more useful. Locavores don't have a blueprint, but we should welcome the ingenuity and challenge that they bring to this urgent redesign. ■

Tom MacMillan *is director of innovation at the Soil Association in Bristol, a UK charity that campaigns for planet-friendly food and farming.*
e-mail: tmacmillan@soilassociation.org

F. COFFRINI/AFP/GETTY

Peter Piot co-discovered the Ebola virus and helped to coordinate the global response to HIV and AIDS.

VIROLOGY

The battle inside

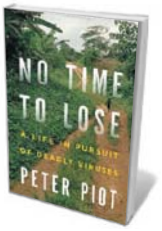
José Esparza enjoys the memoir of a long-term veteran of the virus wars.

In 1933, Nobel-prizewinning physician Charles Nicolle said that infectious diseases "carry the traits of life that seeks to perpetuate itself, evolving and trying to achieve equilibrium". But this evolution has a high price for humans. The war between human and microbe is epic and ongoing.

In *No Time to Lose*, Peter Piot, director of the London School of Hygiene and Tropical Medicine, offers chronicles of two battles from that war: his front-line fights against the Ebola virus, which can trigger a highly lethal haemorrhagic fever, and HIV. The book does not pretend to be a history of those viruses, or a technical manual on infectious diseases generally. It is a memoir — although intertwined with epidemiology, science and politics — and, as such, it is Piot's prerogative to remember and to recognize what he chooses.

We witness Piot's evolution over 35 years, from idealistic young medical scientist in Belgium to skilful United Nations politician and diplomat in Geneva, Switzerland, as director of the Joint United Nations Programme on HIV/AIDS (UNAIDS). Piot is not always diplomatic: he paints a warts-and-all portrait of how science is done and public health protected. And, like many good storytellers, he identifies the good guys and the villains in the threads of his narrative.

Piot's first African adventure was in Zaire, now the Democratic Republic of Congo, in 1976. He was chasing an unusual epidemic caused, he and his colleagues learned, by a previously undiscovered pathogen that came to be known as the Ebola virus. As Piot works towards an understanding of Ebola haemorrhagic fever, the story becomes the stuff of high drama: the writing is so vivid that I felt as if I were beside Piot in the Congolese jungle.



No Time to Lose: A Life in Pursuit of Deadly Viruses
PETER PIOT
Norton: 2012. 304 pp. $28.95, £17.99

The epidemic Piot witnessed was fast and furious, killing 431 people in Zaire and Sudan in the last four months of 1976. As it raged, Piot began to absorb the realities of research: the tensions between competition and collaboration and the need for priority recognition of scientific discoveries. He also started to learn how to communicate with affected populations, including Belgian nuns in the small village of Yambuku, Zaire, four of whom succumbed to Ebola. Rather than just studying it as a pathological phenomenon, Piot probed the epidemic's human dimension — an essential component of modern epidemiology.

During the epidemic, Piot collaborated and competed with several US scientists. These encounters led him to study sexually transmitted infections with epidemiologist King Holmes in Seattle, Washington. ▶

▸ He was at the University of Washington for a little more than a year, but this period was a turning point for Piot, preparing him for his next challenge: the emerging AIDS epidemic. It was to absorb the next 30 years of his life.

Now pursuing HIV, Piot returned to Kinshasa. In 1984, he and his collaborators established Project SIDA, which produced most of the early information on AIDS in Africa. The project was led at first by the epidemiologist Jonathan Mann; in 1986, Mann became the first director of the Global Programme on AIDS of the World Health Organization (WHO).

Piot details the personal differences and changing focus that led to the dissolution of the Global Programme and the launch of UNAIDS. Piot served as its first director from 1996 until 2008 — a period that makes up the bulk of the book. A more definitive overview of these years appears in *AIDS at 30* (Potomac, 2012) by Victoria Harden, a historian at the US National Institutes of Health.

Piot resolved that on the research front, UNAIDS would focus on epidemiology. But it also ran many other activities, particularly coordination of the country-level response to AIDS. Piot's main focus was advocacy, community mobilization, political sensitization and fund-raising, and he found success. I am disappointed, however, that as a medical scientist, he does not use his book to discuss the enormous research effort behind the antiretroviral drugs that significantly improved the prognosis of people living with HIV. Nor does he discuss the other biomedical efforts, including vaccines, which I believe will have a key role in stopping the epidemic.

"Despite the efforts of virus hunters, neither Ebola nor HIV is under control."

Despite the efforts of virus hunters, neither Ebola nor HIV is under control. These viruses continue to strive for the equilibrium suggested by Nicolle. By May 2011, 28 outbreaks of viruses in the Ebola family had occurred in 11 countries, with a total of 2,288 human cases. And by the end of 2010, an estimated 34 million people worldwide were living with HIV.

This book is not the story of two diseases. Rather, it is a fascinating account of the complex behavioural responses that epidemics trigger among their human hosts. ■

José Esparza *is senior adviser on vaccines for the Bill & Melinda Gates Foundation in Seattle, Washington.*
e-mail: jose.esparza@gatesfoundation.org

PSYCHOLOGY

Markets in mind

Investment bankers are addicts on a steroid roller coaster, finds **Richard Lea**.

René Descartes may have sparked the Enlightenment when he proposed that thought is the basis for existence, but Cartesian mind–body dualism has fallen out of favour in philosophy. According to neuroscientist John Coates, however, there is one domain in which the idea of a mind driven by pure rationality persists: economics.

In *The Hour Between Dog and Wolf*, Coates tests this to breaking point, with an area of economics where rationality does not rule. Using physiology and neuroscience, and grounded by 12 years working in New York's financial district, Coates paints a vivid picture of stockbrokers as thrill junkies, surfing waves of boom and bust on steroid hormones.

Coates suggests ways to calm those waves, but his prescription doesn't go far enough. He focuses on strategies for controlling the testosterone highs of the "Masters of the Universe" — as Tom Wolfe styled them in *The Bonfire of the Vanities* (Farrar, Straus and Giroux, 1987) — instead of restructuring a financial system that currently "balances precariously on the mental health of these risk takers".

Coates begins with a vibrant portrait of a Wall Street investment bank as the markets prepare for an interest-rate announcement. He conjures up the excitement of the trading floor, a "parabolic reflector" gathering information and registering early signals. Traders pick up on this information through a hunch or gut feeling, and act on it fast.

Alongside some novelistic vignettes — the head of department surveying the floor like a hound on the scent — Coates describes the neurological and physiological changes that bankers experience. For example, when stockbrokers hear a rumour that interest rates will rise, their brains put them into high alert: they "hear the faintest sound, notice the slightest movement". Their metabolisms accelerate, breathing quickens and blood shunts to major muscle groups. Their bodies are flooded with adrenaline and testosterone.

Turning to the dialogue between brain and body, Coates says that physiology is key in decision-making. Decisions imply action, so "our thoughts come freighted with physical implications". Traders, like soldiers, make snap decisions with much at stake, so they must listen to the signals from their bodies.

Coates presents his own research from a London firm. He found a link between traders' morning levels of testosterone and afternoon profits, as well as a substantial increase in cortisol during volatile markets. He saw this chemistry in action on Wall Street: traders on a roll, in clubs and prowling for sex, and the men's toilets exuding a slaughterhouse stench as the market tumbled.

The Hour Between Dog and Wolf: Risk-taking, Gut Feelings and the Biology of Boom and Bust
JOHN COATES
Fourth Estate/Penguin: 2012. 288/352 pp. £20/$27.95

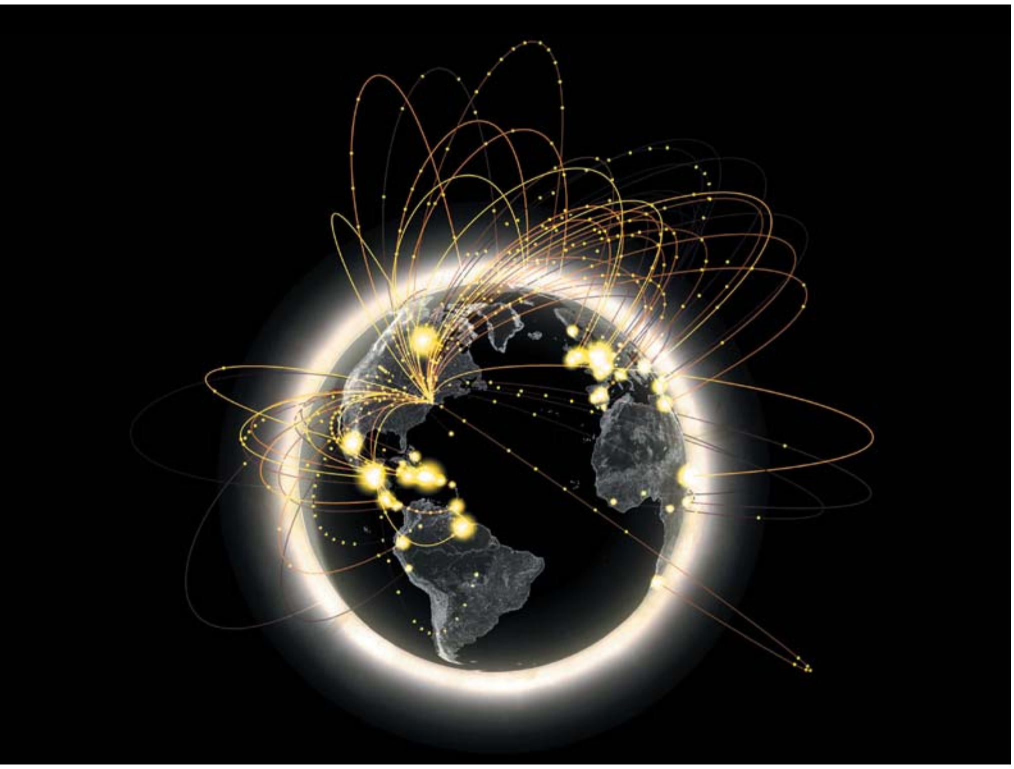
When testosterone levels rise during a bull market, with successes following each other so rapidly that there is no time for hormone levels to return to normal, stockbrokers can fall prey to the "irrational exuberance" that powers a bubble. During a declining 'bear' market, sustained levels of cortisol can fuel the panic before a crash and even, Coates suggests, contribute to hypertension and type 2 diabetes in individuals. A system that evolved to respond to physical threats over seconds or hours — the leopard in the forest, the rival in the group — is unable to cope with threats that evolve over weeks or months.

The drawbacks of Coates's familiarity with the trading floor become plain when he proposes fixes. Perhaps traders will be less stressed if they practise yoga or take up a wind instrument as he suggests, but his broader solutions — such as restraint from middle management, increased gender equality and the sinister idea of sending traders home on the results of a morning blood test — seem inadequate as counters to the powerful physiological influences on decision-making.

Coates's other solutions need some rethinking too. His idea of offering traders rewards on the basis of long-term results should be extended to managers and companies. Instead of adding "audio price feeds" to the trading floor to decrease traders' reaction times, a financial-transaction tax should be introduced to reduce market volatility.

If market swings are as driven by hormones as Coates suggests, it is not enough for politicians to step in when stress in the financial world has become pathological. They must decouple irrational exuberance on Wall Street from misery on the high street. Readers reeling from the effects of the most recent financial Armageddon may feel that the Universe needs some new masters. ■

Richard Lea *is a journalist on the books desk at* The Guardian *in London.*
e-mail: richard.lea@guardian.co.uk



A. KOBLIN

The *New York Talk Exchange* project visualizes the global flow of telephone and Internet data.

Q&A Aaron Koblin

The data visualizer

Aaron Koblin, head of the Data Arts Team in Google's Creative Lab, uses data visualization and crowdsourcing to reveal the changing relationship between people and technology. As he presents his work at the Eyeo Festival of digital creativity and prepares to release a collaboration with Google, London's Tate Modern and artist Chris Milk, he talks about the beauty of big data.

NICK HIGGINS

Eyeo Festival 2012
5–8 June, Minneapolis, Minnesota
http://eyeofestival.com

What place does visualization have in science?
Scientists have an amazing quantity of data, but their presentation can drain the meaning and power from it. They sometimes get stuck on the data, and don't work up to information, knowledge and wisdom in their communications. Human visual perception has been tuned by evolution. Good visualization allows an illuminating god's-eye view of the data. Medical imaging is the most obvious example, but you can see the value of visualization everywhere from statistics to physics and chemistry.

What does leading the Data Arts Team entail?
We create demos and hacks using the latest web technologies. We have a gallery, ChromeExperiments.com. My favourite experiment is *The Wilderness Downtown*, a 2010 music video for the band Arcade Fire, which asks you to enter the address where you grew up, then uses Street View and satellite imagery from Google Maps to zoom into that location as if the narrative were unfolding in your town.

What else have you done with data?
For *Flight Patterns* (2005), I processed data from the US Federal Aviation Administration to create an animation of air-traffic density and movement, showing the beautiful ebb and flow of people moving through pathways in the sky. With the SENSEable City Laboratory at the Massachusetts Institute of Technology in Cambridge, I did a project using real-time long-distance telephone and Internet data from telecommunications company AT&T, called *New York Talk Exchange* (2008). We broke the data down by borough and compared them with ethnographic and demographic data to see how different groups communicate around the world.

Has your visualization work brought other opportunities?
At the Center for Embedded Networked Sensing at the University of California, Los Angeles, I worked for Mark Hansen, a statistician and data artist, writing visualization software for three-dimensional laser scanners that measured environmental conditions, such as sunlight exposure through rainforest canopies and erosion of hillsides. The visuals were so beautiful that I set up an installation in the centre's lobby to show how people moved through the building. A director saw it and asked me to help to make a "music video without video" for the band Radiohead, using laser scanners.

Some of your music videos look like they were filmed through a microscope. How were they made?
In the video for Interpol's 2008 song *Rest My Chemistry*, I used principles of repulsion and attraction to animate vertex-only models of human bodies. It was programmed using algorithms from physics demonstrations online. Much of the same code was used on another (unofficial) interactive video for Dopplereffekt's 2007 song *Hyperelliptic Surfaces*, programmed mostly by Aaron Myers. It was meant to have the look of particles moving under a microscope, using depth effects to simulate the narrow frame of focus in micro-optics. You can see all the forces playing out on the particles, creating complex patterns of motion. We didn't start with real scientific data sets. We created them ourselves.

You've also worked with crowdsourcing?
I thought that Amazon's Mechanical Turk platform, which lets people outsource any computer-based task to people all over the world, was one of the most powerful and disturbing ideas ever to materialize as a product. I was troubled by how the workers often have no idea who they are working for or what they are doing. But I was curious as to how it might be used for art and cultural investigations.

How did you begin to hire strangers online?
In Antoine de Saint-Exupéry's *The Little Prince* (1943), the narrator is asked repeatedly to draw sheep, and learns that imagination is more important than precision. I paid online workers 2 cents each to draw one of 10,000 digital sheep, to create *The Sheep Market* (2006). It was amazing to see them coming in. I have also paid more than 2,000 people to sing single notes from *Daisy Bell*, the song used at Bell Laboratories for the first musical speech synthesis. I resynthesized the song in a distributed chorus that came out sounding like a dystopian horde of gremlins. I called it *Bicycle Built for Two Thousand* (2009).

Is originality possible in data-driven art?
All art is a representation of some influence. With data-driven projects I consider myself more of a 'first viewer' than 'The Artist'. The data tell a story, and I craft and reveal it. I'm not concerned with where that leaves me as an artist. It is more about sharing something that we can all reflect on. ■

INTERVIEW BY JASCHA HOFFMAN

Correspondence

Value of submerged early human sites

Your articles on human dispersal in the late Pleistocene epoch (*Nature* **485,** 23; 2012) overlook the significance of now-submerged archaeological sites on the continental shelf during this period (126,000–11,000 years ago). It is wrong to assume that these were completely destroyed by the sea and that the interpretation of human movements must rely on proxy data, such as DNA or evidence from islands.

More than 3,000 prehistoric sites on the seabed have been found and mapped, and in some cases excavated. They range in age from 500,000 to 5,000 years old, and at locations from the present-day shoreline out to a depth of 130 metres. These sites were extensive, often located on key travel routes and more attractive than arid hinterlands to human settlers.

Marine archaeologists have recovered in-context stone artefacts, animal remains and human fossils from such sites. Some materials, including food remains, organics, bone, DNA and plants, are better preserved underwater than on land.

Questions of early human dispersal will not be resolved until continental shelves are fully investigated — spurred by advances in modern oceanographic technology (see http://splashcos.org).

Nicholas Flemming *National Oceanography Centre, Southampton, UK.*
n.flemming@sheetsheath.co.uk
Geoffrey N. Bailey *University of York, UK.*
Dimitris Sakellariou *Hellenic Centre for Marine Research, Athens, Greece.*

Two faces of marine ecology research

The ecology of animal movement is one field that would benefit from sound evaluation of the risks, benefits and ethics of its important research findings (*Nature* **484,** 415 and *Nature* **484,** 432–434; 2012).

Scientists can now track the complex horizontal and vertical movements of a wide range of marine species, including tuna, sharks and turtles. These results reveal biodiversity hotspots and inform conservation policies by providing insight into animal behaviour and ecology. However, they also guide fishing operations towards resource-rich locations — putting further strain on both target and by-catch species.

Too many species face severe stock depletion because of intense fishing, pollution and other anthropogenic pressures. The detrimental implications of marine ecological research results must be acknowledged.

Juerg Brunnschweiler *Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland.*
juerg@glueckklich.net

Villages project responds to criticism

Some of your criticisms of child-mortality figures from the Millennium Villages project in Africa are unjustified (*Nature* **485,** 147; 2012).

You question the methods we used to generate mortality estimates, which were based on mothers' recall of the numbers of births and deaths of their children (P. M. Pronyk *et al. Lancet* http://dx.doi.org/10.1016/s0140-6736(12)60207-4; 2012). But this is standard practice, and was approved by the paper's reviewers. To avoid any potential systematic bias, we used the same technique and reporting period for the Millennium and comparison villages.

No "alarm bells" sounded over the mortality rate in the comparison villages. These were closely matched to the Millennium sites, chosen because they were poor, often remote, hunger hotspots. We therefore had no expectation that mortality rates at any of these sites would track national trends.

You wrongly suggest that the costs of the project were not properly reported. Our paper presents costs by site, sector and funding source — providing much more information than in most comparable studies. Detailing intervention costs and spending by partners is a core purpose of the project, executed with scrupulous care.

Our results on the decline in child mortality in the Millennium Villages are statistically significant over time and, in relation to the comparison villages, follow protocol. As you point out, the comparisons made with national trends were not tested for statistical significance, which is why they have been withdrawn (P. Pronyk *Lancet* http://dx.doi.org/10.1016/S0140-6736(12)60824-1; 2012).

Since then, the project has been piloting techniques for assessing changes in the under-5 mortality rate in the Millennium Villages using trained community-health workers to collect detailed, accurate birth and death data. Leading experts on public health will be invited to form an independent committee to review and help strengthen the collection and processing of these data.

Paul Pronyk *Center for Global Health and Economic Development, Earth Institute, Columbia University, New York, USA.*
ppronyk@ei.columbia.edu

Pitfalls of Romania's ethics council

Changes to the laws of the Romanian government's National Ethics Council, created in 2004 to control misconduct and plagiarism in scientific research (*Nature* **485,** 289; 2012), have been hailed as a boost to the country's research reforms. But preventable pitfalls threaten the council's prospects for success.

The council's 11 members have impeccable credentials and have issued bold pronouncements. However, the council has only advisory status and no legal powers. Rather than seeking cross-party consensus on membership, the education minister retains the power to appoint council members, who are therefore vulnerable to accusations of political bias.

The council's powers are further restricted because it has no access to anti-plagiarism software or to comprehensive databases. Members must judge the cases brought to them, some of which could be politically motivated and might affect public perception of the council.

As you point out, the government's new anti-plagiarism legislation rules that any academic found guilty of misconduct will lose his or her job. Such sanctions can be retroactive, affecting scientists appointed before the new regulations came into force — a questionable strategy prohibited by most constitutions, including Romania's.

It remains to be seen whether these factors will prevent the ethics council from acting efficiently, asserting its independence and gaining the role it deserves.

Octavian Voiculescu *University of Cambridge, UK.*
ogv20@cam.ac.uk

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines at **http://go.nature.com/cmchno**. Alternatively, readers can comment online on anything published in *Nature*: **www.nature.com/nature**.

AFTER EXPERIMENT SEVEN

Parallel processing.

BY MICHAEL W. LUCHT

'Experiment 6. Apparatus: Smith & Wesson Model 13.' Having written thus, Professor Hillabin began searching for the gun among the piles of books, papers and assignments. His students believed him to be disorganized — Hah! — little realizing that the mess afforded perfect concealment. Besides the gun, it presently occluded vials of cyanide, assorted knives, a parrot and even an electric chair. If the vice-chancellor found out, she would throw a fit — especially about the parrot. Faculty policy strictly prohibited pets in academic offices.

Eventually, Hillabin unearthed the weapon from beneath a pile of decaying term papers. After meticulously filling the chambers with bullets, he faced a problem. His methodology required him to test it, but a gunshot might be noticed, even in the philosophy department. He glanced at his watch; it was past 11 p.m.. Deciding to risk it, he pointed the gun at the wall, aiming between a soaring stack of old journals and an even taller tower of unmarked exam papers.

Missing the gap, the thunderous bang was accompanied by a cloud of confetti. Oh well, only exam papers …

Hillabin had barely replaced the spent shell when Professor Forthington stormed in. "Hillabin! What the hell are you up to now? You've ruined my desk!"

Hillabin reflected on his bad luck that his neighbour, famous for heading home on the dot ever since the great philosophers' strike of '00, had selected this of all nights to work late.

He carefully recorded 'Apparatus working' in his notebook before glancing up. "Sorry old *chum*. I'm conducting a series of quantum suicide experiments."

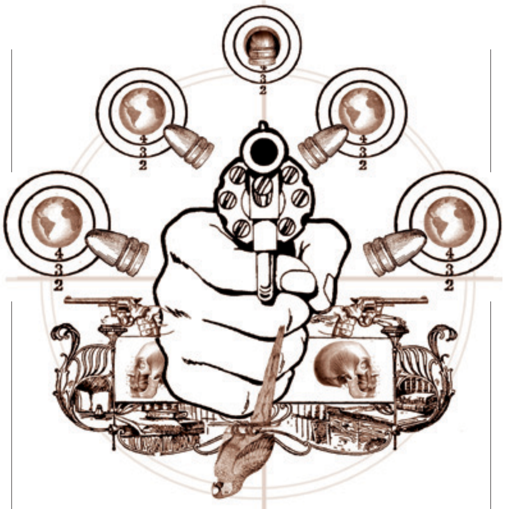
"Sounds positively ghastly! Why don't you do some real work on Kierkegaard's spiritual angst?"

Hillabin sighed. It was hell for an experimental metaphysicist to have an existentialist as a neighbour. "I am probing," he said, trying not to sound excessively pompous, "the nature of reality!"

"Yeah, I'm sure people will find that enormously helpful when dealing with the general meaninglessness of life." Bug-eyed, Forthington stared at the gun. "So, you're trying to kill yourself then?"

"Quantum *suicide* — duh!" Hillabin replied, using the vernacular of his students.

Forthington raised an eyebrow. "May I watch?"



JACEY

"Why not?" Hillabin shrugged, placing the barrel in his mouth.

"Wait!"

"What?" Hillabin asked, gagging.

"Any last words?"

Rather than dignifying the question with an answer, Hillabin pulled the trigger. The hammer clicked — and that was about it.

"See?" Hillabin asked triumphantly.

"I don't know." Forthington pinched his lower lip. "Could be a fluke."

"That's the whole point!" However, to strengthen the result, Hillabin pulled the trigger five more times, producing five more clicks. He turned the gun around. When he again aimed at the wall, the gun emptied without a hitch.

Forthington's initial reaction was to scream: "My poor office!" followed in rapid succession by: "That's amazing."

"It sure is," Hillabin agreed, recording 'Success' in his notebook.

Flipping to the next page, he wrote: 'Experiment 7. Apparatus: Cyanide.' Hillabin uncorked one of the small vials, poured a drop into a tiny bowl and offered it to his parrot. A pity, really, but Hillabin had to eliminate the possibility that the chemistry department had appeased him with some harmless liquid. It had been known to happen.

But not this time. The parrot, stressed by the loud bangs, badly needed a drink. Ten seconds later, it keeled over.

"It's dead," Hillabin muttered.

"No, it's not!" Forthington retorted instantly.

As much as he enjoyed gratuitous Monty Python references, Hillabin was beginning to suspect that his colleague was not treating his work with the gravity it deserved. As there was not much to be done about that right now, he downed the poison in one gulp and leaned back on his chair. After five minutes, he wrote 'Success. Probably survived due to some peculiar immunity. Must remember to ask a biologist.'

"So, you really can't commit suicide?" Forthington asked.

"Not just me. Nobody!"

"Really? Wow!" For once, Forthington looked impressed. "You know, this is so remarkable. Who would have thought? I almost wish that I could give it a try."

Hillabin reloaded the revolver and held it out. "Be my guest."

"Oh, I don't know."

Hillabin once more tried shooting himself, with the same result as before. "This is exactly what you will observe. I guarantee it." He again held out the gun.

"Well, in that case …" With his arm shaking, Forthington finally took the weapon and aimed it at his head. "Should I really?" he asked, sweating.

"Don't be such a logical positivist!"

The shot killed Forthington instantly, making the huge mess in Hillabin's office significantly less palatable.

Despite this, Hillabin was pleased with the progress achieved that day. He had successfully established that the many-worlds interpretation of quantum mechanics was correct. Each time he had tried killing himself, the universe forked into many. In the bulk of them he died. However, as it stood to reason that he could not observe universes in which he did not exist, his consciousness — his point of view, if you will — would always find itself in a universe in which he had survived, no matter how unlikely the odds.

Looking down on Forthington's body, Professor Hillabin felt an unanticipated tinge of guilt. Should he have told him that in the vast majority of universes he would die? Surely, even he must have known?

Then again, did it matter? In at least one parallel universe, Forthington was no doubt already annoying another version of Hillabin by going on and on about how amazing it was that he still lived.

But not in this universe, thank God. Shrugging, Hillabin dialled security. ■

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

Michael W. Lucht *lives in Singapore, where he hopes either to create artificial life or to publish a novel — whichever is easier.*

TECHNOLOGY FEATURE

BUILDING BETTER BIOBANKS

High-quality, data-rich samples are essential for future research. But obtaining and storing these samples is not as straightforward as many researchers think.

L. RUBENSTEIN/BROAD INST.

Larger biobanks have invested in automated storage and retrieval systems to track samples and ensure that they are maintained at a constant temperature.

BY MONYA BAKER

Across the world, freezers and cabinet shelves are full of human samples.

Biobanks — collections of biological material set aside for research — vary tremendously in size, scope and focus. Samples can be collected from the general population, from patients who have had surgery or a biopsy and from people who have recently died. Some collections date back decades. The Aboriginal genome, for instance, was sequenced from a lock of hair originally given to British ethnologist Alfred Cort Haddon in the 1920s; he criss-crossed the world gathering samples that are now housed at the University of Cambridge, UK. Most collections contain dried or frozen blood, but tissues such as eye, brain and nail are also held. Some biobanks address different questions from others: a population-based biobank that collects dried blood and health data may be used to determine the genetic risk factors for breast cancer, whereas a disease biobank that collects tumour samples might be used to reveal different molecular forms of breast cancer.

The number of tissue samples in US banks alone was estimated at more than 300 million at the turn of the century and is increasing by 20 million a year, according to a report[1] from the research organization RAND Corporation in Santa Monica, California. Those numbers are probably an underestimate, says Allison Hubel, director of the Biopreservation Core Resource at the University of Minnesota in Minneapolis.

But many scientists still say they cannot obtain enough samples. A 2011 survey[2] of more than 700 cancer researchers found that 47% had trouble finding samples of sufficient quality. Because of this, 81% have reported limiting the scope of their work, and 60% said they question the findings of their studies.

Whereas researchers would once have inspected biological specimens under a microscope or measured only a handful of chemical constituents, or analytes, now they want to profile hundreds of molecules, including DNA, RNA, proteins and metabolites. The ▶

INDIVUMED

▶ popularity of genome-wide association studies, in which researchers scan the genome to look for genetic markers, has trained scientists to go on statistical hunts that require both more quantitative measurements and greater numbers of samples. "The manner in which biomedical researchers use biospecimens has changed substantially over the past 20 years," says Stephen Hewitt, a clinical investigator at the National Cancer Institute in Bethesda, Maryland, and an expert on sample quality. "Our knowledge of the factors that impact a biospecimen has not kept up, nor has the education of the users about how fragile a biospecimen is."

IN THE COLD

"If you don't treat the sample properly, it can limit what you can do," says Kristin Ardlie, director of the Biological Samples Platform at the Broad Institute in Cambridge, Massachusetts. She recalls a project to isolate RNA from placenta samples, which are full of RNA-degrading enzymes. After several tries and no success, a bit of detective work revealed that a collaborator had put the samples in a −20 °C freezer to begin with, and only moved them to a −80 °C freezer several hours later.

"Researchers think 'freezing is freezing'," says Ardlie, but a typical freezer is not cold enough to stop degradative enzymes. Except for DNA, few biomolecules are preserved well at −20 °C. Most samples can be stored at −80 °C, but certain specimens, such as live cells, need to be kept at temperatures close to −200 °C, at which point enzymes are thought not to be able to function at all[3].

Worse than having nothing to analyse are analytes that change in unpredictable ways. One study[4] showed that the concentration of two cancer biomarkers seemed to increase by around 15% from the time that the serum samples were collected and frozen to when they were thawed and measured again about 10 years later. In another experiment[5], designed to simulate long-term freezing, researchers examined how several cancer biomarkers changed in serum samples that were frozen and then thawed. Some protein biomarkers seemed to be stable for decades even with multiple freeze–thaw cycles. However, vascular endothelial growth factor — an extensively studied biomarker implicated in diabetes, arthritis and cancer — was so unstable that the authors recommended that it should never be measured in samples that have been frozen.

N.G. JOHNSON

Reports have probably underestimated the true number of samples in US biobanks.

Allison Hubel

Liquid-nitrogen freezing stops samples degrading.

Not all biobanks document whether a sample has been thawed for analysis and then restocked, nor do they monitor freezer temperatures, says Daniel Simeon-Dubach, a biobanking consultant based in Switzerland. Even short-term fluctuations in temperature can allow sample-damaging ice crystals to form, but Simeon-Dubach says he has seen researchers hold freezer doors open for minutes at a time to show off their specimens. "I think 'what are you doing? Show me a picture!'."

ON THE SHELF

Many of the larger biobanks are buying sophisticated freezers to maintain a constant temperature. Systems from companies such as Hamilton Storage Technologies in Hopkinton, Massachusetts, start at around US$1 million and can hold between 250,000 and 10 million samples. Rather than opening freezer doors, researchers place sample tubes in a hatch, and a mechanical arm then moves them to interior shelves. Researchers can even use laboratory information-management systems to search for appropriate samples for a particular study, such as those from donors of a particular age or weight, and then transmit their request to be retrieved. The samples are deposited in a delivery hatch and an e-mail is sent when they are ready to be picked up. The −80 °C freezer also records how many times each sample is removed from frozen storage and for how long. Other companies, such as Freezerworks in Mountlake Terrace, Washington and Brooks Automation in Chelmsford, Massachusetts, also sell automated freezers for storing and tracking samples.

Even without such sophisticated equipment, freezer storage can be expensive. A typical epidemiological study might have 100,000 samples from 10,000 patients that would fill five freezers, each of which costs $6,000 a year to maintain properly, says Jim Vaught, deputy director of the National Cancer Institute's Office of Biorepositories and Biospecimen Research (OBBR). And although freezing is considered the best way to preserve biomolecules and live cells, it can distort the appearance of tissues.

To cut storage costs, most researchers study morphology by relying on a preservation technique that harks back more than a century. Tissue taken from a patient is soaked in the preservative formalin, and pieces of the 'fixed' tissue are then embedded in blocks of paraffin. The Joint Pathology Center in Silver Spring, Maryland, has some 28 million of these blocks, dating back to the First World War. The blocks allow a thin slice of tissue to be taken and stained for microscope slides, but biomolecules are not preserved as effectively. "The tissue actually drowns in fixative," explains Hewitt. Hypoxia in the dying cells degrades RNA and alters proteins; formalin crosslinks protein and DNA into complexes, and causes nicks in RNA and DNA. When researchers go to recover biomolecules, removing the paraffin can cause more damage.

Although DNA and RNA have been extracted from paraffin-embedded samples, the quality varies and analysis is difficult. Mike Hogan is vice-president of research at IntegenX in Pleasanton, California, which sells products for storing DNA and RNA at room temperature. He believes that formalin fixation can be modified to preserve biomolecules. The main causes of biomolecular degradation are not due to the formalin directly but to hydrolysis and oxidation, he says. Freezing works because the chemical reactions driving degradation occur more slowly at low temperatures. Scientists at IntegenX and at the University of North Carolina, Chapel Hill, are working on techniques to slow hydrolysis and oxidation by removing water and reactive oxygen-containing molecules. If this technique works, it would allow researchers to study biomolecules but still maintain the morphology standards and staining protocols developed over decades of formalin fixation.

Other approaches focus on removing formalin from the process. In 2009, Qiagen, based in Hilden, Germany, released a product called PAXgene Tissue, which uses a proprietary, alcohol-based fixative to preserve biomolecules and to allow tissue specimens to be embedded in paraffin. The tissue can be stored for up to seven days at room temperature, four weeks at 4 °C and months at −20 °C without compromising morphology or biomolecules, says Daniel Groelz, a senior scientist at Qiagen. Researchers there are working on ways in which pathologists can analyse more kinds of PAXgene-preserved histological samples. "There is a huge amount of specialized staining techniques that have been optimized for

VAN ANDEL RESEARCH INST.

Researchers are investigating the best way to analyse biomolecules in samples stored in paraffin blocks.

formalin," says Groelz. The company is working on adapting protocols for PAXgene, particularly antibody-based protocols to stain proteins, he explains.

This preservative method is starting to be used in place of deep freezing. A pilot project for one of the most ambitious tissue-collection studies, which aims to correlate gene expression and common genetic variation within dozens of tissue types, examined more than 20 tissue types preserved using four different methods. The Genotype-Tissue Expression (GTEx) programme, a collaborative effort involving several groups at the US National Institutes of Health, as well as academic institutions ultimately chose PAXgene. Jeffrey Struewing, a programme director for the US National Human Genome Research Institute in Bethesda, Maryland — who works on the project — explains that not only did the technique preserve RNA, the logistics necessary to ship ultracold samples could have hampered collection. Struewing says that it is too early to know how PAXgene will work over many years or for biomolecules such as proteins. "There is no preservation method that is going to work for every analyte in every sample."

QUALITY COLLECTION

Some of the most intractable difficulties occur before preservation begins, says Carolyn Compton, the first director of the OBBR and chief executive of the Critical Path Institute in Tucson, Arizona — a not-for-profit organization for improving drug development. "Biospecimens are parts of people's bodies that get removed from their setting. They are undergoing biological stresses they would never experience in your body." When cut off from a blood supply and exposed to abrupt changes in temperature, the cells' behaviour becomes hard to predict. Gene expression and protein phosphorylation fluctuate wildly and cellular self-destruct pathways may be activated. Researchers must ask themselves whether analyses of samples reflect the biology of the patients they come from, says Compton. "You can have an absolutely perfect test but still get the wrong answer."

Even if tissue is preserved well, it may not tell the full biological story. "The problem is not just the post-mortem interval," says Hewitt. "What's difficult is the pre-mortem lack of vitality." Tissues collected from patients who have been on ventilators may not resemble those from healthy patients. "If you took a biopsy of muscles in my arm after I went rowing in the morning, my RNA profile will look a lot different than if I've been sleeping for a while," Hewitt says. "You've got to interpret your data within the limit of what they can tell you."

> "You can have an absolutely perfect test but still get the wrong answer."

Blood, urine and saliva samples from non-hospitalized volunteers can be collected during scheduled appointments. But solid-tissue samples are usually collected in hospitals as part of more urgent procedures. Medication, the anaesthesia regime and how blood is shunted from the tissue being removed all affect the sample. So does the length of time the sample stays at room temperature before it is frozen, the time and type of fixative, the rate at which it is frozen and the size and shape of the aliquots.

Medical staff will always be focused on the patient on the operating table, but a greater awareness of the impact that samples can have on medical research and patients' diagnoses is having an effect. At a conference organized by the OBBR this February, Gene Herbek, a pathologist at the Nebraska Methodist Hospital in Omaha, described working with surgical teams so that tissues reached pathologists' laboratories within an hour of excision.

Biotechnology company Indivumed in Hamburg, Germany, collects samples within ten minutes of excision by having designated nurses on surgical teams. These nurses prepare for surgery along with the rest of the team, receiving information about a patient's treatment and condition. Once the tissue is removed, it is taken into a room next to the surgical suite, where it is sectioned into pieces that are then fixed and frozen. "The solution is not technology; it is process," says Helge Bastian, managing director of the company.

"The rule of thumb for how long you have from taking a sample to starting to process it is 15 minutes," says Simeon-Dubach, adding that this is a very ambitious goal. Specifics vary by organ — gastrointestinal organs such as the stomach should be processed much faster.

Speed is also important for tissues collected post-mortem. Staff collecting tissues for GTEx are expected to be ready around the clock so that they can begin work as soon as the team collecting donated organs has finished. More than half of the tissue samples in the programme have been collected and fixed within six hours of death, says Struewing.

Tissues are shipped to the Broad Institute, also part of GTEx, for gene-expression profiling. Before beginning the profiling, Ardlie's team isolates the RNA and checks it for quantity and quality using a metric called the RNA integrity number; it is an imperfect measure, but excluding samples with low integrity numbers maintains consistency.

ASSESSING QUALITY

Researchers need better biomarkers of sample quality both to prevent expensive experiments on inappropriate material and to reduce artefacts, says Hewitt, who is working with Ardlie to find measures of RNA quality that work in paraffin-embedded tissue. Scott Jewell, director of the programme for biospecimen science at the Van Andel Research Institute in Grand Rapids, Michigan, is evaluating markers of oxygen deprivation and various types of cell death (autophagy, apoptosis and necrosis). Documenting how samples are collected and maintained is important, but may be insufficient, he says. Researchers need specific recommendations. "We want markers that can say, 'This is a bad sample. This is a good sample.'" Such a process is important not only for choosing samples to include in a particular study, but also for understanding how best to preserve them.

Many biobanking experts find that researchers give little thought to sample quality. An analysis of 125 biomarker discovery papers published in open-access journals between 2004 and 2009 found that more than half included no information about how specimens had been obtained, stored or processed[6]. Perhaps this is not surprising; biobanking practices have come under scrutiny only recently. The OBBR, established in 2005, released its first official set of best-practice guidelines in 2007, and last year released *Biospecimen Reporting for Improved Study Quality* to guide researchers on documenting how biospecimens are collected, processed and stored. The International Society of Biological and Environmental Repositories in Bethesda, Maryland, published its first edition of best practice in 2005, and a coding system — Standard PREanalytical Code — for describing what tissue had been collected and how in 2010. The European Union has funded a four-year programme called Standardisation and Improvement of Generic Pre-analytical Tools and Procedures for *In Vitro* Diagnostics, a multi-institution project coordinated by Qiagen with the aim of improving and standardizing sample handling for *in vitro* diagnostics. In addition, societies are advocating that journals request this information for peer-reviewed articles.

The College of American Pathologists, a professional society based in Northfield, Illinois, has developed an accreditation programme for biorepositories. It began accepting applications this year and has so far had a good response. Facilities receive a checklist of required practices, such as whether they have tracked if samples have been thawed and refrozen, and whether they have installed freezer alarms. If facilities meet the list of requirements, they

A kingdom's worth of samples and data

In the past few years, more than a half a million people in the United Kingdom have collectively peed into cups, spat into tubes and had needles stuck in their arms. They have spent hours having their weight, blood pressure, memory, lung volume and grip strength measured and recorded, and answering extensive questionnaires about their lifestyles.

Deciding what samples and information to collect was several years in the planning, says Paul Downey, director of operations, who joined the UK Biobank project in 2004, years before the first participants were enrolled. "Tracking 14 million small tubes of biological fluid from the point of creation through to storage for, potentially, 30 years is an ask," he says. Everything had to be completed within the bank's budget and in a way that was acceptable to volunteers. Initially, for example, some researchers hoped for fat biopsies, but that would have discouraged volunteers and increased the cost and complexity of collection.

The decision was made to collect blood, urine and saliva. Samples were collected using seven kinds of blood tubes, containing different mixtures of anticoagulants and preservatives to support anticipated assays. The team decided not to include additives in the saliva and urine samples because this might introduce variability or affect future analyses.

On a typical day, the project processed samples from between 700 and 1,000 volunteers. This took four or five laboratory personnel and more than a dozen robots, which together could do the work of about 50 technicians, says Downey. "It does look like a science-fiction set." One six-armed robot centrifuges blood samples, images the layers that form and then pipettes each layer into its own tube, which is bar-coded and prepared for storage. The robots are still busy processing samples for disease-specific biobanks and follow-up specimens from participants, says Downey.

Each donor's samples are split between two facilities, which serve as a back-up for each other. The first, which holds 9 million samples at −80 °C, is essentially a freezer the size of six double-decker buses, containing a 1.5-tonne robot. Researchers can request samples by bar-code, and the robot will pluck them from their racks and bring them to a delivery port, minimizing the exposure of other samples to warmer temperatures. Nineteen kilometres away, a manual back-up system cooled by liquid nitrogen holds 6 million samples at −196 °C — the temperature recommended for storing live cells. This facility contains 50 round, stainless-steel tanks, each the size of a small car. Eventually, says Downey, the set-up will let researchers compare how various biomolecules survive under the two conditions.

The UK Biobank stores millions of samples from UK citizens.

WELLCOME LIBRARY, LONDON

Steps have also been taken to preserve samples during analysis, says Downey. For example, research facilities for measuring DNA, RNA and certain proteins are available on-site, minimizing the need to ship precious samples to investigators. Eventually, the biobank itself may measure parameters that will probably be of interest to many researchers, including blood lipid levels. Informatics systems will allow data from the experiments to be shared; they will also combine requests for analyses into batches to minimize the number of times a sample is thawed and refrozen.

The project, which is nearly ten years in the making, has so far cost £85 million (US$134 million), and is still in its infancy. The UK Biobank began accepting research proposals for ways to study these samples and associated data in March and received some 150 "credible applications" within the first three weeks, says Downey. Scientists from universities and health-care companies in any country are eligible. Samples will grow more valuable as data accumulate. In ten years, an estimated 9,000 of the original donors will have developed Alzheimer's disease, 10,000 will have breast cancer and 28,000 will have died from heart disease. **M.B.**

can apply for accreditation and schedule an inspection. "I've been hearing for ten years that someone should step up and do this," says Nilsa Ramirez, director of the biopathology centre at Nationwide Children's Hospital in Columbus, Ohio, and co-chair of the accreditation working group. "I think it will allow investigators to have a sense that what they are dealing with is the highest possible quality."

One difficulty with these efforts is that published guidelines are generally based on researchers' impressions and experience, not dedicated experiments that test for the best ways to preserve samples, says Vaught, whose office is now awarding grants for assessing and developing storage technology, and is maintaining a hand-curated database of relevant peer-reviewed literature. Biobank professionals often develop their own practices after a few pilot studies but do not publish them, he says. "There are no international standards based on solid research."

Indeed, resources for research and facilities for preserving biospecimen quality are in short supply. "People want to allocate funds for the research project and analysis, not the infrastructure that supports it," says Compton. First-year start-up costs for a mere 50,000 samples will probably be between $3 million and $5 million, not including information systems. Ten-year operating costs[7] could be more than $10 million. Obtaining funds for ongoing expenses is also a challenge: academics are used to getting samples from their colleagues, rather than paying a repository for high-quality samples.

WELLCOME LIBRARY, LONDON

Bar-coded samples can be scanned and tracked from collection to analysis.

PRESERVING PATIENT DATA

Several organizations exist to help researchers access the samples needed for their studies. The Cooperative Human Tissue Network — a network of divisions across the United States initiated by the US National Cancer Institute to improve access to human tissue — asks researchers to collect tissue and fluids from routine surgery and autopsies, as does the National Disease Research Interchange in Philadelphia, Pennsylvania, which specializes in rarer specimens, such as eyes. Several governments have initiated large biorepository projects (see 'A kingdom's worth of samples and data'). The Biobanking and Biomolecular Resources Research Infrastructure (BBMRI), a network of European biobanks, is creating policies to allow researchers to share specimens and data. The Public Population Project in Genomics (P^3G) based in Quebec, Canada, offers not only open-source software for documenting some aspects of informed consent, sample collection and processing, but also a database of biorepositories and their collections. Future research questions will require larger numbers of samples for rigorous statistical analysis, says Isabel Fortier, director of research and development at P^3G and a researcher at McGill University Health Centre in Quebec. "There is no way to think that just one study, even a big study, will have enough samples," she says. "We need to give a second life to the data that we have already collected."

Ongoing health information about a donor is increasingly desired by researchers, along with information on the preservation of any particular class of biomolecule. This has already prompted considerable reanalysis of appropriate informed consent and data policies[8], as well as innovations in how data can be stored and mined. "It's really going to have to be the wave of the future for biobanking," says Jewell. "Without knowing how to manage the continuous flow of data, you'll be a static biobank. We want to be able to constantly update the clinical record."

> "People want to allocate funds for the research project and analysis, not the infrastructure that supports it."

The more information that is available about a specimen, the more valuable it becomes to other researchers. Scientists studying the effects of a particular gene on a cancer pathway could save years of effort and thousands of dollars if they have ready access to a collection of tumour samples with mutations of interest. David Cox, a senior vice-president at drug-company Pfizer and a member of the BBMRI's scientific advisory board, believes that the way to get the most out of biological specimens is not prospectively banking samples but finding ways to reuse samples that researchers have already collected for their own questions. "You can't store everything. This concept that you're going to get all the samples and store them and then decide what to do is too expensive and it's hard to maintain the quality." At the same time, he says, individual negotiations with every group that collects specimens is also inefficient. He envisions loosely coordinated 'centres of excellence', in which researchers store samples and track clinical information for their own research questions, but also agree to a common structure for sharing samples and maintaining their quality.

One problem is who pays for what, says Cox. "People are trying to make money off of these individual pieces instead of trying to get them all together." Government funding is tight; pharmaceutical companies are willing to fund studies that can lead to new products, and individuals are generally willing to donate specimens and data for the public good, but not for corporate profit. One idea is that research would be conducted for pharmaceutical companies within the biobanking infrastructure, but that companies would not retain the exclusive rights to the data; however, it is too early to say whether this would be viable. There needs to be a way to link infrastructure and information "in a precompetitive fashion, so we can understand the biology better, and we can make better medicines," says Cox. Perhaps the hardest problem of all will be establishing — and maintaining — investment. ■

Monya Baker *is technology editor for* Nature *and* Nature Methods.

1. Eiseman, E. & Haga, S. *Handbook of Human Tissue Sources* (RAND, 1999).
2. Massett, H. A. *et al. J. Natl Cancer Inst. Monogr.* **2011,** 8–15 (2011).
3. Hubel, A., Aksan, A., Skubitz, A. P. N., Wendt, C. & Zhong, X. *Biopreserv. Biobank.* **9,** 237–244 (2011).
4. Kugler, K. G. *et al. J. Clin. Bioinforma.* **1,** 9 (2011).
5. Kisand, K., Kerna, I., Kumm, J., Jonsson, H. & Tamm, A. *Clin. Chem. Lab. Med.* **49,** 229–235 (2011).
6. Simeon-Dubach, D. & Perren, A. *Nature* **475,** 454–455 (2011).
7. Vaught, J., Rogers, J., Carolin, T. & Compton, C. *J. Natl Cancer Inst. Monogr.* **2011,** 24–31 (2011).
8. Scott, C. T., Caulfield, T., Borgelt, E. & Illes, J. *Nature Biotechnol.* **30,** 141–147 (2012).

BIODIVERSITY

Remote responsibility

International trade is the underlying cause of 30% of threatened animal species extinctions, according to a modelling analysis of the impact of global supply chains and consumption patterns on biodiversity. SEE LETTER P.109

EDGAR HERTWICH

Biodiversity loss has just become a little more personal. Your freshly brewed cup of coffee is implicated in causing a significant number of threats of animal extinctions, according to a study by Lenzen *et al.*[1] on page 109 this issue. The authors present an analysis of species threats associated with internationally traded commodities, based on a detailed model of the global supply chains that connect final consumption to economic activities — and thus, for example, coffee drinking to species vulnerability.

If you buy a set of chess figures carved from ivory, you can suspect that you have contributed to killing an elephant. But if you buy a sausage, you cannot know whether the pig that was turned into the sausage was fed soy meal sourced from a farm that had just expanded into elephant habitat. The effects on species diversity, however, are similar. Understanding the complete causality chains leading to animal species extinctions has proven an intractable problem. Although the causes of individual threats to species are routinely identified when these species are 'red-listed' as vulnerable, endangered or extinct, the driving forces behind these immediate causes have until now escaped quantification. This incomplete understanding has hindered us from seeing the big picture and appropriately identifying the importance of different drivers.

The difficulties in linking proximate causes, such as the consumption of specific goods by identifiable groups of people, to immediate threats to biodiversity, such as habitat change, arise from both the sheer complexity of causal relationships that run through interconnected environmental and human systems, and from a lack of adequate indicators. Lenzen and colleagues present two significant advances in making such connections. The first is their model, the most detailed yet to describe the economic relationships between production and consumption. The second is their use of the threat causes recorded by the International Union for Conservation of Nature's Red Lists to identify direct links between threatened species and economic production activity.

Multiregional input–output models trace the multiple inputs required by manufacturing industry and other producing sectors, even across international borders, and have become the tool of choice for analysing the environmental pressures of consumption activities[2,3]. For example, such models have linked greenhouse-gas emissions from production activities in emerging economies to consumption in affluent countries[3], and shown that the emissions resulting from importation to affluent countries are increasing at a faster rate than the emissions associated with exported goods[4]. Lenzen *et al.* present a new multiregional input–output model, which they constructed 'from the bottom up' using a wide range of data sources — primarily national input–output tables and trade data. Their modelling combines powerful computation with novel approaches for reconciling conflicting data and estimating data points for which no primary data exist.

The authors then used their model to link economic activity to biodiversity (Fig. 1). Conventionally, this type of assessment links economic activity to individual environmental pressures that have been identified[5] as threats to biodiversity, such as land use, water use, or the over-fertilization of land and water. This approach allows the contribution of consumption to different environmental pressures to be quantified[6]. Impacts are then assessed using mechanistic models that connect the environmental pressures with an intermediate or final indicator of ecosystem impact, such as species threats. There are many such mechanisms, however, and some are highly site-dependent, so that this assessment approach is not able to provide a satisfactory picture of global biodiversity impacts[7]. An alternative approach is to link biodiversity models to environmental pressures[8], but such analysis has not yet, to my knowledge, been connected to models of global supply chains.

Pragmatically, Lenzen *et al.* circumvent any attempt to model the causal relationship between environmental pressure and ecosystem impact, and rely instead on threat causes provided in the Red Lists, such as 'smallholder farming' and 'logging and wood harvesting'. These causes are thereby connected in the input–output tables to specific industries, such as farming and forestry. When more than one industry can be connected to a cause, the responsibility is distributed in proportion to the economic importance of the industry. The fundamental unit of measurement in the authors' system is thus national species-threat

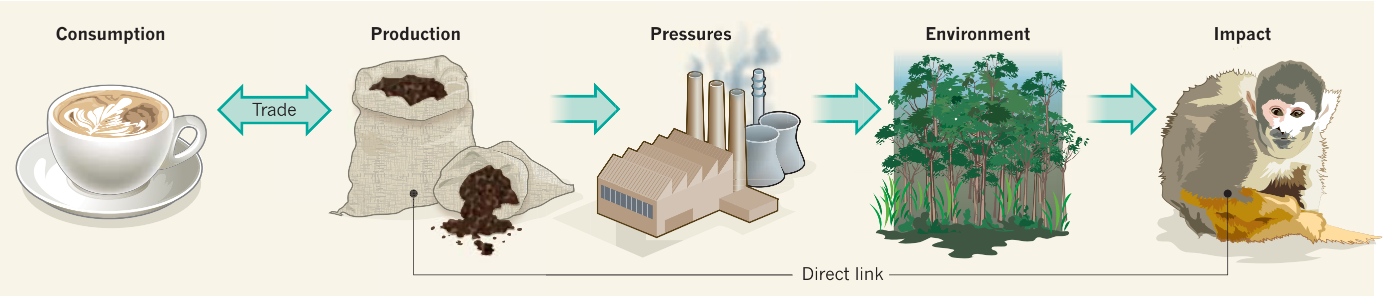


Figure 1 | Exporting species threats. The causal link between consumption and biodiversity loss involves the driving forces of economic activities (production, trade and consumption); the pressures exerted by these activities (such as resource extraction, pollution and land use); and the environmental processes, such as habitat change, that link these pressures to impacts, which include species threats. Impact-assessment methods typically trace causality from pressure to impact, but Lenzen *et al.*[1] have used established causes of observed species threats to link biodiversity loss directly to economic activity.

records; in other words, the instances of a species being put on a Red List in a given country. In the model, fractional responsibilities are then redistributed, using the input–output calculations, to final consumers all over the world. An example from the study is the Central American spider monkey *Ateles geoffroyi*, the red-listing of which is specified as resulting from habitat loss linked to coffee and cocoa plantations (Fig. 1).

Lenzen and colleagues' results indicate that 30% of instances of red-listed species worldwide are caused by internationally traded commodities, and that the United States, Japan and European countries are the main net 'importers' of species threats, whereas southeast Asian countries are the main net 'exporters'— the region in which the most species threats arising from trade occur. The authors show that the contribution of trade to biodiversity threats is similar to its contribution to global carbon dioxide emissions[3,4], although it is China, Russia and South Africa that are the largest emissions exporters.

There is some risk that this research overemphasizes the effect of international trade because, in developing countries, the production of cash crops for export results in a higher added value than subsistence agriculture, so that species threats may be disproportionally allocated to exported crops. Starting a cause–effect analysis from the effect side, as Lenzen and colleagues have done, is a novel and interesting approach. However, their results should be corroborated by further research exploring the linkage of pressures on biodiversity through global trade to consumption.

This study provides an indication of which areas of consumption need to be targeted to reduce biodiversity threats, which is a valuable contribution. The fundamental question that remains is whether the current (and increasing) scale of consumption will inevitably cause these threats, or whether ways could be found to satisfy this consumption but allow affluent consumers to reduce their impact, such as improved labelling systems and lower-impact production methods. ■

Edgar Hertwich *is in the Department of Energy and Process Engineering, Norwegian University of Science and Technology, 7941 Trondheim, Norway.*
e-mail: edgar.hertwich@ntnu.no

1. Lenzen, M. *et al. Nature* **486,** 109–112 (2012).
2. Wyckoff, A. W. & Roop, J. M. *Energy Policy* **22,** 187–194 (1994).
3. Peters, G. P. & Hertwich, E. G. *Environ. Sci. Technol.* **42,** 1401–1407 (2008).
4. Peters, G. P. *et al. Nature Clim. Change* **2,** 2–4 (2012).
5. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Synthesis* (Island, 2005).
6. Hertwich, E. *et al. Assessing the Environmental Impacts of Consumption and Production: Priority Products and Materials* (UNEP, 2010).
7. Curran, M. *et al. Environ. Sci. Technol.* **45,** 70–79 (2011).
8. Pereira, H. M. *et al. Science* **330,** 1496–1501 (2010).

CELL BIOLOGY

High-tech yeast ageing

A method commonly employed to study replicative ageing in yeast is laborious and slow. The use of miniaturized culture chambers opens the door for automated molecular analyses of individual cells during ageing.

MICHAEL POLYMENIS & BRIAN K. KENNEDY

Similarly to many cells in our body, the cells of budding yeast cannot replicate indefinitely. On division, a yeast cell gives rise to a mother cell and a 'fresh' daughter cell. The mother cell can produce, on average, only about 25 daughters before it dies. A test that measures the replicative lifespan of yeast cells has become a popular way to study ageing processes, and researchers have used it to identify genes and pathways that were later confirmed to have roles in longevity in animals[1–4]. However, such an assay is labour intensive and cannot be implemented in a high-throughput fashion[5]. Two studies, one by Lee *et al.*[6] in *Proceedings of the National Academy of Sciences* and another by Xie *et al.*[7] in *Aging Cell*, offer modified versions of the assay that are amenable to automation and that allow the study of ageing processes in yeast cells to be made in unprecedented detail. The techniques use tiny chambers to retain mother cells and wash away daughters, coupled to powerful microscopes capable of time-lapse photography.

In the conventional replicative ageing assay, the experimenter must look through a microscope and painstakingly remove each daughter cell after division using a small needle on the surface of thick, solid culture media (Fig. 1a). Moreover, just as with other organisms, there is significant variation in lifespan between individual yeast cells, even when they are genetically identical. This means that a minimum of 40 cells have to be interrogated to generate a reliable lifespan data set, which necessitates the manual removal of approximately 1,000 daughter cells.

Lee *et al.* and Xie *et al.* replaced the manual approach with transparent microfluidic devices that consisted of submillimetre-scale channels and tunnels through which nutrient broth flows in a controlled manner (Fig. 1b). Such a set-up allowed the authors to apply high-resolution microscopy techniques for tracking individual cells and molecular markers.

Subtle differences exist between the two systems, however. Lee and colleagues suspended the yeast cells between silicone micropads and thin cover glass. The micropads were slightly lifted by the hydrostatic pressure of the broth during loading of a cell suspension, and they held the mother cells after release of the pressure. Daughter cells were washed away because of their smaller size.

By contrast, Xie *et al.* trapped the cells in 'micro-jails' from which the daughters could escape through gates. The researchers also attached biotin molecules to the mothers' cell walls, causing these cells to adhere to the chambers' surfaces, which had been coated with avidin (a protein that binds biotin with high affinity). This ensured that only mother cells remained trapped, as the synthesis of new cell wall in yeast is confined to daughter cells, and no biotin was supplied after the initial labelling of the mother cells.

Interestingly, both groups of authors studied

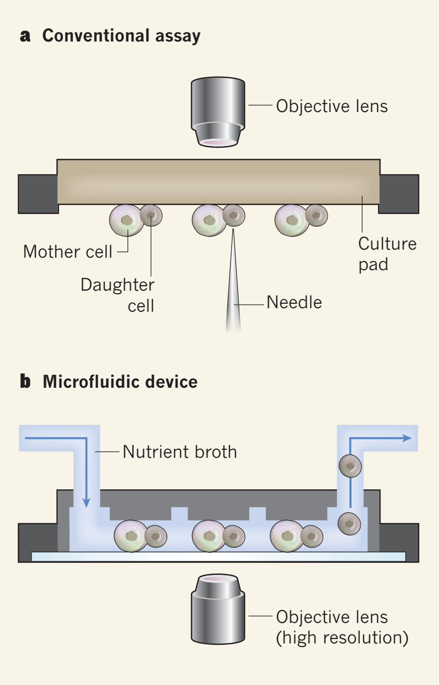


Figure 1 | Watching how cells age. Budding yeast divides by forming a bigger mother cell and a smaller daughter cell. As a measure of lifespan in yeast, researchers count the number of daughters produced by each mother. **a,** In a conventional assay, yeast cells are grown on the surface of thick culture media, and the researcher removes daughter cells — one by one — using a needle and a microscope. **b,** Lee *et al.*[6] and Xie *et al.*[7] developed transparent microfluidic devices that trap mother cells in small chambers, whereas daughters are washed away by a controlled flow of nutrient broth. The authors used high-resolution microscopes to track changes associated with ageing in individual cells. (Figure modified from ref. 6).

the same yeast strain but came up with contrasting results: the mean lifespan observed by Lee *et al.* was 25 cell divisions, whereas that found by Xie *et al.* was 18. A mean lifespan of approximately 27 was previously reported for the same strain when measured using the conventional replicative ageing assay[8]. Such discrepancies remain unexplained, but could be due to the different methods used. Importantly, however, the two microfluidic assays recapitulated the effects of known longevity mutants; for example, a strain lacking the gene *FOB1* was found to have an increased lifespan, as expected.

Both studies documented significant diversity among individual cells of the same population as the cells aged and died. At the time of death, a spherical cell shape correlated with shorter lifespan, whereas an elongated shape was linked to longer lifespan. Furthermore, ageing and death were associated with profound changes in the morphology and function of intracellular organelles such as vacuoles (which became fragmented with age[6]) and mitochondria (which showed increased dysfunction[7]).

Xie and colleagues went one step further by demonstrating that they could track fluorescent molecular markers in individual ageing cells using high-resolution microscopy. In this way, they found that potentially harmful reactive oxygen species (ROS) increased in old cells. They also observed that expression of the Hsp104 chaperone — a protein that assists other proteins in folding — was inversely correlated with lifespan, although not with an accompanying increase in aggregates of damaged proteins as might have been anticipated.

The continuous imaging afforded by the two microfluidic approaches is likely to identify new and better molecular landmarks of cellular ageing, which are not accessible with the conventional replicative ageing assay. In addition, with further development, the microfluidic platforms will greatly facilitate high-throughput applications. Nevertheless, the conventional assay, albeit laborious, may remain the least biased of the three. For example, both microfluidic ageing assays rely on differences in size between mother and daughter cells. Because cell size seems to affect lifespan[9], it needs to be established how the microfluidic assays perform in cases where the differences between mother and daughter may be diminished, or with mutants of altered cell size or shape. This might be less of a concern in Xie and colleagues' method, in which the labelling of the mothers with biotin helps confine them to the chamber, independently of their larger size.

It should also be noted that the microfluidic systems do not solve another shortcoming of the conventional assay: the poor yield of old cells, which are required if the researchers want to do biochemical tests on them. A different replicative ageing assay, the mother-enrichment programme[10], may be better for exploring biochemical questions. In this system, yeast cells are genetically engineered to inhibit daughter-cell division in liquid culture, allowing the accumulation of large numbers of aged mother cells.

Clearly, by using the microfluidic systems and the mother-enrichment programme, it is becoming possible to combine biochemistry, cell biology and high-resolution lifespan measurement to make the most of the yeast replicative ageing model. Combining all these tools will enable scientists not only to identify molecular markers of ageing, but also to dissect the causal role of these factors in the ageing process. Furthermore, the anticipated increase in the throughput of the replicative ageing assays will usher in large-scale screening to identify small molecules that could modulate cellular ageing. With these tools in place, yeast replicative ageing is becoming an ideal experimental model that, when coupled to systems-biology approaches, may yield for the first time a holistic understanding of ageing in an organism. ■

Michael Polymenis *is in the Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77845, USA.*
Brian K. Kennedy *is at the Buck Institute for Research on Aging, Novato, California 94945, USA, and at the Aging Research Institute, Guangdong Medical College, Dongguan, China.*
e-mails: polymenis@tamu.edu; bkennedy@buckinstitute.org

1. Kaeberlein, M. *Nature* **464,** 513–519 (2010).
2. Kaeberlein, M. *et al. Science* **310,** 1193–1196 (2005).
3. Kaeberlein, M., McVey, M. & Guarente, L. *Genes Dev.* **13,** 2570–2580 (1999).
4. Smith, E. D. *et al. Genome Res.* **18,** 564–570 (2008).
5. Steffen, K. K., Kennedy, B. K., Kaeberlein, M. *J. Vis. Exp.* (28), e1209 (2009).
6. Lee, S. S., Vizcarra, I. A., Huberts, D. H. E. W., Lee, L. P. & Heinemann, M. *Proc. Natl Acad. Sci. USA* **109,** 4916–4920 (2012).
7. Xie, Z. *et al. Aging Cell* http://dx.doi.org/10.1111/j.1474-9726.2012.00821.x (2012).
8. Kaeberlein, M., Kirkland, K. T., Fields, S. & Kennedy, B. K. *PLoS Biol.* **2,** 1381–1387 (2004).
9. Yang, J. *et al. Cell Cycle* **10,** 144–155 (2011).
10. Lindstrom, D. L. & Gottschling, D. E. *Genetics* **183,** 413–422 (2009).

ASTROPHYSICS

Young dwarfs date an old halo

An ingenious way of measuring the ages of stellar populations in the halo of the Milky Way will allow astronomers to obtain direct information on the timing of the Galaxy's evolution. SEE LETTER P.90

TIMOTHY C. BEERS

In a paper[1] in this issue, Kalirai describes a method for estimating the ages of the progenitors of newly formed white-dwarf stars in the halo of our Galaxy, the extended region outside the plane of the Galaxy's disk (Fig. 1). This new chronometer, reported on page 90, provides a means to determine the ages of stellar populations in the halo, and will increase our knowledge of how and where the Galaxy's stars have formed and evolved*.

Advances in the understanding of how stars evolve, and detailed observations of large numbers of globular clusters (tightly bound, dense stellar systems) in our Galaxy, have yielded age estimates for individual clusters of between 10 billion and 13 billion years, with claimed precisions of between 0.5 billion and 1.0 billion years[2].

Current methods for measuring the ages of the halo 'field' population — that is, of stars not in clusters — rely on one of two approaches. The first involves theoretical predictions of the relationship between the age, composition, temperature and luminosity of a given star, which place the star at a unique position in an observed two-dimensional diagram of colour versus apparent brightness[3]. The second approach uses empirical comparisons between the locations on the diagram of collections of stars that have just exhausted the supply of hydrogen in their core (main-sequence turn-off stars) with the locations of similar stars in globular clusters.

Both of these methods typically yield age estimates that are similar to those obtained for globular clusters, but with precisions no better than 1 billion to 2 billion years[4] — twice as large as for the clusters. These estimates are also subject to other systematic uncertainties due to evolutionary effects on a star's atmosphere, such as atomic diffusion. Such uncertainties may alter the expected position of main-sequence turn-off stars on the diagram[5].

Alternative techniques for calculating the ages of individual halo-field stars include use of the measured abundances of radioactive species, such as uranium and thorium, in ancient stars that have low metallicities (they

*This article and the paper[1] under discussion were published online on 30 May 2012.

2MASS/J. CARPENTER, T. H. JARRET & R. HURT

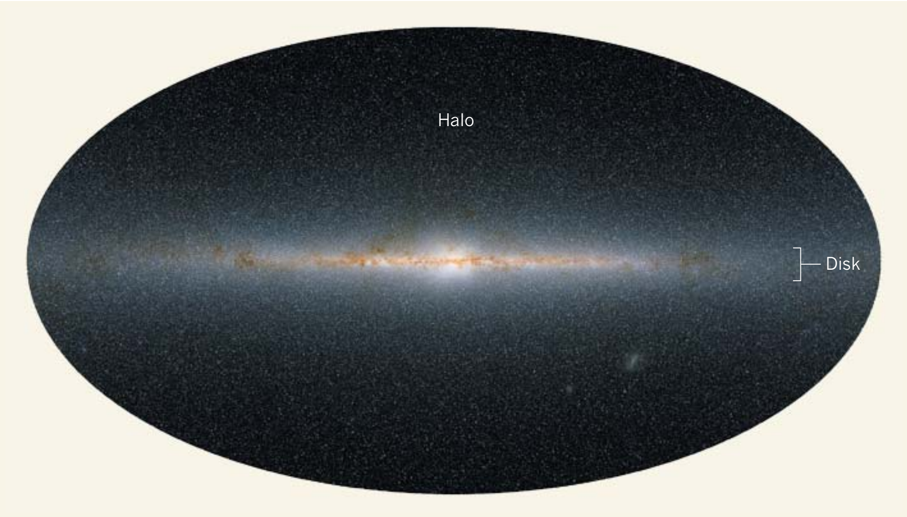


Figure 1 | Edge-on infrared view of the Milky Way. Kalirai[1] has estimated the ages of stars in the Galactic halo — the extended region outside the plane of the Galaxy's disk — from measurements of newly formed halo white dwarfs. The halo is likely to consist of at least two stellar populations, including one that dominates the inner part of the halo and another that dominates the outer halo.

are deficient in elements heavier than hydrogen and helium). However, such techniques are limited to precisions no better than about 2 billion to 3 billion years[6]. This limitation is due to the difficulty of measuring the weak signatures of these elements in stellar spectra, and to incomplete knowledge of the production ratios of such species in nuclear reactions that involve rapid neutron capture.

In his study, Kalirai used an ingenious approach. The technique is based on estimates of the masses of stars that have just turned into white dwarfs and are found in the halo. Because more-massive dwarfs originate from younger stars, whereas less-massive dwarfs come from older stars, the author could estimate the ages of their progenitors. This step required calibration of the relationship between mass and age, which was obtained by comparing these dwarfs with newly formed white dwarfs observed in a globular cluster called Messier 4, which is estimated to be 12.5±0.5 billion years old[2]. These recently minted white dwarfs are the immediate descendants of stars that have exhausted all sources for nuclear fusion in their cores. Over time, they cool to invisibility as they slowly release the internal heat of the former core of their parent main-sequence (hydrogen-fusing) star.

By focusing on white dwarfs that have just formed, Kalirai's mass and age estimates obviate the need to understand the complex cooling process of a white dwarf, which depends on the composition of the white dwarf's outer layers and on changes in its structure that occur over time. His age estimate for the progenitors of a few members of the stellar population in the inner-halo field that are found relatively close to Earth is 11.4±0.7 billion years; because of their low luminosities, only white dwarfs that are nearby can be studied in detail. Improved estimates will come from adding larger numbers of inner-halo white dwarfs, and, in particular, from comparisons with globular clusters that have different ages and metallicities from Messier 4.

The importance of this deft tool for estimating the ages of white dwarfs, and the stellar populations of the Galactic halo with which they are associated, lies in its ability to resolve details of how the halo-field stars formed and evolved during the period between 10 billion and 13 billion years ago, a spread in age comparable to the precisions of previous age-determination methods[4–6]. Recent observations[7–9] and numerical models of the formation of the Galactic halo[10–12] strongly suggest that the halo consists of at least two stellar populations: an inner-halo and an outer-halo population. These would have differing spatial distributions, kinematics and composition, and would also contain gravitationally bound debris from recent mergers of the Galaxy with smaller individual galaxies, such as the Sagittarius dwarf galaxy.

Current understanding suggests that the Galaxy was assembled from hierarchical mergers of small proto-Galactic fragments. Stars of the inner-halo population, to which the motions of Kalirai's white dwarfs suggest they belong, are thought to have originated from the collective mergers of proto-Galactic fragments of relatively high mass and metallicity[11,13]. These progenitor fragments were able to attain higher metallicities because they could retain an interstellar medium throughout many bursts of star formation, with each burst polluting this material with the nucleosynthetic products from which newer stars formed. Such fragments may have been similar to the more-massive surviving satellites of the Milky Way (such as the Fornax, Sculptor, Sextens and Carina dwarf spheroidal galaxies).

By contrast, stars of the outer-halo population are likely to have formed from fragments of relatively low mass and metallicity[11]. Such low metallicities suggest that outer-halo stars formed in these fragments early in the history of the Galaxy, and that the fragments underwent a single burst of star formation that either consumed or drove out any interstellar medium capable of forming newer stars[14]. These outer-halo progenitors may resemble the ultra-faint dwarf spheroidal galaxies — such as Ursa Major II and Leo IV — that have been discovered by the Sloan Digital Sky Survey[15]. Such spheroidal galaxies have subsequently been shown to contain relatively large numbers of stars that have extremely low metallicities. Some of the stars in these ultra-faint dwarf galaxies have metallicities close to the lowest yet found in the Galactic halo[16].

The identification and analysis of white dwarfs among halo-field stars in addition to those investigated by Kalirai will, in principle, allow a distinction to be made — on the basis of their differing kinematics — between dwarfs that could be associated with inner-halo progenitors and those with outer-halo progenitors. Assuming that the interpretation of a dual halo applies, one would expect to see differences in the inferred ages of the different populations (the inner halo being somewhat younger than the outer halo), and in the derived spread of the ages of the white dwarfs associated with the two populations. Strong constraints could then be placed on the duration of star formation in the inner-halo population, and the expectation that outer-halo progenitors experienced only a single burst of star formation could be tested empirically. Ongoing and future large-scale surveys, both from the ground and in space, will supply the required samples of halo white dwarfs for such investigations. Kalirai's white-dwarf chronometer provides a valuable tool for exploring this anticipated wealth of information. ■

Timothy C. Beers *is at the Kitt Peak National Observatory and the National Optical Astronomy Observatory, Tucson, Arizona 85719, USA.*
e-mail: beers@noao.edu

1. Kalirai, J. S. *Nature* **486,** 90–92 (2012).
2. Dotter, A. *et al. Astrophys. J.* **708,** 698–716 (2010).
3. Holmberg, J., Nordström, B. & Andersen, J. *Astron. Astrophys.* **501,** 941–947 (2009).
4. Jofré, P. & Weiss, A. *Astron. Astrophys.* **533,** A59 (2011).
5. Salaris, M., Groenewegen, M. A. T. & Weiss, A. *Astron. Astrophys.* **355,** 299–307 (2000).
6. Frebel, A. *et al. Astrophys. J.* **660,** L117–L120 (2007).
7. Carollo, D. *et al. Nature* **450,** 1020–1025 (2007).
8. Carollo, D. *et al. Astrophys. J.* **712,** 692–727 (2010).
9. Beers, T. C. *et al. Astrophys. J.* **746,** 34 (2012).
10. Font, A. S. *et al. Mon. Not. R. Astron. Soc.* **416,** 2802–2820 (2011).
11. Tissera, P. B., White, S. D. M. & Scannapieco, C. *Mon. Not. R. Astron. Soc.* **420,** 255–270 (2012).
12. McCarthy, I. G. *et al. Mon. Not. R. Astron. Soc.* **420,** 2245–2262 (2012).

13. Bullock, J. S. & Johnston, K. V. *Astrophys. J.* **635,** 931–949 (2005).
14. Qian, Y.-Z. & Wasserburg, G. J. *Proc. Natl Acad. Sci. USA* (in the press); preprint available at http://arxiv.org/abs/1202.3202 (2012).
15. York, D. G. *et al. Astron. J.* **120,** 1579–1587 (2000).
16. Frebel, A. & Norris, J. E. in *Planets, Stars and Stellar Systems* Vol. 5 (ed. Gilmore, G.) (Springer, in the press); preprint available at http://arxiv.org/abs/1102.1748 (2012).

GEOCHEMISTRY

A dash of deep nebula on the rocks

The cocktail of noble-gas isotopes in an Icelandic rock suggests that the upper mantle does not, and never did, receive gas from a deeper mantle reservoir. This challenges ideas of deep Earth's behaviour and formation. SEE LETTER P.101

CHRIS J. BALLENTINE

The pattern of isotopic abundances of inert and rare noble gases, trapped in small bubbles in volcanic rock, act as a 'fingerprint' of how and where our planet first acquired its gas. Furthermore, the type of volcanic setting, and the way that parts of the fingerprint change with time, offer insight into the workings of deep Earth. Squeezing out this information from lava derived from the deepest parts of our planet — possibly some 2,900 kilometres beneath our feet — is a challenge. Yet this is precisely what S. Mukhopadhyay[1] has done, and in spectacular fashion. On page 101 of this issue, he reports the long-awaited detailed secrets of the planet's deepest gases, based on an isotope analysis performed using a new-generation mass spectrometer.

Helium is light enough to be lost from the atmosphere to space, and so its atmospheric concentration is very low. Its isotopic composition in mantle rocks (measured as the ratio of the gas's two isotopes, ^{4}He and ^{3}He) is therefore the easiest to ascertain of all the noble gases, because measurements are not swamped by background 'noise' from contaminating atmospheric helium. The ^{4}He/^{3}He ratios at mid-ocean ridges — the 65,000 km of interconnected underwater volcanic systems that spew magma from the uppermost mantle to build new ocean crust — are almost constant around the globe. But lower ratios have been measured in rocks produced by certain 'hotspot' volcanoes, such as those in Hawaii and Iceland, which are thought to tap the deepest mantle.

The existence of different ^{4}He/^{3}He ratios underpins the idea that there are at least two geochemical reservoirs in the mantle[2]: a deep reservoir rich in gases and volatile compounds feeds material into an upper reservoir, which is the convecting part of the mantle that supplies magma to mid-ocean ridges (Fig. 1). Although ideas about the depth, size and nature of the deepest reservoir have changed substantially, the two-reservoir model has dominated attempts to explain observations of mantle geochemistry for the past 30 years.

Obtaining robust information about the isotopic composition of the heavy noble gases in the mantle (neon, argon, krypton and xenon) has been far harder to do than it was for helium. This is because the atmospheric concentrations of these gases are much higher than that of helium, greatly increasing the background noise caused by air contamination of mantle samples. (No magma can erupt into the ocean or the atmosphere without the resulting basalt rock becoming contaminated by air — it would be like jumping into a swimming pool and expecting to stay dry.) Our detailed understanding of the upper mantle's heavy noble gases has therefore come almost entirely from only two rare samples in which such contamination is minimized: a single gas-rich basalt dredged from the mid-Atlantic ridge[3]; and volcanic gas trapped in a deep carbon dioxide gas field[4,5] in New Mexico.

Even fewer traces of noble gases are found in basalts produced by hotspot volcanism than in those produced at mid-ocean ridges[6], making hotspot rocks highly susceptible to air contamination. However, in one area of Iceland, a basalt has been found[7] in which more mantle gas is known to have been preserved than in most basalts, in part because it erupted under an ice cap. Mukhopadhyay[1] has now re-examined this basalt by applying a technique that allows large samples of the rock to be crushed under vacuum (to protect the isotopic fingerprint of gases in the rock from air contamination), and then analysing the released gases using one of a new generation of mass spectrometers that greatly increases the precision with which isotope ratios are measured[5]. In this way, he has teased out a veritable cornucopia of fresh information.

The isotopic composition of neon in the basalt suggests that the deep Iceland mantle gases originated from the solar nebula — the cloud of dust and gas from which the planets of the Solar System formed. Such gas was around only for the first few tens of millions of years

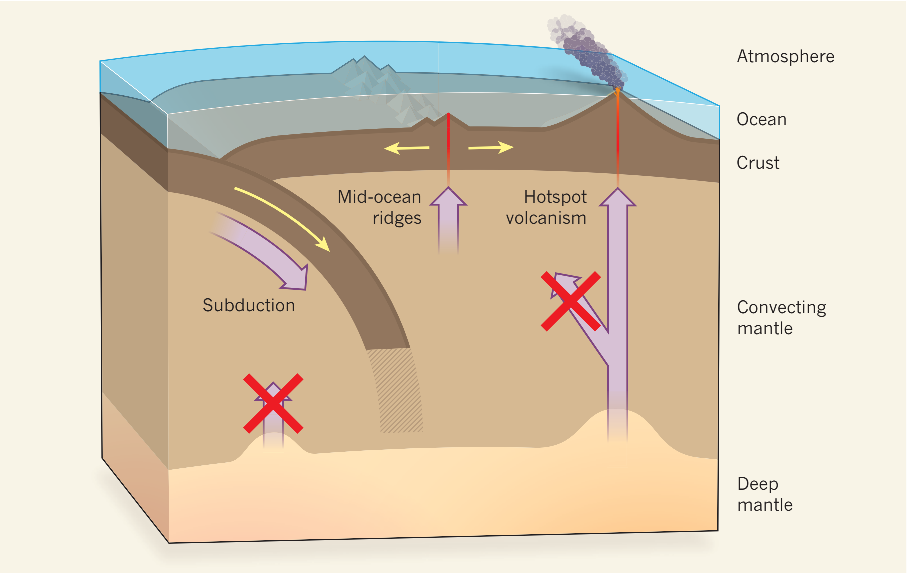


Figure 1 | Mantle movement. Magma from the upper part of the convecting mantle erupts at mid-ocean ridges, whereas that from a deep reservoir is thought to erupt at 'hotspot' volcanoes. Subduction processes transfer material from the ocean crust back into the convecting mantle, and possibly also into the deep mantle. The isotopic composition and amount of helium in the upper mantle suggest a flux of ^{3}He through this region[13], entering from the deep reservoir and exiting to the oceans and atmosphere. Mukhopadhyay reports[1] that the isotopic and elemental compositions of noble gases from an Iceland mantle plume that is thought to originate from the deep mantle are fundamentally different from those in the convecting mantle. If the Iceland plume is representative of the deep-mantle reservoir, this rules out the possibility of a large transfer of noble gases from the deep to the convecting mantle (red crosses), overturning a long-standing model of mantle-gas geochemistry. Yellow arrows, crust movement; pink arrows, gas transport.

of the Solar System's formation, after which it was either pulled into the Sun or blasted out of the Solar System when the Sun ignited. By contrast, the noble gases in the upper mantle came from meteorites[4,5]. Mukhopadhyay's findings, together with those of others, allow us to appreciate the true complexity of gas delivery to our planet from different sources at different stages of Earth's infancy[1,4,5,8].

But the questions of how gas from the solar nebula was trapped in the solid parts of growing planets, and how the gas was preserved through early accretionary events, will certainly test our models of accretion. Some of the noble-gas isotopes from the Icelandic deep mantle came from long-dead radioactive isotopes of iodine and plutonium that were present in the early Solar System. Mukhopadhyay[1] compared these noble-gas isotopes with those in the convecting mantle[9], and concluded that the Iceland deep mantle formed in a drier environment, and preserved a higher proportion of its plutonium-decay gases, than did the convecting mantle. This chimes with the idea of a process or location in the deep mantle that has preserved the earliest geochemical signals of accretion exceptionally well. Mukhopadyay's findings may also help to connect theories of how a planet starts to obtain its gas with evidence[10] from other isotope systems that also points to the very early formation of reservoirs hidden in the deep mantle.

Although many geochemists have argued that Earth contains a deep, gas-rich reservoir, they have struggled to pinpoint where it should be. Ever since it became apparent from seismic tomography that Earth's mantle was not nicely layered[11], the location or processes that could prevent such a deep reservoir from mixing into the convecting mantle and disappearing completely have remained enigmatic. Wherever this reservoir might be, it has survived the cataclysmic Moon-forming event (in which Earth was struck by a Mars-sized body)[12]; avoided mixing with volatile compounds brought to Earth by meteorites; and withstood continual removal of material by mantle plumes.

One result from Mukhopadhyay's work is touched on only lightly by the author, but might have the greatest impact on how we think the mantle behaves. If the isotopic composition of the basalt analysed by Mukhopadhyay — and therefore of the Iceland plume from which this hotspot rock is derived — is indeed representative of a deep mantle reservoir, then this reservoir cannot also be the source of ^{3}He needed to explain the ^{4}He/^{3}He ratio in the upper mantle, because the heavy noble gases in the basalt don't match those in the upper mantle. The two-reservoir mantle model must therefore be modified. Mukhopadhyay's data about the cocktail of mantle noble gases, however, will endure. ■

Chris J. Ballentine *is at the School of Earth, Atmospheric and Environmental Sciences, The University of Manchester, Manchester M13 9PL, UK.*
e-mail: chris.ballentine@manchester.ac.uk

1. Mukhopadhyay, S. *Nature* **486,** 101–104 (2012).
2. Kurz, M. D., Jenkins, W. J. & Hart, S. R. *Nature* **297,** 43–47 (1982).
3. Moreira, M., Kunz, J. & Allegre, C. J. *Science* **279,** 1178–1181 (1998).
4. Ballentine, C. J., Marty, B., Sherwood Lollar, B. & Cassidy, M. *Nature* **433,** 33–38 (2005).
5. Holland, G., Cassidy, M. & Ballentine, C. J. *Science* **326,** 1522–1525 (2009).
6. Gonnermann, H. M. & Mukhopadhyay, S. *Nature* **449,** 1037–1040 (2007).
7. Harrison, D., Burnard, P. & Turner, G. *Earth Planet. Sci. Lett.* **171,** 199–207 (1999).
8. Marty, B. *Earth Planet. Sci. Lett.* **313,** 56–66 (2012).
9. Pepin, R. O. & Porcelli, D. *Earth Planet. Sci. Lett.* **250,** 470–485 (2006).
10. Boyet, M. & Carlson, R. W. *Science* **309,** 576–581 (2005).
11. van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. *Nature* **386,** 578–584 (1997).
12. Canup, R. M. & Asphaug, E. *Nature* **412,** 708–712 (2001).
13. Kellogg, L. H. & Wasserburg, G. J. *Earth Planet. Sci. Lett.* **99,** 276–289 (1990).

NEUROSCIENCE

Sibling neurons bond to share sensations

Two studies show how electrical coupling between sister neurons in the developing cerebral cortex might help them to link up into columnar microcircuits that process related sensory information. SEE LETTERS P.113 & P.118

THOMAS D. MRSIC-FLOGEL & TOBIAS BONHOEFFER

A pioneering set of experiments in the 1950s and 1960s inspired generations of neuroscientists to explore how the anatomy of the brain gives rise to its function[1–3]. When researchers lowered electrodes into the cerebral cortices of cats and monkeys, they found that neurons lying above and below each other form functional columns — that is, they respond in a similar way to certain stimuli, such as touch on specific areas of the skin or the orientation of an elongated visual stimulus.

Even though such cortical columns have long been considered to be exemplars of basic computational units of cortical organization, the precise relationship between their anatomy and function has been difficult to define and remains the subject of debate[4–5]. This is particularly true in rodents, in which the cortex seems to lack functional columns almost entirely. What is common to rodents and other mammals, however, is a highly specific organization of cortical connections that link neurons across layers in the cortex to relay and process related sensory information[6–8]. Reporting in this issue, Yu *et al.*[9] (page 113) and Li *et al.*[10] (page 118) reveal some of the developmental events that could give rise to such precisely arranged functional circuits.

It has long been known that, during embryonic development of the cortex, neuronal progenitor cells give birth to daughter cells that migrate towards the brain surface to form strings of 'sibling' neurons that span the cortical layers (Fig. 1a). These radially aligned clones, referred to as radial units or ontogenetic columns, have been proposed to constitute the basis of the functional columns in the mature brain[11]. However, a direct link between cellular lineage, microcircuit development and the sensory preference of neurons had not been demonstrated.

Yu and colleagues[9] used viruses to label sibling neurons in the developing cortex of mouse embryos with a fluorescent protein, and then recorded the cells' electrical activity in brain slices prepared shortly after birth. The authors showed that gap junctions — small pores that couple adjacent cells electrically by bridging their membranes — formed transiently between sibling neurons in the same radial unit, very early in development (Fig. 1b). Gap junctions had previously been observed between clusters of excitatory neurons in the developing cortex and had been proposed to contribute to the establishment of neuronal assemblies[12], but the ancestry and significance of such cell clusters were unknown. Moreover, other work had revealed that, later in development, neurons in radial clones mostly connect to one another through chemical synapses[13] mediated by neurotransmitter molecules (Fig. 1c). Yu *et al.* showed that gap-junction inactivation abolished the formation of such synapses, and report that transient electrical coupling is thus essential for the establishment of chemical synapses between sibling neurons.

Li and colleagues[10] used the same method to label radial clones, and then used a microscopy technique known as two-photon calcium imaging[14] to monitor the activity of sibling neurons in the cortex of live mice in response to visual stimuli. The authors observed that clonally related neurons, when compared with a random subset of neighbouring cells,

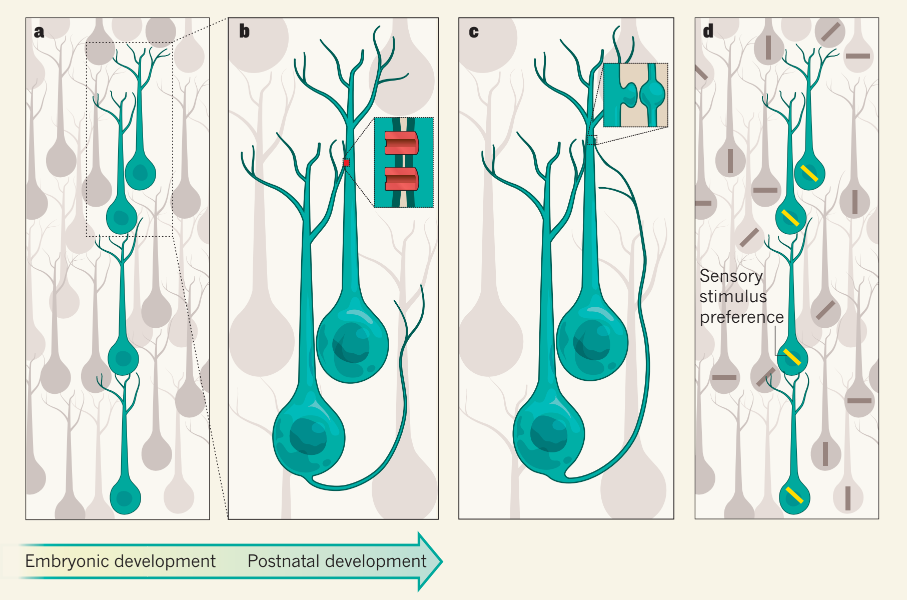


Figure 1 | A link between neuronal lineage, connectivity and sensory preference. **a**, During embryonic development, newly born neurons migrate towards the surface of the cortex and form strings of sibling neurons that span the cortical layers. Yu *et al.*[9] and Li *et al.*[10] used viruses that express a green fluorescent protein to label neurons derived from a single progenitor cell. **b**, Yu *et al.* show that, early in development, sibling neurons are preferentially connected by small pores called gap junctions (inset) that enable electrical currents to pass directly between them. **c**, As development proceeds, gap junctions disappear and chemical synapses (inset) are preferentially established between sibling neurons. **d**, Li *et al.* describe how, in a later developmental phase, sibling neurons respond to similar sensory features, such as the orientation of visual stimuli.

were more likely to respond to stimuli of the same orientation in the animals' visual field (Fig. 1d). Moreover, the blockade of gap junctions eliminated the shared preference for stimulus orientation, which further supports the idea that electrical coupling between sibling neurons plays a part in influencing the functional organization of the cortex.

The two studies are intriguing because they lend support to the involvement of genetic lineage in the assembly of precise columnar circuits in the cortex. An extreme interpretation of the results is that ontogenetic columns constitute an elementary unit of functional organization in the cortex; that is, a basic, repeating circuit comprising excitatory neurons that process related sensory information. However, the relevance of ontogenetic columns for sensory processing is still unclear; an excitatory neuron in the mammalian cortex receives inputs from at least 1,000 others, but only a handful of these connections are with sibling neurons. It will be important to assess how much the connections between siblings actually contribute to shaping their sensory responses.

The small size of mouse ontogenetic columns, as reported by Li *et al.* and Yu *et al.*, might explain why functional columns have not previously been described in the visual cortex of rodents, in which neurons with different sensory preferences seem to be locally intermixed[15,16]. But the authors' findings also raise the question of whether there is any relationship between ontogenetic columns and the much larger functional columns in the cortices of other mammals such as cats or primates. Large functional columns could form as aggregates of multiple ontogenetic mini-columns, or from larger radial clones containing many more neurons than those of rodents, or by a different mechanism altogether.

Regardless, the two studies demonstrate that at least some of the connection specificity in cortical microcircuits is established intrinsically by clonal lineage. The studies also show that cellular lineage could influence neurons' development of a similar sensory preference during early postnatal life — but how might this be achieved? Yu and colleagues' results suggest a close interplay between clonal lineage and early neuronal activity. Electrical coupling is likely to influence the formation and/or stabilization of chemical synapses between neurons that share gap junctions, because it is known that synapses can grow stronger or weaker if the cells' electrical activities are correlated or uncorrelated, respectively — a process known as synaptic plasticity. It is tempting to speculate that the sensory preference of these sub-networks might then be developed by similar mechanisms: electrically coupled neurons could select and stabilize a common set of sensory inputs, which would endow these cells with a shared preference for certain sensory features. Future experiments are required to determine the developmental events that define how non-sibling neurons with similar stimulus preferences become connected to each other in the cortical circuit.

Other important questions remain unanswered. To what extent is the electrical coupling between cortical neurons necessary for the establishment of stimulus selectivity? In other words, did the early blockade of gap junctions between sibling neurons result in a fundamentally altered visual cortex? Li and colleagues' results indicate that gap-junction blockade does not prevent the emergence of orientation preference, but more subtle features — such as the range of orientations that a neuron detects — could depend on gap-junction connectivity.

Another issue relates to the fact that, during embryonic development, clonally related neurons not only migrate radially towards the cortical surface — some are also distributed tangentially. The connectivity and functional fate of these displaced sibling neurons remains undetermined. Do they also establish functional sub-networks with their siblings, or do they form local connections with non-sibling neighbours? How does the interplay between sensory input and synapse plasticity link up neurons from different ontogenetic columns with similar sensory preferences to form larger functional assemblies? And do displaced sibling neurons have a role in this process? Whatever the answers to these questions, the two studies reveal the fascinating way in which neurons emerging from the same progenitor cell are destined to share functional properties, and thus show how the earliest developmental events influence the elaborate functional circuitry of the brain. ■

Thomas D. Mrsic-Flogel *is in the Department of Neuroscience, Physiology and Pharmacology, University College London, London WC1E 6DE, UK.* **Tobias Bonhoeffer** *is at the Max-Planck Institute of Neurobiology, 82152 München-Martinsried, Germany.*
e-mails: t.mrsic-flogel@ucl.ac.uk; tobias.bonhoeffer@neuro.mpg.de

1. Mountcastle, V. B. *J. Neurophysiol.* **20,** 408–434 (1957).
2. Hubel, D. H. & Wiesel, T. N. *J. Physiol.* **165,** 559–568 (1963).
3. Hubel, D. H. & Wiesel, T. N. *J. Physiol.* **195,** 215–243 (1968).
4. da Costa, N. M. & Martin, K. A. *Front. Neuroanat.* **4,** 16 (2010).
5. Horton, J. C. & Adams, D. L. *Phil. Trans. R. Soc. Lond. B* **360,** 837–862 (2005).
6. Binzegger, T., Douglas, R. J. & Martin, K. A. *J. Neurosci.* **24,** 8441–8453 (2004).
7. Yoshimura, Y., Dantzker, J. L. M. & Callaway, E. M. *Nature* **433**, 868–873 (2005).
8. Ko, H. *et al. Nature* **473,** 87–91 (2011).
9. Yu, Y. C. *et al. Nature* **486,** 113–117 (2012).
10. Li, Y. *et al. Nature* **486,** 118–121 (2012).
11. Rakic, P. *Science* **241,** 170–176 (1988).
12. Yuste, R., Nelson, D. A., Rubin, W. W. & Katz, L. C. *Neuron* **14,** 7–17 (1995).
13. Yu, Y. C., Bultje, R. S., Wang, X. & Shi, S. H. *Nature* **458,** 501–504 (2009).
14. Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. *Proc. Natl Acad. Sci. USA* **100,** 7319–7324 (2003).
15. Ohki, K. *et al. Nature* **433,** 597–603 (2005).
16. Mrsic-Flogel, T. D. *et al. Neuron* **54,** 961–972 (2007).

doi:10.1038/nature11115

Electrocatalyst approaches and challenges for automotive fuel cells

Mark K. Debe[1]

Fuel cells powered by hydrogen from secure and renewable sources are the ideal solution for non-polluting vehicles, and extensive research and development on all aspects of this technology over the past fifteen years has delivered prototype cars with impressive performances. But taking the step towards successful commercialization requires oxygen reduction electrocatalysts—crucial components at the heart of fuel cells—that meet exacting performance targets. In addition, these catalyst systems will need to be highly durable, fault-tolerant and amenable to high-volume production with high yields and exceptional quality. Not all the catalyst approaches currently being pursued will meet those demands.

The current performances of the small test fleets of vehicles powered by automotive fuel cells are impressive, reflecting 15 years of intense development of all aspects of proton exchange membrane (PEM) fuel cells that have brought the technology close to pre-commercial viability[1]. But to move towards a genuinely practical technology that can be mass-produced cost-effectively, important further improvements are needed. This calls for a critical look at how we need to develop key components determining fuel-cell performance, durability and cost. A pivotal component is the electrocatalyst system that underpins fuel-cell operation, and excellent reviews of progress made in Pt-based fuel-cell catalyst development for automotive applications have been written from both an academic perspective focused on fundamentals and from a perspective focused on the requirements of the automotive companies[2–6]. This review is provided from the perspective of a fuel-cell component supplier who needs to consider all factors that any electrocatalyst approach will need to meet if it is to be commercially successful. Following an overview of fuel cells and the challenges they need to meet for commercialization, I will consider the electrocatalyst system and the different approaches taken to ensure its performance meets automotive fuel-cell requirements. In my view, focusing only on catalytic activity targets will not be sufficient to meet the challenge posed by large-scale automotive fuel-cell commercialization, which requires the manufacture of catalyst electrodes at high rates, high quality and low costs.

Fuel-cell components

An automotive fuel cell produces electricity from the electrochemical oxidation of hydrogen. It is manufactured as a stack of identical repeating unit cells comprising a membrane electrode assembly (MEA) in which hydrogen gas (H_2) is oxidized on the anode and oxygen gas (O_2) is reduced on the MEA cathode, all compressed by bi-polar plates that introduce gaseous reactants and coolants to the MEA and harvest the electric current (Fig. 1). The electrochemical reactions occur in the MEA electrodes, each attached to a solid polymer ion exchange membrane that conducts protons but not electrons. The cathode oxygen reduction reaction (ORR) and anode hydrogen oxidation reaction both occur on the surfaces of platinum (Pt)-based catalysts. Pure water and heat are the only byproducts. Porous gas diffusion layers transport reactants and product water between the flow fields and catalyst surfaces while exchanging electrons between them.

The challenges for commercialization

Fuel-cell MEAs must meet three major criteria: cost, performance and durability. The cathode ORR is six or more orders of magnitude slower than the anode hydrogen oxidation reaction and thus limits performance, so almost all research and development focuses on improving the cathode catalysts and electrodes. Most MEA catalysts used today are based on Pt (in the form of nanoparticles dispersed on carbon black supports), with the high price of this scarce precious metal having a decisive impact on costs. Fuel-cell vehicles in the test fleets monitored by the United States Department of Energy (DOE) have used 0.4 mg of Pt per square centimetre (mg Pt cm^{-2}) or more on the cathode, and in these vehicles the catalyst/MEA stability has still been short of the 5,000-hour durability target[7]. How to reduce costs by reducing cathode loadings to $<$0.1 mg Pt cm^{-2} without loss of performance or durability is the subject of most electrocatalyst research. Current US DOE 2017 targets[1] for electrocatalysts aim for a total (anode + cathode) platinum group metals (PGM) loading of 0.125 mg cm^{-2} on MEAs able to produce rated stack

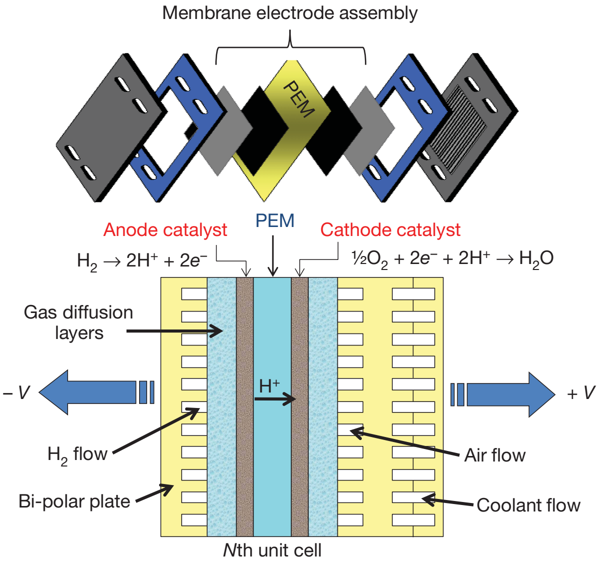


Figure 1 | Fuel-cell components. Unit cell cross-section of the *N*th unit cell in a fuel-cell stack, showing the components of an expanded MEA.

[1] 3M Fuel Cell Components Program, 3M Center, St Paul, Minnesota 55144, USA.

Table 1 | Development criteria for automotive fuel-cell electrocatalysts

Performance	• Must meet beginning-of-life performance targets at full and quarter power. • Must meet end-of-life performance targets after 5,000 h or 10 years operation. • Must meet performance, durability and cost targets and have less than 0.125 mg PGM per cm^2. • Corrosion resistance of both Pt and the support must withstand tens of thousands of start-up/shut-down events. • Must have low sensitivity to wide changes in relative humidity. • Must withstand hundreds of thousands of load cycles. • Must have adequate cool start, cold start and freeze tolerance. • Must enable rapid break-in and conditioning (the period needed to achieve peak performance).
Materials	• Must have high robustness, meaning tolerance of off-nominal conditions and extreme-load transient events. • Must produce minimal H_2O_2 production from incomplete ORR. • Must have high tolerance to external and internal impurities (for example, Cl^-) and ability to fully recover. • Must have statistically significant durability, meaning individual MEA lifetimes must enable over 99.9% of stacks to reach 5,000-hour lifetimes. • Electrodes must be designed for cost-effective Pt recycling. • Environmental impact of manufacturing should be minimal at hundreds of millions of square metres per year.
Process	• Environmental impact must be low over the total life-cycle of the MEAs. • Manufacturing rates will need to approach several MEAs per second. • MEA manufacturing quality must achieve over 99.9% failure-free stacks at beginning of life (one faulty MEA in 30,000 for just 1% stack failures). • Proven high-volume manufacturing methods and infrastructure will be required. • Catalyst-independent processes will be preferred, to enable easy insertion of new-generation materials.

power densities of 8.0 kW g^{-1} Pt. This gives 8 g of PGM per vehicle, similar to what is in an internal combustion engine today.

Vehicle operation imposes severe durability and performance constraints on the fuel-cell cathode electrocatalysts[1] beyond the fundamental requirement for high ORR activity. System integrators require that the MEA produce at least 0.6 V at 1.5–2 A cm^{-2} owing to radiator size and related cooling constraints. Catalysts must survive hundreds of thousands of load cycles and tens of thousands of start-up and shut-down events over the 5,000-hour lifetime of the stack[1]. Although durability is beyond the scope of this review, serious degradation is associated with the tendency for the cathode to reach potentials above the onset of oxidation of carbon in contact with Pt, during even the short times when H_2/air waves move through the flow fields during start-up or shut-down[8]. Some of the countermeasures being developed are using more stable graphitized carbon, using catalyst supports that will not electrochemically corrode, and adding oxygen evolution catalysts to the mix to clamp the potentials at the start of water oxidation[9,10]. Table 1 summarizes these major catalyst requirements, along with secondary criteria and manufacturing considerations.

Some fundamental electrochemical considerations

MEA performance is reflected in its polarization curve, a plot of cell voltage versus current density (Fig. 2). As current is drawn from a cell, its voltage decreases as a result of three primary sources of power loss: ORR kinetic losses of the cathode, current times resistance (*iR*) losses due to material and interface resistances, and mass transport overpotentials at high current densities when it is difficult for the catalyst to get enough oxygen from air. For the polarization curve conditions in Fig. 2, the theoretical open-circuit voltage (zero current) is 1.169 V; note that measured open-circuit voltages are lower than theoretical ones owing to imperfect separation of the gases by the membrane and its finite electronic resistance. Increases in load current cause the cell voltage to decrease logarithmically owing to kinetic losses (green line in Fig. 2), with the large cell-voltage loss of about 370 mV in going from open-circuit voltage to practical currents of just 0.1 A cm^{-2} reflecting the sluggish ORR kinetics on Pt. The measured cell resistance multiplied by current density gives the *iR* losses, which can be added to the measured polarization curve (blue symbols) to give the *iR*-free curve (red symbols). The remaining difference between the *iR*-free curve and the ideal kinetic line represents the sum of all mass transport losses. In the range of practical current densities, 0.1–2.5 A cm^{-2}, improvements in MEA resistance can have a much larger impact on actual cell voltage than improvements in kinetics. Cell resistances have therefore been researched, and have been reduced nearly as much as is possible using current membranes and gas diffusion media. Kinetic losses are more challenging because an order-of-magnitude improvement in ORR activity would gain only 60–70 mV, and progress in catalyst development so far has achieved only modest cell voltage gains of tens of millivolts. Reducing mass transport overpotentials by the same amount is less difficult.

Targeted catalyst development benefits from a detailed understanding of the metal-catalysed electrochemical reduction of oxygen to water, $O_2 + 4H^+ + 4e^- \rightarrow 2H_2O$, which is mechanistically complicated. It is usually thought to involve different reaction pathways such as direct $4e^-$ reduction of adsorbed oxygen to water; or a $2e^-$ reduction to adsorbed H_2O_2 that then either desorbs or undergoes a second $2e^-$ reduction to water[11]. Irrespective of mechanistic detail, the kinetic current density *i*, normalized by the surface area of the Pt electrode and generated at a potential *E*, has been proposed[12] to be a function of the Gibbs energy of adsorption ΔG_{ad}

$$i = nFKc_{\mathrm{O_2}}(1 - \Theta_{\mathrm{ad}})^x \exp\left(-\frac{\beta FE}{RT}\right) \exp\left(-\frac{\gamma \Delta G_{\mathrm{ad}}}{RT}\right) \qquad (1)$$

where n, F, K, x, β, γ and R are constants, $c_{\mathrm{O_2}}$ is the molecular oxygen concentration, and Θ_{ad} the fraction of electrode surface sites covered with adsorbates. A key assumption in the development of the $(1 - \Theta_{\mathrm{ad}})$ pre-exponential factor is that the ORR rate-limiting step is the first electron transfer step, $\mathrm{Pt(O_2)_{ad}} + e^- \rightarrow \mathrm{Pt(O_2{}^-)_{ad}}$, with

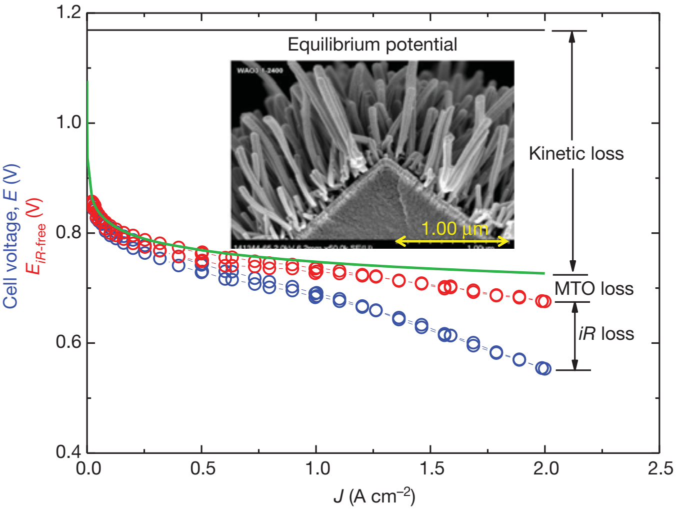


Figure 2 | Fuel-cell polarization curve. Measured PEM fuel-cell MEA polarization curve (blue) and *iR*-free (red) compared to a hypothetical curve for kinetic losses only (green). The difference gives the losses due to mass transport overpotentials (MTO). The polarization curve is from an MEA having electrodes based on the NSTF PtCoMn catalyst (inset) under 150 kPa H_2/air. Other conditions are given in ref. 23.

oxygen adsorption onto a surface primarily covered by impurities (hydroxyls) rather than reactive intermediates. This suggests that anything done to the surface's atomic or electronic structure that delays hydroxyls adsorbing and blocking O_2 adsorption sites will have a positive impact on the specific activity, because the $(1 - \Theta_{ad})$ term in equation (1) will be larger. This model underpins many approaches used to improve ORR electrocatalysts, although some recent model calculations and measurements are consistent with adsorbed oxygen or hydroxyl species playing a more active part as intermediates[6].

Electrocatalyst properties determining the observed current densities J (in A per planar cm^2) in MEAs are the electrode's electrochemically active surface area S (in cm^2 of Pt per planar cm^2), the kinetic current density i as in equation (1) that reflects the catalyst's activity, and a recently proposed[13] collision frequency scaling factor $f(\lambda, \rho_S)$ that depends on the mean free path length λ above the catalyst surface and on a spatial distribution function of surface area that to first order can be approximated by the catalyst surface area volumetric density ρ_S (in cm^2 Pt per cm^3). Or, in brief, $J = f \cdot S \cdot i$. The value of S is generally determined experimentally using cyclic voltammograms (see Box 1, which also provides catalyst activity definitions and DOE targets) and normalized by the electrode surface area to give the surface-area enhancement factor (in cm^2 of Pt per planar cm^2). Kinetic activity is measured *ex situ* in rotating disc electrode (RDE) apparatus[3,14,15] or *in situ* in fuel cells, with both methods requiring care to obtain reproducible results[3]. The two methods sample different surface states (that is, different amounts of hydroxyl or oxygen adsorption on platinum), so comparisons of RDE and fuel-cell activities are not always straightforward. The factor $f(\lambda,\rho_S)$, calculated from catalyst electrode physical properties, captures one way in which geometry has a differentiating role in comparing different electrocatalyst designs.

The basic electrocatalyst designs

PEM fuel-cell electrocatalyst technology has relied almost exclusively on either Pt blacks (metal particles so tiny that they absorb light very well and appear black, having a high surface to volume ratio, ideal for catalysts) or Pt nanoparticles, 2–5 nm in size, dispersed onto larger carbon black particles. Neither will meet the DOE 2017 performance and durability targets at PGM loadings that meet the cost targets. But our fundamental understanding of what controls catalytic activity has dramatically improved in the last few years and is now guiding next-generation catalyst development. Most emerging approaches focus on controlling the surface structure and composition of catalytic nanoparticles to achieve higher ORR activity with less Pt, and new synthetic routes have delivered such 'designer nanoparticles' that meet or exceed the DOE 2017 targets for ORR activity. However, the newest approaches are barely out of the 'test-tube' stage: they have not yet been tested extensively in actual fuel-cell MEAs, and it remains to be seen which approaches can also meet the other practical MEA requirements (see Table 1).

The basic designs for platinum catalysts are summarized in Fig. 3, categorized by overall geometry of the catalyst and its support and then further subdivided according to structural morphology and composition. This approach illustrates that kinetic activity can change by nearly an order of magnitude when the catalyst is a discrete nanoparticle or a polycrystalline thin film, and that catalyst surface area per unit volume can affect the maximum achievable current density. Also, the volume occupied by the non-active support influences current density[13], while aspect ratios determine the packing of the catalyst supports and hence porosity and free-radical scavenging.

BOX 1

Definitions and activity targets

■ The surface-area enhancement factor is the Pt catalyst surface area S measured by the charge generated from an under-potential deposited monolayer of hydrogen atoms on the Pt catalyst surface divided by the planar area of the sample (cm^2 Pt per planar cm^2).
■ Pt loading is the number of mg of Pt per planar cm^2 in an MEA electrode layer.
■ Absolute ORR kinetic activity is currently defined as the current density measured at 900 mV under one atmosphere of fully saturated pure oxygen, at 80 °C. For an MEA this means 150 kPa absolute pressure, due to 50 kPa of water vapour.
■ The area-specific activity A_s (A per cm^2 Pt) is determined by dividing the absolute activity by the surface-area enhancement factor.
■ The mass activity A_m (A per mg of Pt) is determined by dividing the absolute activity by the Pt loading.
■ The DOE's 2017 target for A_m is 0.44 A per mg Pt and its 2015 target for A_s is 0.7 mA per cm^2 Pt (ref. 1).

Extended surface area catalysts

Extended surface area (ESA) catalysts comprise large area surfaces extended in two dimensions like thin films. Their advantages over conventional Pt/C approaches are larger radii of curvature that make them more resistant to surface area loss via mechanisms such as Pt dissolution and redeposition, reduced mass transport losses, and in some cases the potential to eliminate corrosion of the catalyst support and simpler manufacturing. The area-specific ORR activities (A_s) of ESA catalysts are about ten times higher than nanoparticle-on-carbon catalysts, owing to electronic structure properties of thin films versus nanoparticles. Classic ESA examples are polycrystalline or single-crystalline bulk catalyst surfaces[16–20]; developed at Lawrence Berkeley and Argonne National Laboratory, these serve as valuable model systems that have shown the effect of structural aspects like single-crystal facet orientation, size, type and composition on activity. Some advanced ESA alloy catalysts exhibit a roughened 'Pt-skeleton' surface with Pt atoms covering the average bulk composition, or a highly coordinated pure 'Pt-skin' overlying a modulated surface structure in which the underlying three layers' composition oscillates towards the bulk composition[16,20]. Such a Pt-skin surface structure on Pt_3Ni_1 {111} was reported to have the world-record A_s of 18 mA per cm^2 Pt (non-*iR* corrected), about 90 times higher than a standard commercial dispersed Pt/C in comparative RDE measurements[16]. The peculiar arrangement of Pt and Ni atoms in the top three layers of this system causes different amounts of shift in the *d*-band centre relative to the Fermi level of the topmost layer of Pt atoms, which is believed to affect the adsorption coverage of hydroxyl spectator species that interfere with ORR (see equation (1)). The quest is how to achieve this in a practical catalyst.

Nanostructured thin-film (NSTF) catalysts are the only practical ESA catalysts found so far[13,21–24]. The support is a thin monolayer of an oriented array of crystalline organic whiskers, less than 1 μm tall and 30 nm × 55 nm in cross-section. It is applied to a roll-good substrate (a material made by a roll-to-roll process) with a number density of 30–40 whiskers per square micrometre, and then magnetron sputter-coated with catalyst thin films of choice (Fig. 2, inset). The organic support whisker is non-conductive, does not support corrosion currents and its body-centred cubic crystalline nature influences the subsequent nucleation and growth of the catalyst thin films coating them[25]. NSTF electrocatalysts have documented basic advantages in A_s of the catalyst[23,26], surface area utilization and stability[27,28], performance with ultralow PGM loadings[24], support corrosion resistance[29], high-volume manufacturability[30], and their release of F^- ions from the ionomer (perfluorosulphonic acid) used in the membrane or electrode is up to three orders of magnitude lower[24,31]. (Incomplete oxygen reduction to water on Pt produces H_2O_2, which creates free radicals that attack the membrane ionomer. Membrane degradation rates are proportional to the F^- ion release rates, with suppressed F^- release showing that the unique geometry of NSTF catalysts facilitates scavenging of radicals before they reach the membrane.) The low-volume support makes NSTF electrodes 10 to 20 times thinner than equivalently loaded Pt/C dispersed electrodes, which means better access to all of the catalyst surface area at all current densities. It also means less volume for

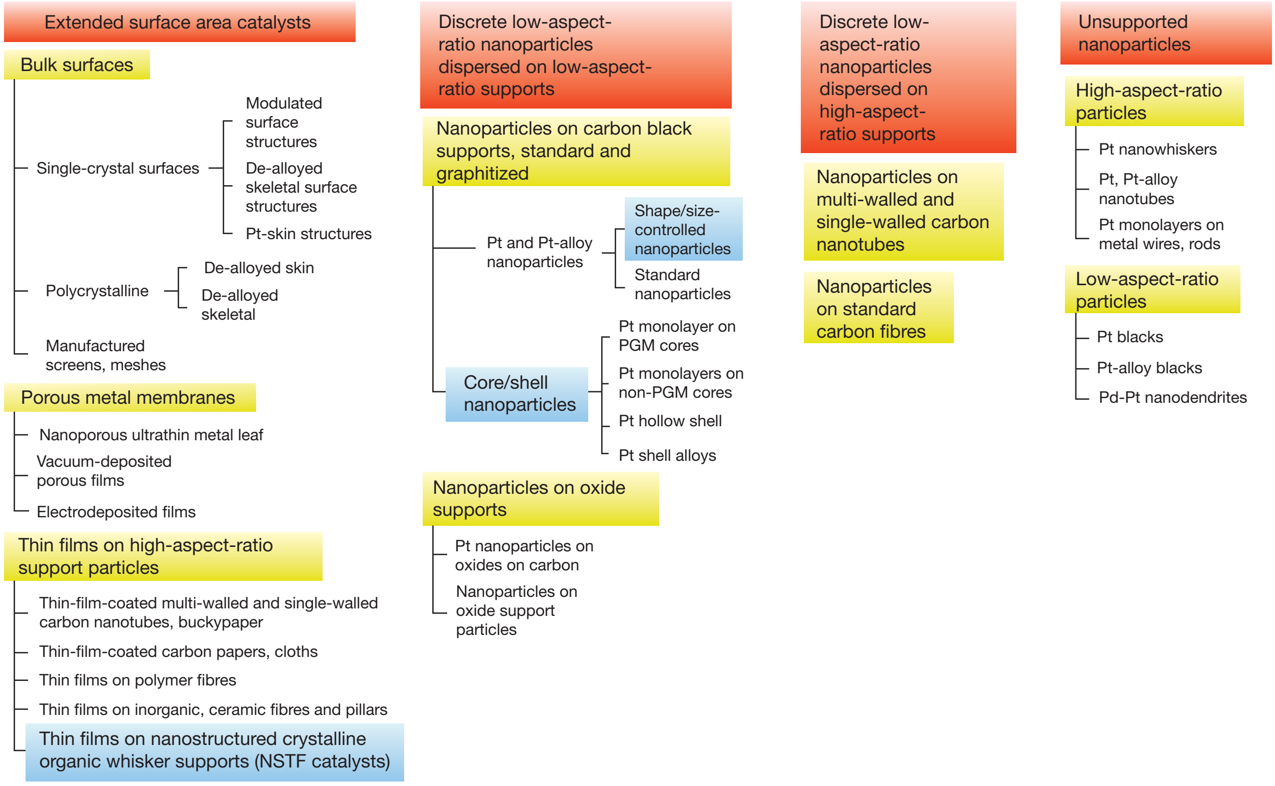

Extended surface area catalysts
Bulk surfaces
Single-crystal surfaces
Modulated surface structures
De-alloyed skeletal surface structures
Pt-skin structures
Polycrystalline
De-alloyed skin
De-alloyed skeletal
Manufactured screens, meshes
Porous metal membranes
Nanoporous ultrathin metal leaf
Vacuum-deposited porous films
Electrodeposited films
Thin films on high-aspect-ratio support particles
Thin-film-coated multi-walled and single-walled carbon nanotubes, buckypaper
Thin-film-coated carbon papers, cloths
Thin films on polymer fibres
Thin films on inorganic, ceramic fibres and pillars
Thin films on nanostructured crystalline organic whisker supports (NSTF catalysts)
Discrete low-aspect-ratio nanoparticles dispersed on low-aspect-ratio supports
Nanoparticles on carbon black supports, standard and graphitized
Pt and Pt-alloy nanoparticles
Shape/size-controlled nanoparticles
Standard nanoparticles
Core/shell nanoparticles
Pt monolayer on PGM cores
Pt monolayers on non-PGM cores
Pt hollow shell
Pt shell alloys
Nanoparticles on oxide supports
Pt nanoparticles on oxides on carbon
Nanoparticles on oxide support particles
Discrete low-aspect-ratio nanoparticles dispersed on high-aspect-ratio supports
Nanoparticles on multi-walled and single-walled carbon nanotubes
Nanoparticles on standard carbon fibres
Unsupported nanoparticles
High-aspect-ratio particles
Pt nanowhiskers
Pt, Pt-alloy nanotubes
Pt monolayers on metal wires, rods
Low-aspect-ratio particles
Pt blacks
Pt-alloy blacks
Pd-Pt nanodendrites

Figure 3 | Basic platinum-based heterogeneous electrocatalyst approaches. The four PEM fuel-cell electrocatalyst approaches (developed or under investigation) for the performance-limiting cathode ORR are shown, with Pt and Pt-alloy electrocatalysts listed according to the basic geometric structure of the catalyst particles and their supports. The main subcategories are highlighted in yellow. Catalyst approaches with the highest demonstrated activities are highlighted in blue.

storing any condensed product water that can lead to challenges in freeze-start, load transients (such as engine idling to full power as fast as possible) and cold operation, requiring different water-management strategies, anode gas diffusion layers, or modifications of operating protocols[32]. NSTF Pt alloys show the same activity gains over pure Pt as dispersed nanoparticles. De-alloying (see Box 2 for details) can endow the NSTF catalyst coating with core–shell properties as well, with sublayers affecting the lattice properties of the outermost Pt surface. The NSTF alloy catalyst with the largest production to date uses the $Pt_{68}Co_{29}Mn_3$ composition, which for loadings of 0.05–0.15 mg cm^{-2} delivers A_s values of 1.3–1.8 mA cm^{-2} Pt, and mass activities (A_m) of 0.15–0.17 A cm^{-2} Pt. Still under development are NSTF $Pt_{1-x}Ni_x$ alloy catalysts with an unusually sharply peaked gain in activity for $x = 0.7$ (ref. 33). They reach $A_s \approx 2.4$ mA cm^{-2} Pt and A_m from 0.24 to over 0.5 A mg^{-1} Pt in 50-cm^2 fuel-cell measurements, with loadings below 0.15 mg Pt cm^{-2} (refs 23, 24).

In related approaches, thin catalyst films are applied to single- or multi-wall carbon nanotubes[34,35] to achieve higher activity on a more durable support. These systems are more durable[35] than catalysts on high-surface-area carbons, but performances and activities have not improved significantly and the nanotubes will still ultimately corrode at high potentials.

Porous metal membranes[36–40], the third class of extended surface area catalysts, include nanoporous ultrathin metal leaves and vacuum-deposited, electrodeposited or laser-deposited porous films. They can be de-alloyed (see Box 2 for more details) to create nanoporosity, or to obtain specific modulated surface layer compositions similar to that of the Pt_3Ni_1 {111} system. These catalysts show high specific A_s values similar to those of polycrystalline bulk surfaces but with higher surface area, and enable valuable studies of the de-alloying processes, but are mostly not amenable to high-volume manufacturing.

Nanoparticles on low-aspect-ratio supports

This catalyst category is dominated by roughly spherical Pt or Pt alloy nanoparticles dispersed on standard or graphitized carbon black support particles, including conventional homogeneous Pt and Pt transition-metal alloy nanoparticles[3,19,41–43] and designer nanoparticles with size, shape and radial composition controlled to increase activity and reduce Pt.

BOX 2

De-alloying of Pt transition-metal alloys

The de-alloying of less-noble elements from Pt-alloys is a key strategy for creating Pt-based catalysts in all design categories. This means less stable elements initially alloyed with Pt (usually at a high atomic percentage of transition metal) are intentionally or unintentionally dissolved out to leave a nanoporous film, or skeletal surface or core–shell particle configuration[37,38,88] overlying a composition closer to a stable bulk alloy composition such as Pt_3M_1, where M is a transition metal. The process increases surface area, and creates in the outer few layers a modulated surface composition that leaves the surface Pt lattice contracted or its electronic structure favourably modified, as illustrated by lattice-strain control of the Pt shells formed around de-alloyed cores[94]. In the case of $Pt_{1-x}Ni_x$ bulk alloys with high Ni content, electrochemical de-alloying provide a means of studying[95–97] the evolution of nanoporosity. The resultant material, with average pore and ligament sizes of 3–5 nm and filled with hydrophobic protic liquids, displayed $S_m = 44$ m^2 per g Pt, and $A_m = 0.4$ A per mg Pt[95]. Strasser and colleagues have pioneered voltammetric de-alloying of electrode nanoparticles, composed of base-metal-rich bimetallic and trimetallic alloys[88,91,92,94], directly in the MEA and reported A_s and A_m values exceed the DOE 2017 and 2015 targets (Box 1). Although such *in situ* methods are not practical, understanding the de-alloying process is critical for optimizing the catalyst properties and *ex situ* de-alloying of these alloy systems is under investigation[93].

The state-of-the-art of conventional Pt and Pt-alloy/C electrocatalysts in dominant use today consist of Pt nanoparticles with diameters of 2–4 nm, dispersed onto larger high-surface-area carbon black support particles at Pt/C weight percentages of 20% to 60%. Commercial pure Pt/C catalysts have surface areas of 80–120 $\mathrm{m^2\,g^{-1}}$ Pt, specific activities of 0.15–0.2 $\mathrm{mA\,cm^{-2}}$ Pt and mass activities of 0.1 to 0.12 $\mathrm{A\,mg^{-1}}$ Pt measured in MEAs. Homogeneous Pt-alloy nanoparticles on carbon have historically been observed to increase ORR activity over pure Pt/C by about a factor of 2 to 2.5. The commercially available, heat-treated 30 weight per cent PtCo/C system routinely provides MEA measurement values of specific and mass activity in our 3M laboratory that are close to the DOE targets, for example, 1.2 $\mathrm{mA\,cm^{-2}}$ Pt and 0.39 $\mathrm{A\,mg^{-1}}$ Pt. For a wide range of carbon-based supports, initial Pt surface areas range from 20 to over 70 $\mathrm{m^2\,g^{-1}}$ Pt, but converge after stability testing to 20–30 $\mathrm{m^2\,g^{-1}}$ Pt (ref. 44).

Uniformly sized octahedra, cubes or other polyfaceted shapes identically terminated with {111} and {100} facets[45–48]can be produced, often using capping agents selectively to control facet growth rates. Truncated-octahedral Pt_3Ni particles with predominantly {111} facets show A_m values up to four times those of commercial Pt/C (ref. 46), and much higher than {100} bounded cubes[47]. Surface-specific activity strongly depends on the fraction of {111} surface exposed and is about a tenth that of bulk single-crystal Pt_3Ni\{111\} surfaces, suggesting that larger shape-controlled particles with higher surface coordination (fewer surface defects where oxides preferentially form) or compositional gradients could improve activities[49]. In the case of monodispersed Pt_3Co nanoparticles, these factors increase grain size and A_s monotonically for particle sizes of 3–9 nm, while A_m peaks at about 4.5 nm for optimum particle annealing at 500 °C (refs 45, 50). Monodispersed and homogeneous $Pt_{1-x}Ni_x$ nanoparticles with controlled Ni depletion from the outer surface layers exhibit a Pt-skeleton-type surface structure, with the improvement factor over Pt/C peaking at 4.5 nm for monodispersed PtNi (ref. 51). For similarly structured PtCoNi alloy particles, mass activities by RDE are reported to exceed 2.5 $\mathrm{A\,mg^{-1}}$ Pt (ref. 52).

Core–shell nanoparticle electrocatalysts pioneered at Brookhaven National Laboratory constitute a highly promising subcategory of catalysts[52–64]. These systems exhibit higher mass activities because Pt is eliminated from the core of the catalyst particles, and higher specific activities because the core material influences the outer Pt monolayer and optimizes its surface electronic and structural properties. Examples include Pt monolayers on Pd and non-PGM cores, de-alloyed cores, 'Swiss-Pt' de-alloyed cores (full of pores), and hollow Pt or Pt-alloy shells[53,54,57–63]. The core material would ideally consist of non-PGM elements, though the successful Pt-monolayer core–shell catalysts so far have cores containing Pd, Au-Ni, Pd-Co, Pd_3Co, Pd_3Fe, Pd-Ir, Ir, Pd-Au, $AuNi_{0.5}Fe$, Pd-Nb, Pd-V, Pd-W and Ru monolayers on Pd cores. Fuel-cell MEA measurements with these novel catalysts at two industrial locations yield mass activities lower than the RDE values measured by the Brookhaven National Laboratory group[59,64], which may be due to non-optimized electrodes. Key development opportunities for this subcategory include development of scalable synthetic routes for generating 'pin-hole'-free Pt monolayers to protect the non-PGM core material from leaching; increasing specific activity while using less expensive core materials and optimizing electrode ink formulations; and further increased resistance to Pt dissolution due to repeated start-stop induced voltage cycling events.

Support corrosion is a problem for all these systems, pushing development towards dispersing the nanoparticle catalysts on graphitized, lower-surface-area carbons or single-wall or multi-wall carbon nanotubes (see below). Another strategy uses inherently more stable inorganic oxides as support. If these provide adequate conductivity, for example, where the oxide support is formed on carbon black, promising mass activities are obtained: 0.2 $\mathrm{A\,mg^{-1}}$ Pt for Pt/C, 0.3 $\mathrm{A\,mg^{-1}}$ Pt for Pt-$TaOPO_4$/C and 0.45 $\mathrm{A\,mg^{-1}}$ Pt for heat-treated Pt-$TaOPO_4$/C (where C = Vulcan carbon)[65].

Nanoparticles on high-aspect-ratio supports

This category includes Pt or Pt-alloy nanoparticles dispersed on standard carbon fibres as well as on single- or multi-walled carbon nanotubes[66]. As in the case of thin-film catalysts coated onto these types of supports, the objective is usually improved durability because the activity of the nanoparticles is not expected to change much with the aspect ratio of the support. In the case of carbon nanotube supports, their higher electrical conductivity is not really any advantage because electronic impedance of the electrodes is a very small contributor to the overall overpotential losses, but carbon nanotube supports can improve transport or water management[67]. However, concerns about the possible adverse health effects of carbon nanotubes without the counterweight of a significant functional performance or processing advantage may make them unattractive for high-volume electrode manufacturing.

Unsupported nanoparticles

Unsupported nanoparticle catalysts include the traditional low-aspect-ratio particles of Pt and Pt alloy blacks[41], and new concepts that use Pt and Pt-alloys in the form of high-aspect-ratio structures such as nanotubes[68,69]and Pt monolayers on Pd nanowires and nanorods[70,71]. Although they are just at the initial-concept phase, the unsupported metal nanotube catalysts are claimed to improve A_s values over bulk polycrystalline Pt by a factor of eight (ref. 69), and the monolayer systems are claimed to exhibit impressive mass activities of 0.4 $\mathrm{A\,mg^{-1}}$ PGM[70,71](measured by RDE).

Pt-free electrocatalysts

Owing to the scarcity and high cost of Pt, Pd-based catalysts (recently reviewed by Shao[72]) and Pd-based transition metal alloy catalysts (explored in the Myers group[73]) have been considered. But the best activities reported for Pd are barely equivalent to typical Pt/C, with the best use of Pd appearing to be in combination with Pt as in the core–shell configurations discussed above. In any case, replacement of most of the Pt by Pd may not significantly address the cost issue, owing to demand/price fluctuations[2].

Non-precious metal (NPM) catalysts eliminate PGMs completely to reduce costs, and have recently shown dramatic ORR activity improvements. Among the many approaches being explored[74–83], pyrolysis of cobalt- and iron-containing heteroatom polymer precursors (for example, porphyrins) has been the dominant route towards NPM catalysts. Active sites are believed to comprise metals coordinated to several nitrogen atoms, MN_x (where M = Co, Fe). The Dodelet group's impregnation and pyrolysis process combining Fe precursors and a nitrogen precursor has led to the largest beginning-of-life activity increases[79,81]. The nature of the active site in these NPM catalysts is still debated, as is the question of whether Fe is associated with the active site or with the means of creating such a site. There is no way to measure area density of active sites as for Pt-based catalysts, so kinetic activity is reported as a volumetric current density. Kinetic activities measured at 0.8 V using pure O_2 have rapidly increased from 2.7 $\mathrm{A\,cm^{-3}}$ in 2008 (ref. 82), to 99 $\mathrm{A\,cm^{-3}}$ in 2009 (ref. 79), and 230 $\mathrm{A\,cm^{-3}}$ in 2011 (ref. 81). The most recent increase was due to a large reduction in mass transport loss: the performance of a new NPM Fe-based cathode on Nafion 117 ionomer under 2-bar gauge H_2/air was 0.25 $\mathrm{A\,cm^{-2}}$ at 0.5 V after 100 hours, about a tenth of the current density obtained with Pt-based cathodes. The key issues for NPM catalysts are mass transport losses and stability. A polyaniline FeCo-C catalyst with relatively high durability (700 hours) at 0.4 V in a H_2/air fuel cell and a peak power density of 0.55 $\mathrm{W\,cm^{-2}}$ at 0.38 V under pressurized pure oxygen has been reported[83]. But NPM catalyst durability is worse at higher potentials—and for automotive purposes, performances at less than 0.6 V are largely irrelevant.

Comparison of kinetic activities

As illustrated by the preceding discussion, a wide range of catalyst systems has been explored and Fig. 4 shows composite plots of A_m versus

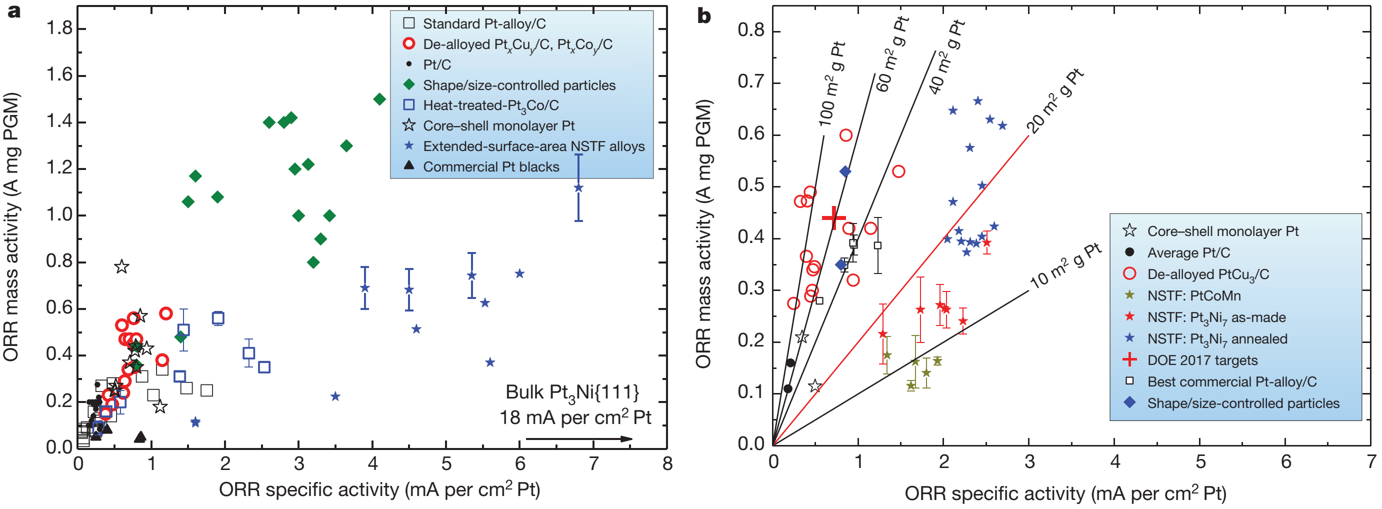


Figure 4 | Kinetic activities of the main Pt-based electrocatalyst systems. The ORR A_m versus A_s are shown for the major Pt-based electrocatalyst approaches listed in Fig. 3. **a**, Activities are measured by RDE at 900 mV for the following catalysts: standard Pt-alloys/C (refs 42, 43, 87), de-alloyed PtM$_3$/C (where M = Cu, Co) (refs 88–90), Pt/C (refs 3, 42, 44, 46–48, 50), shape- and size-controlled particles (refs 45, 47, 48, 50, 51), heat-treated Pt$_3$Co/C (ref. 42), core–shell monolayer Pt (refs 44, 56, 60, 62–64, 71), extended-surface-area NSTF alloys (refs 23,26) and commercial Pt blacks (refs 3, 48). **b**, Activities are measured in MEAs at 900 mV, 80 °C and 150 kPa saturated O_2 for the following catalysts: core–shell Pt monolayer (refs 49, 64), average Pt/C (refs 3, 91, 92), de-alloyed PtCu$_3$/C (refs 91–93), three extended-surface-area NSTF alloys (ref. 23), the DOE 2017 and 2015 targets, best commercial Pt-alloys/C (data from GM, 3M), and shape- and size-controlled particles (refs 46, 52). The scattering of activity values for any one type or reference represent different catalyst compositions, loadings or preparation and process treatments, not statistical variations in measurement.

A_s measured by both RDE and in MEAs for the most promising ones. RDE values are generally larger than those measured in MEAs, reflecting not only the more complex environment of the catalyst in a working fuel cell but also the simple difference in protocol that allows RDE measurements on a clean surface while MEA values are obtained with oxidized surfaces. (Note that RDE measurements will also depend on temperature, voltage scan rate, *iR* and diffusion-correction factors.) For standard Pt/C or Pt alloy/C catalysts, only a few data points are given to represent generally accepted measured values. The figures show that several catalyst systems seem to be able to meet the DOE 2017 A_m target of 0.44 A mg^{-1} Pt; some systems come very close, including the best commercial Pt-alloy/C catalyst, for which durability is still an issue. The highest-performing systems in Fig. 4 are not yet practical, in that they are not able to simultaneously satisfy all the electrocatalyst requirements for high current density and durability at low loadings.

Manufacturing and scalability

Manufacturability and quality control of MEAs produced at the rates ultimately required for true commercialization have not been an issue for the small fuel-cell vehicle fleets produced to date. But the lead time to get an MEA component qualified for insertion into a vehicle stack is about five years, and publicly announced future emergent fuel-cell vehicle volumes are tens to hundreds of thousands of vehicles per year by 2020. Clearly, the scalability of any new catalyst approach now has to be seriously considered from the outset. For many of the new catalyst approaches being evaluated at the ‘test-tube’ stage, it is not clear that the processes used to generate them would be scalable with the quality levels required. The DOE cost targets are based on 500,000 vehicles per year, and a particular MEA technology advanced to that emergent market level will not be easily discarded in favour of something totally new that may be better able to reach a more mature production level at say 10% of the world market in 2030 (or 15 million vehicles).

To illustrate the scales involved, consider that annual production of 15 million fuel-cell vehicles each with a stack containing 300 (versus the 400 required today) MEAs (each around 300 cm^2) would require 4.50 billion MEAs a year. With a production line at full capacity operating three shifts per day, 360 days per year, or about 8,000 hours per year with 80% average up-time to account for maintenance, repair and lot changes of input materials, the required production rate is about 11,700 MEAs per minute. To match car production at about one vehicle per minute per production line, or 300 MEAs per stack per minute, would require 20 full-capacity MEA production lines each producing 585 MEAs per minute or about 10 MEAs per second (with more production lines requiring more capital and operating costs). Individual piece part processing is out of the question.

Target loadings of about 0.1 mg Pt per cm^2 mean electrode thicknesses will be less than 2 μm, requiring precision coating methods with critical limits on debris and tolerances. High-volume roll-to-roll widths up to a metre in width should be possible, so even with ten MEAs across the web width, each measuring 10 cm × 30 cm, line speeds of 20 m min^{-1} will be required. These MEAs have to be made with extraordinary quality control (one fatal MEA defect in 30,000 MEAs for 1% stack failures). The catalyst and catalyst/membrane integration manufacturing processes have to be simple, robust, few in number and have wide process parameter windows, given that the yields per step multiply. Just four sequential process steps with 90% yields would increase costs by 30% without recycling. Process steps for electrode formation that require hot bonding, annealing, solvent evaporation or drying steps lasting for minutes will require proportionately long manufacturing lines. Build-up of residues on the coaters, 8-hour shift lengths, ability to handle jumbo roll-goods, safety and catalyst batch-size limitations will be factors affecting batch sizes, throughputs and labour costs. At loading targets of even 0.125 mg cm^{-2} on MEAs with 300 cm^2 active areas, the above line speeds require catalyst flow-through rates of 1.5 kg of Pt per hour—roughly US$2 million worth of Pt per day per manufacturing line, at a metal price of US$2,000 troy ounces. On-site recycling of scrap will probably be justified. Recycling of Pt, once it reaches the target stack loadings equivalent to the PGM amounts used in current internal combustion engines, will be justified by the same economic arguments currently used for recycling PGMs in catalytic converters. Ink mixing of dispersions for catalyst coating, if used, would have to keep pace and use chemistries compatible with coating line speeds and quality levels. Safety, environmental and cost control requirements will probably exclude coating processes using flammable solvents. I believe these requirements and the required production rates and levels of quality will ultimately result in catalyst layer coating by all-dry vacuum-coating methods, such as the method used to produce over 80% of the world’s multi-layer optical-film-coated glass that is used in low-emissivity windows and produced at a volume of 250 million square metres already in 2005 (ref. 84). All the above considerations regarding essential process requirements apply to catalyst-membrane integration and the fabrication and integration of other stack components as well.

Perspective

Several catalyst systems have ORR kinetics adequate for automotive applications, including commercially available heat-treated PtCo/C. But at the required low PGM loadings, none are able yet to deliver in a fuel-cell environment the necessary high power, durability or robustness. Recent progress with 'designer' catalyst particles and ESA catalysts suggests that they will reach over the next few years their maximum kinetic performance, something close to the specific activity of 18 mA per cm^2 Pt measured by RDE on bulk crystals of Pt_3Ni\{111\}. The consistent difference between RDE and MEA values (see Fig. 4) implies that this ultimate performance maximum will be about 6–9 mA per square centimetre of Pt in a fuel-cell environment, well above the DOE 2017 target. But as the discussion of additional practical requirements and manufacturing issues has shown, impressive kinetic activity will not suffice to make a catalyst system attractive for large-scale automotive MEA production. I therefore consider that realizing target ORR activities may no longer be the most important goal, despite being the overwhelming focus of many researchers in the field. Instead, we should aim to achieve (at the required low PGM loadings) the durability and power targets using something that can be manufactured at high volumes with the requisite quality, throughput and yields. This poses a considerable challenge to the community, particularly when considering that for most catalyst systems durability, high power performance, quality and yield decrease as loadings decrease.

Grounds for optimism are the recent pace and diversity in the development of interesting new approaches to PEM fuel-cell electrocatalysts that optimize ORR kinetic activities without sacrificing durability or cost. A few material and design concepts are guiding these efforts, namely the use of extended-surface-area catalyst geometries; the use of alloys, particularly with Ni; the synthesis of catalysts with large, highly coordinated \{111\} facets; and the use of de-alloyed or annealed surface structures with increased surface area and a modulated composition that improves the activity of the topmost Pt layer. The ideal catalyst would embody all of these improvement characteristics[22].

These recent developments illustrate that the level and quality of fundamental research in the field need to continue unabated. Particularly beneficial would be a clear understanding of surface area and activity loss mechanisms, and insight into the causes of durability loss[85,86] associated with externally and internally generated impurities. I also hope for future electrocatalyst research exploring how membrane and microporous materials in contact with the catalyst and off-nominal operating events affect performance. We also need to know whether there are any fundamental processing rate limitations on high-volume manufacturability. These are the issues that will arise when the rubber meets the road, and will dictate the ultimate costs and customer acceptance of fuel-cell vehicles. It has been eighteen years since the first PtCo/C catalyst was investigated, but it is still not generally accepted for use in current fuel-cell vehicles; it may take even longer to implement any of the newer catalyst approaches into realistic electrodes.

1. The. US Department of Energy (DOE). *Energy Efficiency and Renewable Energy* http://www.eere.energy.gov/hydrogenandfuelcells/mypp/pdfs/fuel_cells.pdf and the *US DRIVE Fuel Cell Technical Team Technology Roadmap* (revised 25 January 2012) www.uscar.org/guest/teams/17/Fuel-Cell-Tech-Team.
These websites define the most critical performance, durability and cost targets for the PEM fuel-cell MEA and each of its components, as well as stack and system requirements.
2. Wagner, F. T., Lakshmanan, B. & Mathias, M. F. Electrochemistry and the future of the automobile. *J. Phys. Chem. Lett.* **1,** 2204–2219 (2010).
3. Gasteiger, H., Kocha, S., Sompalli, B. & Wagner, F. Activity benchmarks and requirements for Pt, Pt-alloy, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl. Catal. B* **56,** 9–35 (2005).
This paper first defined and explained the ORR activity targets and requirements for the PEM fuel-cell cathodes, particularly for fuel-cell vehicles.
4. Markovic, N., Schmidt, T., Stamenkovic, V. & Ross, P. Oxygen reduction reaction on Pt and Pt bimetallic surfaces: a selective review. *Fuel Cells* **1,** 105–116 (2001).
5. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chem.* **1,** 37–46 (2009).
6. Greeley, J. *et al.* Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nature Chem.* **1,** 552–556 (2009).
7. Wipke, K, *et al. Controlled Hydrogen Fleet and Infrastructure Analysis: 2011 DOE Hydrogen Program Annual Merit Review and Peer Evaluation Meeting* http://www.hydrogen.energy.gov/pdfs/review11/tv001_wipke_2011_o.pdf (National Renewable Energy Laboratory, 2011).
8. Reiser, C. A. *et al.* A reverse-current decay mechanism for fuel cells. *Electrochem. Solid-State Lett.* **8,** A273 (2005).
This explains the basic mechanism by which fuel starvation or start-up and shut-down events in a PEM fuel cell can cause carbon corrosion on the cathode.
9. Atanasoska, L. L., Vernstrom, G. D., Haugen, G. M. & Atanasoski, R. T. Catalyst durability for fuel cells under start-up and shutdown conditions: evaluation of Ru and Ir sputter-deposited films on platinum in PEM environment. *ECS Trans.* **41,** 785–795 (2011).
10. Halalay, I. C. *et al.* Anode materials for mitigating hydrogen starvation effects in PEM fuel cells. *J. Electrochem. Soc.* **158,** B313–B321 (2011).
11. Sepa, D. B., Vojnovic, M. V. & Damjanovic, A. Reaction intermediates as a controlling factor in the kinetics and mechanism of oxygen reduction at platinum electrodes. *Electrochim. Acta* **26,** 781–793 (1981).
12. Markovic, N. M. & Ross, P. N. Surface science studies of model fuel cell electrocatalysts. *Surf. Sci. Rep.* **45,** 117–229 (2002).
13. Debe, M. K. Effect of electrode surface area distribution on high current density performance of PEM fuel cells. *J. Electrochem. Soc.* **159,** B54–B67 (2012).
14. Mayrhofer, K. J. J. *et al.* Measurement of oxygen reduction activities via the rotating disc electrode method: from Pt model surfaces to carbon-supported high surface area catalysts. *Electrochim. Acta* **53,** 3181–3188 (2008).
15. Garsany, Y., Barurina, O. A., Swider-Lyons, K. E. & Kocha, S. S. Experimental methods for quantifying the activity of platinum electrocatalysts for the oxygen reduction reaction. *Anal. Chem.* **82,** 6321–6328 (2010).
16. Stamenkovic, V. R. *et al.* Improved oxygen reduction activity on Pt_3Ni(111) via increased surface site availability. *Science* **315,** 493–497 (2007).
This paper showed that the fundamental kinetic activity for oxygen reduction on bulk Pt–Ni alloy surfaces could be nearly two orders of magnitude higher than the standard dispersed Pt on carbon.
17. Stamenkovic, V. R., Mun, B. S., Mayrhofer, K. J. J., Ross, P. N. & Markovic, N. M. Effect of surface composition on electronic structure, stability and electrocatalytic properties of Pt-transition metal alloys: Pt-skin versus Pt-skeleton surfaces. *J. Am. Chem. Soc.* **128,** 8813–8819 (2006).
This paper demonstrates the sensitivity and specificity of ORR activity to the fundamental surface structure and composition of the top few layers of Pt transition metal alloys.
18. Stamenkovic, V. R. *et al.* Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces. *Nature Mater.* **6,** 241–247 (2007).
19. Paulus, U. A. *et al.* Oxygen reduction on high surface area Pt-based alloy catalysts in comparison to well defined smooth bulk alloy electrodes. *Electrochim. Acta* **47,** 3787–3798 (2002).
20. Stamenković, V., Schmidt, T. J., Ross, P. N. & Markovic, N. M. Surface composition effects in electrocatalysis: kinetics of oxygen reduction on well-defined Pt_3Ni and Pt_3Co alloy surfaces. *J. Phys. Chem. B* **106,** 11970–11979 (2002).
21. Debe, M. K. in *Handbook of Fuel Cells—Fundamentals, Technology and Applications* (eds Vielstich, W., Lamm, A. & Gasteiger, H. A.) Ch. 45 (John Wiley & Sons, 2003).
22. Debe, M. K., Atanasoski, R. T. & Steinbach, A. J. Nanostructured thin film electrocatalysts—current status and future potential. *ECS Trans.* **41,** 937–954 (2011).
23. Debe, M. K. *2009–2011 Annual Merit Reviews DOE Hydrogen and Fuel Cells and Vehicle Technologies Programs: Advanced Cathode Catalysts and Supports for PEM Fuel Cells* http://www.hydrogen.energy.gov/pdfs/review11/fc001_debe_2011_o.pdf (DOE, 2011).
24. Debe, M. K. Nanostructured thin film electrocatalysts for PEM fuel cells—a tutorial on the fundamental characteristics and practical properties of NSTF catalysts. *ECS Trans.* **45** (2), 47–68 (2012).
This paper defines all the catalyst and MEA measured properties and published papers so far for the NSTF type catalyst electrodes.
25. Gancs, L., Kobayashi, T., Debe, M. K., Atanasoski, R. & Wieckowski, A. Crystallographic characteristics of nanostructured thin film fuel cell electrocatalysts—a HRTEM study. *Chem. Mater.* **20,** 2444–2454 (2008).
26. van. der Vliet, D. *et al.* Platinum-alloy nanostructured thin film catalysts for the oxygen reduction reaction. *Electrochim. Acta* **56,** 8695–8699 (2011).
27. Debe, M. K., Schmoeckel, A. K., Vernstrom, G. D. & Atanasoski, R. High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J. Power Sources* **161,** 1002–1011 (2006).
28. Debe, M. K., Steinbach, A. J. & Noda, K. Stop-start and high-current durability testing of nanostructured thin film catalysts for PEM fuel cells. *ECS Trans.* **3,** 835–853 (2006).
29. Debe, M. K. *et al.* Durability aspects of nanostructured thin film catalysts for PEM fuel cells. *ECS Trans.* **1,** 51–56 (2006).
30. Debe, M. K. *et al.* in *Proc. 50th Annual Technical Conference of the Society of Vacuum Coaters* 175–185 (The Society of Vacuum Coaters, 2006).
31. Haugen, G., Barta, S., Emery, M., Hamrock, S. & Yandrasits, M. in *Fuel Cell Chemistry and Operation* (eds Herring, A. M., Zawodzinski Jr., T. A. & Hamrock, S. J.) 137 (ACS Symposium Series 1040, 2010).
32. Steinbach, A. *et al.* Influence of anode GDL on PEMFC ultra-thin electrode water management at low temperatures. *ECS Trans.* **41,** 449–457 (2011).

33. Debe, M. K. *et al.* Extraordinary oxygen reduction activity of Pt_3Ni. *J. Electrochem. Soc.* **158,** B910–B918 (2011).
34. Park, S. *et al.* Polarization losses under accelerated stress test using multiwalled carbon nanotube supported Pt catalyst in PEM fuel cells. *J. Electrochem. Soc.* **158,** B297–B302 (2011).
35. Wang, S., Jiang, S. P., White, T. J. & Wang, X. Synthesis of Pt and Pd nanosheaths on multi-walled carbon nanotubes as potential electrocatalysts of low temperature fuel cells. *Electrochim. Acta* **55,** 7652–7658 (2010).
36. Yang, R., Leisch, J., Strasser, P. & Toney, M. F. Structure of dealloyed $PtCu_3$ thin films and catalyst activity for oxygen reduction. *Chem. Mater.* **22,** 4712–4720 (2010).
37. Erlebacher, J. & Snyder, J. Dealloyed nanoporous metals for PEM fuel cell catalysis. *ECS Trans.* **25,** 603–612 (2009).
38. Erlebacher, J., Aziz, M., Karma, A., Dimitrov, N. & Sieradzki, K. Evolution of nanoporosity in dealloying. *Nature* **410,** 450–453 (2001).
39. Moffat, T. P., Mallett, J. J. & Hwang, S.-M. Oxygen reduction kinetics on electrodeposited $Pt_{100-x}Ni_x$, and $Pt_{100-x}Co_x$. *J. Electrochem. Soc.* **156,** B238–B251 (2009).
40. Imbeault, R., Antonio, P., Garbarino, S. & Guay, D. Oxygen reduction kinetics on Pt_xNi_{100-x} thin films prepared by pulsed laser deposition. *J. Electrochem. Soc.* **157,** B1051–B1058 (2010).
41. Ralph, T. R. & Hogarth, M. P. Catalysis for low temperature fuel cells. *Platin. Met. Rev.* **46,** 3–14 (2002).
42. Schulenburg, H. *et al.* Heat-treated PtCo nanoparticles as oxygen reduction catalysts. *J. Phys. Chem. C* **113,** 4069–4077 (2009).
43. Thompsett, D. in *Handbook of Fuel Cells—Fundamentals, Technology and Applications* (eds Vielstich, W., Lamm, A. & Gasteiger, H. A.) Ch. 37 (John Wiley & Sons, 2003).
44. Wagner, F. T. Automotive Challenges and Opportunities for Oxygen Reduction Catalysts. In *First CARISMA Intl Conf.* (La Grande Motte, France, 23 September 2008).
45. Wang, C. *et al.* Monodisperse Pt_3Co nanoparticles as electrocatalyst: the effects of particle size and pretreatment on electrocatalytic reduction of oxygen. *Phys. Chem. Chem. Phys.* **12,** 6933–6939 (2010).
46. Wu, J. B. *et al.* Truncated octahedral Pt_3Ni ORR electrocatalysts. *J. Am. Chem. Soc.* **132,** 4984–4985 (2010).
47. Zhang, J., Yang, H., Fang, J. & Zou, S. Synthesis and oxygen reduction activity of shape-controlled Pt_3Ni nanopolyhedra. *Nano Lett.* **10,** 638–644 (2010).
48. Lim, B. *et al.* Pd-Pt bimetallic nanodendrites with high activity for oxygen reduction. *Science* **324,** 1302–1305 (2009).
49. Gasteiger, H. A. & Markovic, N. M. Just a dream—or future reality? *Science* **324,** 48–49 (2009).
50. Wang, C. *et al.* Monodisperse Pt_3Co nanoparticles as a catalyst for the oxygen reduction reaction: size-dependent activity. *J. Phys. Chem. C* **113,** 19365–19368 (2009).
51. Wang, C. *et al.* Correlation between surface chemistry and electrocatalytic properties of monodispersed Pt_xNi_{1-x} nanoparticles. *Adv. Funct. Mater.* **21,** 147–152 (2011).
52. Markovic, N. Nanosegregated cathode catalysts with ultra-low platinum loading. In *2010 DOE Hydrogen Program Annual Merit Review* FC-006, http://www.hydrogen.energy.gov/pdfs/review10/fc008_markovic_2010_o_web.pdf (2011).
53. Shao, M., Sasaki, K., Marinkivic, N. S., Zhang, L. & Adzic, R. R. Synthesis and characterization of platinum monolayer oxygen-reduction electrocatalysts with Co-Pd core-shell nanoparticle supports. *Electrochem. Commun.* **9,** 2848–2853 (2007).
54. Bliznakov, S. T., Vukmirovic, M. B., Yang, L., Sutter, E. A. & Adzic, R. R. Pt monolayer on electrodeposited Pd nanostructures—advanced cathode catalysts for PEM fuel cells. *ECS Trans.* **41,** 1055 (2011).
55. Vukmirovic, M. B. *et al.* Platinum monolayer electrocatalysts for oxygen reduction. *Electrochim. Acta* **52,** 2257–2263 (2007).
56. Shao, M. H., Sasaki, K., Lui, P. & Adzic, R. R. Pd_3Fe and Pt monolayer Pd_3Fe electrocatalysts for oxygen reduction. *Z. Phys. Chem.* **221,** 1175–1190 (2007).
57. Zhang, J. *et al.* Platinum monolayer electrocatalysts for O_2 reduction: Pt monolayer on Pd(111) and on carbon-supported Pd nanoparticles. *J. Phys. Chem. B* **108,** 10955–10964 (2004).
58. Russell, A. E. *et al.* In situ XAS studies of core-shell PEM fuel cell catalysts: the opportunities and challenges. *ECS Trans.* **41,** 55–67 (2011).
59. Haug, A. *et al.* Stability of a Pt-Pd core-shell catalyst: a comparative fuel cell and RDE study. *218th ECS Meeting* abstr. 743 (The Electrochemical Society, 2010).
60. Knupp, S. L. *et al.* Platinum monolayer electrocatalysts for O_2 reduction: Pt monolayer on carbon-supported PdIr Nanoparticles. *Electrocatalysis* **1,** 213–223 (2010).
61. Xing, Y. *et al.* Enhancing oxygen reduction reaction activity via Pd-Au alloy sublayer mediation of Pt monolayer electrocatalysts. *J. Phys. Chem. Lett.* **1,** 3238–3242 (2010).
62. Wang, J. X. *et al.* Oxygen reduction on well-defined core-shell nanocatalysts: particle size, facet and Pt shell thickness effects. *J. Am. Chem. Soc.* **131,** 17298–17302 (2009).
This is an exemplary paper in a long series by the Adzic group developing core–shell nanoparticle catalysts having Pt monolayer skins, controlled size and surface facets.
63. Gong, K., Su, D. & Adzic, R. Platinum-monolayer shell on $AuNi_{0.5}Fe$ nanoparticle core electrocatalyst with high activity and stability for the oxygen reduction reaction. *J. Am. Chem. Soc.* **132,** 14364–14366 (2010).
64. Ball, S. *et al.* Structure and activity of novel Pt core-shell catalysts for the oxygen reduction reaction. *ECS Trans.* **25,** 1023–1036 (2009).
65. Korovina, A., Garsany, Y., Epshteyn, A., Swider-Lyons, K. E. & Ramaker, D. E. Insight into oxygen reduction on platinum-tantalum oxyphosphate electrocatalysts. *218th ECS Meeting* abstr. 687 (The Electrochemical Society, 2010).
66. Park, S. *et al.* Polarization losses under accelerated stress test using multiwalled carbon nanotube supported Pt catalyst in PEM fuel cells. *J. Electrochem. Soc.* **158,** B297–B302 (2011).
67. Wang, X., Waje, M. & Yan, Y. CNT-based electrodes with high efficiency for PEMFCs. *Electrochem. Solid-State Lett.* **8,** A42–A44 (2005).
68. Chen, Z., Waje, M., Li, W. & Yan, Y. Supportless Pt and PtPd nanotubes as electrocatalysts for oxygen-reduction reactions. *Angew. Chem. Int. Edn* **46,** 4060–4063 (2007).
69. van der Vliet, D. *et al.* Metallic nanotubes with tunable composition and structure as advanced electrocatalysts. *Nature Mater.* (submitted).
70. Zhou, H., Zhou, W.-P., Adzic, R. & Wong, S. S. Enhanced electrocatalytic performance of one-dimensional metal nanowires and arrays generated via an ambient surfactantless synthesis. *J. Phys. Chem. C* **113,** 5460–5466 (2009).
71. Adzic, R. Contiguous platinum monolayer oxygen reduction electrocatalysts on high-stability-low-cost supports. In *2011 DOE Hydrogen Program Annual Merit Review* FC-009, http://www.hydrogen.energy.gov/pdfs/review11/fc009_adzic_2011_o.pdf (2011).
72. Shao, M. Palladium-based electrocatalysts for hydrogen oxidation and oxygen reduction reactions. *J. Power Sources* **196,** 2433–2444 (2011).
73. Myers, D. Non-platinum bimetallic cathode electrocatalysts. In *2008–2010 DOE Hydrogen Program Annual Merit Reviews* http://www.hydrogen.energy.gov/pdfs/review10/fc004_myers_2010_o_web.pdf (2010).
74. Atanasoski, R. & Dodelet, J.-P. in *Encyclopedia of Electrochemical Power Sources* (eds Garche, J. *et al.*) Vol. 2 639–649 (Elsevier, 2009).
75. Lei, M., Li, P. G., Li, L. H. & Tang, W. H. A highly ordered Fe-N-C nanoarray as a non-precious oxygen-reduction catalyst for proton exchange membrane fuel cells. *J. Power Sources* **196,** 3548–3552 (2011).
76. Wang, S., Yu, D. & Dai, L. Polyelectrolyte functionalized carbon nanotubes as efficient metal-free electrocatalysts for oxygen reduction. *J. Am. Chem. Soc.* **133,** 5182–5185 (2011).
77. Zelenay, P. Advanced cathode catalysts. In *2010 DOE Hydrogen Program Annual Merit Review*, http://www.hydrogen.energy.gov/pdfs/review10/fc005_zelenay_2010_o_web.pdf (2010).
78. Ishihara, A., Ohgi, Y., Matsuzawa, K., Mitsushima, S. & Ota, K. Progress in non-precious metal oxide-based cathode for polymer electrolyte fuel cells. *Electrochim. Acta* **55,** 8005–8012 (2010).
79. Lefevre, M., Proietti, E., Jaouen, F. & Dodelet, J.-P. Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* **324,** 71–74 (2009).
80. Bashyam, R. & Zelenay, P. A class of non-precious metal composite catalysts for fuel cells. *Nature* **443,** 63–66 (2006).
81. Proietti, E. *et al.* Iron-based cathode catalyst with enhanced power density in polymer electrolyte membrane fuel cells. *Nature Commun.* **2,** 416 (2011).
This paper is the latest in a long series by these authors that show an amazing rate of improvement in non-precious metal catalysts' beginning-of-life performances under pure oxygen.
82. Wood, T. E., Tan, Z., Schmoeckel, A. K., O'Neill, D. & Atanasoski, R. Non-precious metal oxygen reduction catalyst for PEM fuel cells based on nitroaniline precursor. *J. Power Sources* **178,** 510–516 (2008).
83. Wu, G., More, K. L., Johnston, C. M. & Zelenay, P. High-Performance electrocatalysts for oxygen reduction derived from polyaniline, iron and cobalt. *Science* **332,** 443–447 (2011).
84. Global *and China Low-E Glass Industry Report* http://pressexposure.com/Global_and_China_Low-E_Glass_Industry_Report,_2010_-_Published_by_ResearchInChina-205310.html (ResearchInChina, 2010).
85. Chen, S., Gasteiger, H. A., Hayakawa, K., Tada, T. & Shao-Horn, Y. Platinum-alloy cathode catalyst degradation in proton exchange membrane fuel cells: nanometer-scale compositional and morphological changes. *J. Electrochem. Soc.* **157,** A82–A97 (2010).
86. Kongkanand, A., Liu, Z., Dutta, I. & Wagner, F. T. Electrochemical and microstructural evaluation of aged nanostructured thin film fuel cell electrocatalyst. *J. Electrochem. Soc.* **158,** B1286–B1291 (2011).
87. Wagner, F. T. *et al.* Catalyst development needs and pathways for automotive PEM fuel cells. *ECS Trans.* **3,** 19 (2006).
88. Koh, S., Hahn, N., Yu, C. & Strasser, P. Effects of composition and annealing conditions on catalytic activities of dealloyed Pt-Cu nanoparticle electrocatalysts for PEMFC. *J. Electrochem. Soc.* **155,** B1281–B1288 (2008).
89. Oezaslan, M., Hasche, F. & Strasser, P. Structure-activity relationship of dealloyed $PtCo_3$ and $PtCu_3$ nanoparticle electrocatalyst for oxygen reduction reaction in PEMFC. *ECS Trans.* **33,** 333–341 (2010).
90. Strasser, P., Hahn, N. T. & Koh, S. Corrosion and ORR activity of Pt alloy electrocatalysts during voltammetric pretreatment. *ECS Trans.* **3,** 139–149 (2006).
91. Mani, P., Srivastava, R. & Strasser, P. Dealloyed binary PtM_3 (M = Cu, Co, Ni) and ternary $PtNi_3M$ (M = Cu, Co, Fe, Cr) electrocatalysts for the oxygen reduction reaction: performance in polymer electrolyte membrane fuel cells. *J. Power Sources* **196,** 666–673 (2011).
92. Neyerlin, K. C., Srivastava, R., Yu, C. & Strasser, P. Electrochemical activity and stability of dealloyed Pt-Cu and Pt-Cu-Co electrocatalysts for the oxygen reduction reaction (ORR). *J. Power Sources* **186,** 261–267 (2009).
93. Wagner, F. T. High-activity dealloyed catalysts. *2011 DOE Hydrogen Program Annual Merit Review* FC-087, http://www.hydrogen.energy.gov/pdfs/review11/fc087_wagner_2011_o.pdf (2011).

94. Strasser, P. *et al.* Lattice-strain control of the activity in dealloyed core-shell fuel cell catalysts. *Nature Chem.* **2,** 454–460 (2010).

95. Snyder, J., Fujita, T., Chen, M. W. & Erlebacher, J. Oxygen reduction in nanoporous metal-ionic liquid composite electrocatalysts. *Nature Mater.* **9,** 904–907 (2010).
This paper shows that porosity on the nanometre scale can be controlled in Ni/Pt alloys, describes the spontaneous formation of core/shell catalysts during de-alloying and illustrates a new concept for enhancing the activity of solid surfaces in contact with ionic liquids.

96. Erlebacher, J. & Seshardi, R. Hard materials with tunable porosity. *MRS Bull.* **34,** 561–568 (2009).

97. Snyder, J. & Erlebacher, J. The active surface area of nanoporous metals during oxygen reduction. *ECS Trans.* **41,** 1021–1030 (2011).

Acknowledgements I gratefully acknowledge support by the Fuel Cell Technologies Program in the Office of Energy Efficiency and Renewable Energy at the US Department of Energy, for grant DE-FG36-07GO17007.

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 the author (mkdebe1@mmm.com).

doi:10.1038/nature11018

Approaching a state shift in Earth's biosphere

Anthony D. Barnosky[1,2,3], Elizabeth A. Hadly[4], Jordi Bascompte[5], Eric L. Berlow[6], James H. Brown[7], Mikael Fortelius[8], Wayne M. Getz[9], John Harte[9,10], Alan Hastings[11], Pablo A. Marquet[12,13,14,15], Neo D. Martinez[16], Arne Mooers[17], Peter Roopnarine[18], Geerat Vermeij[19], John W. Williams[20], Rosemary Gillespie[9], Justin Kitzes[9], Charles Marshall[1,2], Nicholas Matzke[1], David P. Mindell[21], Eloy Revilla[22] & Adam B. Smith[23]

Localized ecological systems are known to shift abruptly and irreversibly from one state to another when they are forced across critical thresholds. Here we review evidence that the global ecosystem as a whole can react in the same way and is approaching a planetary-scale critical transition as a result of human influence. The plausibility of a planetary-scale 'tipping point' highlights the need to improve biological forecasting by detecting early warning signs of critical transitions on global as well as local scales, and by detecting feedbacks that promote such transitions. It is also necessary to address root causes of how humans are forcing biological changes.

Humans now dominate Earth, changing it in ways that threaten its ability to sustain us and other species[1–3]. This realization has led to a growing interest in forecasting biological responses on all scales from local to global[4–7].

However, most biological forecasting now depends on projecting recent trends into the future assuming various environmental pressures[5], or on using species distribution models to predict how climatic changes may alter presently observed geographic ranges[8,9]. Present work recognizes that relying solely on such approaches will be insufficient to characterize fully the range of likely biological changes in the future, especially because complex interactions, feedbacks and their hard-to-predict effects are not taken into account[6,8–11].

Particularly important are recent demonstrations that 'critical transitions' caused by threshold effects are likely[12]. Critical transitions lead to state shifts, which abruptly override trends and produce unanticipated biotic effects. Although most previous work on threshold-induced state shifts has been theoretical or concerned with critical transitions in localized ecological systems over short time spans[12–14], planetary-scale critical transitions that operate over centuries or millennia have also been postulated[3,12,15–18]. Here we summarize evidence that such planetary-scale critical transitions have occurred previously in the biosphere, albeit rarely, and that humans are now forcing another such transition, with the potential to transform Earth rapidly and irreversibly into a state unknown in human experience.

Two conclusions emerge. First, to minimize biological surprises that would adversely impact humanity, it is essential to improve biological forecasting by anticipating critical transitions that can emerge on a planetary scale and understanding how such global forcings cause local changes. Second, as was also concluded in previous work, to prevent a global-scale state shift, or at least to guide it as best we can, it will be necessary to address the root causes of human-driven global change and to improve our management of biodiversity and ecosystem services[3,15–17,19].

Basics of state shift theory

It is now well documented that biological systems on many scales can shift rapidly from an existing state to a radically different state[12]. Biological 'states' are neither steady nor in equilibrium; rather, they are characterized by a defined range of deviations from a mean condition over a prescribed period of time. The shift from one state to another can be caused by either a 'threshold' or 'sledgehammer' effect. State shifts resulting from threshold effects can be difficult to anticipate, because the critical threshold is reached as incremental changes accumulate and the threshold value generally is not known in advance. By contrast, a state shift caused by a sledgehammer effect—for example the clearing of a forest using a bulldozer—comes as no surprise. In both cases, the state shift is relatively abrupt and leads to new mean conditions outside the range of fluctuation evident in the previous state.

Threshold-induced state shifts, or critical transitions, can result from 'fold bifurcations' and can show hysteresis[12]. The net effect is that once a critical transition occurs, it is extremely difficult or even impossible for the system to return to its previous state. Critical transitions can also result from more complex bifurcations, which have a different character from fold bifurcations but which also lead to irreversible changes[20].

Recent theoretical work suggests that state shifts due to fold bifurcations are probably preceded by general phenomena that can be characterized mathematically: a deceleration in recovery from perturbations ('critical slowing down'), an increase in variance in the pattern of within-state fluctuations, an increase in autocorrelation between fluctuations, an increase in asymmetry of fluctuations and rapid back-and-forth shifts ('flickering') between states[12,14,18]. These phenomena can theoretically be

[1]Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [2]Museum of Paleontology, University of California, Berkeley, California 94720, USA. [3]Museum of Vertebrate Zoology, University of California, Berkeley, California 94720, USA. [4]Department of Biology, Stanford University, Stanford, California 94305, USA. [5]Integrative Ecology Group, Estación Biológica de Doñana, CSIC, Calle Américo Vespucio s/n, E-41092 Sevilla, Spain. [6]TRU NORTH Labs, Berkeley, California 94705, USA. [7]Department of Biology, The University of New Mexico, Albuquerque, New Mexico 87131, USA. [8]Department of Geosciences and Geography and Finnish Museum of Natural History, PO Box 64, University of Helsinki, FI-00014 Helsinki, Finland. [9]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA. [10]Energy and Resources Group, University of California, Berkeley, California 94720, USA. [11]Department of Environmental Science and Policy, University of California, One Shields Avenue, Davis, California 95616, USA. [12]Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile. [13]Instituto de Ecología y Biodiversidad, Casilla 653, Santiago, Chile. [14]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [15]Facultad de Ciencias Biologicas, Pontificia Universidad Catolica de Chile, Alameda 340, Santiago, Chile. [16]Pacific Ecoinformatics and Computational Ecology Lab, 1604 McGee Avenue, Berkeley, California 94703, USA. [17]Department of Biological Sciences, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. [18]California Academy of Sciences, 55 Music Concourse Drive, San Francisco, California 94118, USA. [19]Department of Geology, University of California, One Shields Avenue, Davis, California 95616, USA. [20]Department of Geography, University of Wisconsin, Madison, Wisconsin 53706, USA. [21]Department of Biophysics and Biochemistry, University of California, San Francisco, California 94102, USA. [22]Department of Conservation Biology, Estación Biológica de Doñana, CSIC, Calle Américo Vespucio s/n, E-41092 Sevilla, Spain. [23]Center for Conservation and Sustainable Development, Missouri Botanical Garden, 4344 Shaw Boulevard, Saint Louis, Missouri 63110, USA.

assessed within any temporally and spatially bounded system. Although such assessment is not yet straightforward[12,18,20], critical transitions and in some cases their warning signs have become evident in diverse biological investigations[21], for example in assessing the dynamics of disease outbreaks[22,23], populations[14] and lake ecosystems[12,13]. Impending state shifts can also sometimes be determined by parameterizing relatively simple models[20,21].

In the context of forecasting biological change, the realization that critical transitions and state shifts can occur on the global scale[3,12,15–18], as well as on smaller scales, is of great importance. One key question is how to recognize a global-scale state shift. Another is whether global-scale state shifts are the cumulative result of many smaller-scale events that originate in local systems or instead require global-level forcings that emerge on the planetary scale and then percolate downwards to cause changes in local systems. Examining past global-scale state shifts provides useful insights into both of these issues.

Hallmarks of global-scale state shifts

Earth's biosphere has undergone state shifts in the past, over various (usually very long) timescales, and therefore can do so in the future (Box 1). One of the fastest planetary state shifts, and the most recent, was the transition from the last glacial into the present interglacial condition[12,18], which occurred over millennia[24]. Glacial conditions had prevailed for ∼100,000 yr. Then, within ∼3,300 yr, punctuated by episodes of abrupt, decadal-scale climatic oscillations, full interglacial conditions were attained. Most of the biotic change—which included extinctions, altered diversity patterns and new community compositions—occurred within a period of 1,600 yr beginning ∼12,900 yr ago. The ensuing interglacial state that we live in now has prevailed for the past ∼11,000 yr.

Occurring on longer timescales are events such as at least four of the 'Big Five' mass extinctions[25], each of which represents a critical transition that spanned several tens of thousands to 2,000,000 yr and changed the course of life's evolution with respect to what had been normal for the previous tens of millions of years. Planetary state shifts can also substantially increase biodiversity, as occurred for example at the 'Cambrian explosion'[26], but such transitions require tens of millions of years, timescales that are not meaningful for forecasting biological changes that may occur over the next few human generations (Box 1).

Despite their different timescales, past critical transitions occur very quickly relative to their bracketing states: for the examples discussed here, the transitions took less than ∼5% of the time the previous state had lasted (Box 1). The biotic hallmark for each state change was, during the critical transition, pronounced change in global, regional and local assemblages of species. Previously dominant species diminished or went extinct, new consumers became important both locally and globally, formerly rare organisms proliferated, food webs were modified, geographic ranges reconfigured and resulted in new biological communities, and evolution was initiated in new directions. For example, at the Cambrian explosion large, mobile predators became part of the food chain for the first time. Following the K/T extinction, mammalian herbivores replaced large archosaur herbivores. And at the last glacial–interglacial transition, megafaunal biomass switched from being dominated by many species to being dominated by *Homo sapiens* and our domesticated species[27].

All of the global-scale state shifts noted above coincided with global-scale forcings that modified the atmosphere, oceans and climate (Box 1). These examples suggest that past global-scale state shifts required global-scale forcings, which in turn initiated lower-level state changes that local controls do not override. Thus, critical aspects of biological forecasting are to understand whether present global forcings are of a magnitude sufficient to trigger a global-scale critical transition, and to ascertain the extent of lower-level state changes that these global forcings have already caused or are likely to cause.

Present global-scale forcings

Global-scale forcing mechanisms today are human population growth with attendant resource consumption[3], habitat transformation and fragmentation[3], energy production and consumption[28,29], and climate change[3,18]. All of these far exceed, in both rate and magnitude, the forcings evident at the most recent global-scale state shift, the last glacial–interglacial transition (Box 1), which is a particularly relevant benchmark for comparison given that the two global-scale forcings at that time—climate change

BOX 1

Past planetary-scale critical transitions and state shifts

Last glacial–interglacial transition[18,24]. The critical transition was a rapid warm–cold–warm fluctuation in climate between 14,300 and 11,000 yr ago, and the most pronounced biotic changes occurred between 12,900 and 11,300 yr ago[24,27,30,54].

The major biotic changes were the extinction of about half of the species of large-bodied mammals, several species of large birds and reptiles, and a few species of small animals[30]; a significant decrease in local and regional biodiversity as geographic ranges shifted individualistically, which also resulted in novel species assemblages[37,49,53,54]; and a global increase in human biomass and spread of humans to all continents[27].

The pre-transition global state was a glacial stage that lasted about 100,000 yr and the post-transition global state is an interglacial that Earth has been in for approximately 11,000 yr. The global forcings were orbitally induced, cyclic variations in solar insolation that caused rapid global warming. Direct and indirect of effects of humans probably contributed to extinctions of megafauna and subsequent ecological restructuring.

'Big Five' mass extinctions[25]. The respective critical transitions ended at ∼443,000,000, ∼359,000,000, ∼251,000,000, ∼200,000,000 and ∼65,000,000 yr ago. They are each thought to have taken at most 2,000,000 yr to complete but could have been much shorter; the limitations of geological dating preclude more precision. The most recent transition (the K/T extinction, which occurred at the end of the Cretaceous period) may have been the catastrophic result of a bolide impact, and could have occurred on a timescale as short as a human lifetime.

The major biotic changes were the extinction of at least 75% of Earth's species; a major reorganization of global and local ecosystems as previously rare lifeforms gained evolutionary dominance; and the return to pre-extinction levels of biodiversity over hundreds of thousands to millions of years.

The pre- and post-transition global states lasted ∼50,000,000–100,000,000 yr. We are now 65,000,000 yr into the present state on this scale, in an era known as the Cenozoic or the Age of Mammals. The global forcings all corresponded to unusual climate changes and shifts in ocean and atmospheric chemistry, especially in concentrations of carbon dioxide and, in one case, hydrogen sulphide. Intense volcanic activity seems to have been important at some extinction events. A bolide impact is well documented as a cause of the K/T event and has been postulated as a cause of some of the others.

Cambrian explosion[26,81]. The critical transition began ∼540,000,000 yr ago and lasted about 30,000,000 yr.

The major biotic changes were evolutionary innovations resulting in all phyla known today; a conversion of the global ecosystem from one based almost solely on microbes to one based on complex, multicellular life; and diversity increased, but on a timescale that is far too long to be meaningful in predicting the biotic future over human generations.

The pre-transition global state lasted ∼2,000,000,000 yr and was characterized by primary lifeforms consisting of prokaryotic and eukaryotic microbes. The post-transition global state is about 540,000,000 yr old and ongoing. The global forcings were the increase of atmospheric oxygen to levels sufficient for the metabolic processes required to sustain complex, multicellular life, and evolutionary innovations that included large size, predation and complex locomotion.

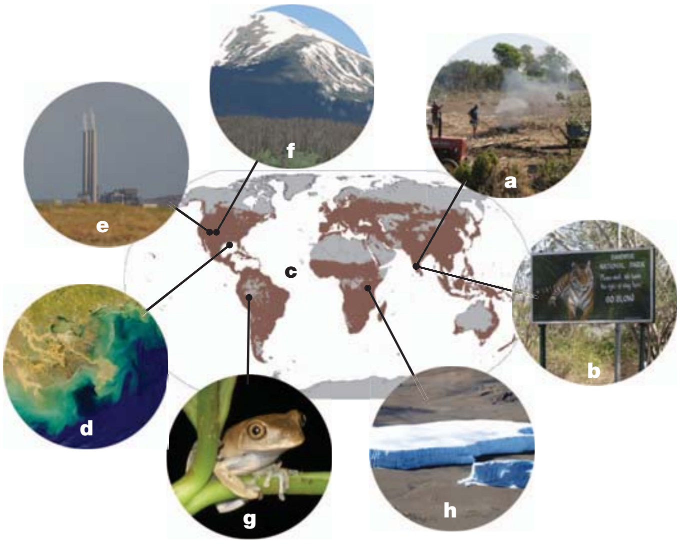


Figure 1 | Drivers of a potential planetary-scale critical transition. **a**, Humans locally transform and fragment landscapes. **b**, Adjacent areas still harbouring natural landscapes undergo indirect changes. **c**, Anthropogenic local state shifts accumulate to transform a high percentage of Earth's surface drastically; brown colouring indicates the approximately 40% of terrestrial ecosystems that have now been transformed to agricultural landscapes, as explained in ref. 34. **d**, Global-scale forcings emerge from accumulated local human impacts, for example dead zones in the oceans from run-off of agricultural pollutants. **e**, Changes in atmospheric and ocean chemistry from the release of greenhouse gases as fossil fuels are burned. **f**–**h**, Global-scale forcings emerge to cause ecological changes even in areas that are far from human population concentrations. **f**, Beetle-killed conifer forests (brown trees) triggered by seasonal changes in temperature observed over the past five decades. **g**, Reservoirs of biodiversity, such as tropical rainforests, are projected to lose many species as global climate change causes local changes in temperature and precipitation, exacerbating other threats already causing abnormally high extinction rates. In the case of amphibians, this threat is the human-facilitated spread of chytrid fungus. **h**, Glaciers on Mount Kilimanjaro, which remained large throughout the past 11,000 yr, are now melting quickly, a global trend that in many parts of the world threatens the water supplies of major population centres. As increasing human populations directly transform more and more of Earth's surface, such changes driven by emergent global-scale forcings increase drastically, in turn causing state shifts in ecosystems that are not directly used by people. Photo credits: E.A.H. and A.D.B. (**a**–**c**, **e**–**h**); NASA (**d**).

and human population growth[27,30]—are also primary forcings today. During the last glacial–interglacial transition, however, these were probably separate, yet coincidental, forcings. Today conditions are very different because global-scale forcings including (but not limited to) climate change have emerged as a direct result of human activities.

Human population growth and per-capita consumption rate underlie all of the other present drivers of global change. The growth in the human population now (∼77,000,000 people per year) is three orders of magnitude higher than the average yearly growth from ∼10,000–400 yr ago (∼67,000 people per year), and the human population has nearly quadrupled just in the past century[31–33]. The most conservative estimates suggest that the population will grow from its present value, 7,000,000,000, to 9,000,000,000 by 2045[31] and to 9,500,000,000 by 2050[31,33].

As a result of human activities, direct local-scale forcings have accumulated to the extent that indirect, global-scale forcings of biological change have now emerged. Direct forcing includes the conversion of ∼43% of Earth's land to agricultural or urban landscapes, with much of the remaining natural landscapes networked with roads[1,2,34,35]. This exceeds the physical transformation that occurred at the last global-scale critical transition, when ∼30% of Earth's surface went from being covered by glacial ice to being ice free.

The indirect global-scale forcings that have emerged from human activities include drastic modification of how energy flows through the global ecosystem. An inordinate amount of energy now is routed through one species, *Homo sapiens*. Humans commandeer ∼20–40% of global net primary productivity[1,2,35] (NPP) and decrease overall NPP through habitat degradation. Increasing NPP regionally through atmospheric and agricultural deposition of nutrients (for example nitrogen and phosphorus) does not make up the shortfall[2]. Second, through the release of energy formerly stored in fossil fuels, humans have substantially increased the energy ultimately available to power the global ecosystem. That addition does not offset entirely the human appropriation of NPP, because the vast majority of that 'extra' energy is used to support humans and their domesticates, the sum of which comprises large-animal biomass that is far beyond that typical of pre-industrial times[27]. A decrease in this extra energy budget, which is inevitable if alternatives do not compensate for depleted fossil fuels, is likely to impact human health and economies severely[28], and also to diminish biodiversity[27], the latter because even more NPP would have to be appropriated by humans, leaving less for other species[36].

By-products of altering the global energy budget are major modifications to the atmosphere and oceans. Burning fossil fuels has increased atmospheric CO_2 concentrations by more than a third (∼35%) with respect to pre-industrial levels, with consequent climatic disruptions that include a higher rate of global warming than occurred at the last global-scale state shift[37]. Higher CO_2 concentrations have also caused the ocean rapidly to become more acidic, evident as a decrease in pH by ∼0.05 in the past two decades[38]. In addition, pollutants from agricultural run-off and urban areas have radically changed how nutrients cycle through large swaths of marine areas[16].

Already observable biotic responses include vast 'dead zones' in the near-shore marine realm[39], as well as the replacement of >40% of Earth's formerly biodiverse land areas with landscapes that contain only a few species of crop plants, domestic animals and humans[3,40]. Worldwide shifts in species ranges, phenology and abundances are concordant with ongoing climate change and habitat transformation[41]. Novel communities are becoming widespread as introduced, invasive and agricultural species integrate into many ecosystems[42]. Not all community modification is leading to species reductions; on local and regional scales, plant diversity has been increasing, owing to anthropogenic introductions[42], counter to the overall trend of global species loss[5,43]. However, it is unknown whether increased diversity in such locales will persist or will eventually decrease as a result of species interactions that play out over time. Recent and projected[5,44] extinction rates of vertebrates far exceed empirically derived background rates[25]. In addition, many plants, vertebrates and invertebrates have markedly reduced their geographic ranges and abundances to the extent that they are at risk of extinction[43]. Removal of keystone species worldwide, especially large predators at upper trophic levels, has exacerbated changes caused by less direct impacts, leading to increasingly simplified and less stable ecological networks[39,45,46].

Looking towards the year 2100, models forecast that pressures on biota will continue to increase. The co-opting of resources and energy use by humans will continue to increase as the global population reaches 9,500,000,000 people (by 2050), and effects will be greatly exacerbated if per capita resource use also increases. Projections for 2100 range from a population low of 6,200,000,000 (requiring a substantial decline in fertility rates) to 10,100,000,000 (requiring continued decline of fertility in countries that still have fertility above replacement level) to 27,000,000,000 (if fertility remains at 2005–2010 levels; this population size is not thought to be supportable; ref. 31). Rapid climate change shows no signs of slowing. Modelling suggests that for ∼30% of Earth, the speed at which plant species will have to migrate to keep pace with projected climate change is greater than their dispersal rate when Earth last shifted from a glacial to an interglacial climate[47], and that dispersal will be thwarted by highly fragmented landscapes. Climates found at present on 10–48% of the planet are projected to disappear within a century, and climates that contemporary organisms have never experienced are likely to cover 12–39% of Earth[48]. The mean global temperature by 2070 (or possibly a few decades earlier) will be higher than it has been since the human species evolved.

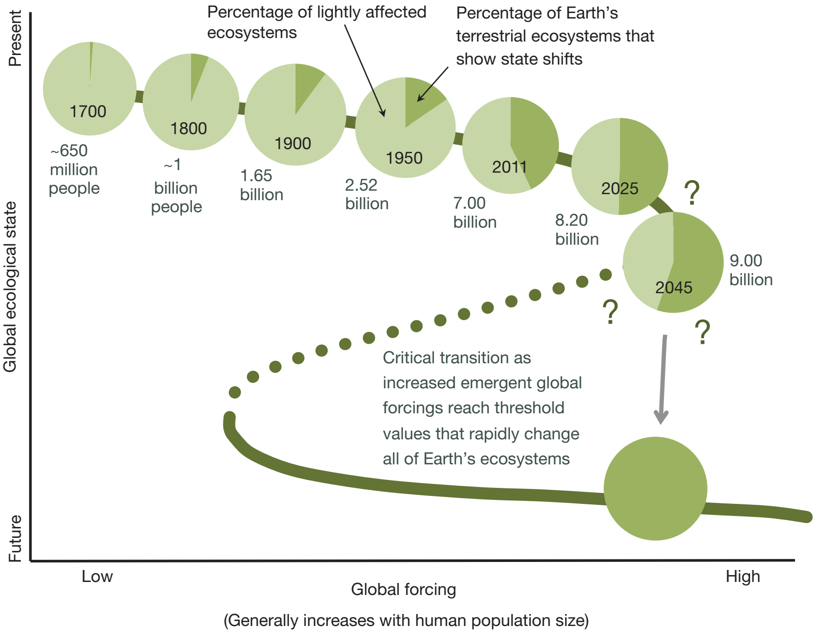


Figure 2 | Quantifying land use as one method of anticipating a planetary state shift. The trajectory of the green line represents a fold bifurcation with hysteresis[12]. At each time point, light green represents the fraction of Earth's land that probably has dynamics within the limits characteristic of the past 11,000 yr. Dark green indicates the fraction of terrestrial ecosystems that have unarguably undergone drastic state changes; these are minimum values because they count only agricultural and urban lands. The percentages of such transformed lands in 2011 come from refs 1, 34, 35, and when divided by 7,000,000,000 (the present global human population) yield a value of approximately 2.27 acres (0.92 ha) of transformed land for each person. That value was used to estimate the amount of transformed land that probably existed in the years 1800, 1900 and 1950, and which would exist in 2025 and 2045 assuming conservative population growth and that resource use does not become any more efficient. Population estimates are from refs 31–33. An estimate of 0.68 transformed acres (0.28 ha) per capita (approximately that for India today) was used for the year 1700, assuming a lesser effect on the global landscape before the industrial revolution. Question marks emphasize that at present we still do not know how much land would have to be directly transformed by humans before a planetary state shift was imminent, but landscape-scale studies and theory suggest that the critical threshold may lie between 50 and 90% (although it could be even lower owing to synergies between emergent global forcings). See the main text for further explanation. Billion, 10^9.

Expecting the unexpected

The magnitudes of both local-scale direct forcing and emergent global-scale forcing are much greater than those that characterized the last global-scale state shift, and are not expected to decline any time soon. Therefore, the plausibility of a future planetary state shift seems high, even though considerable uncertainty remains about whether it is inevitable and, if so, how far in the future it may be. The clear potential for a planetary-scale state shift greatly complicates biotic forecasting efforts, because by their nature state shifts contain surprises. Nevertheless, some general expectations can be gleaned from the natural experiments provided by past global-scale state shifts. On the timescale most relevant to biological forecasting today, biotic effects observed in the shift from the last glacial to the present interglacial (Box 1) included many extinctions[30,49–51]; drastic changes in species distributions, abundances and diversity; and the emergence of novel communities[49,50,52–54]. New patterns of gene flow triggered new evolutionary trajectories[55–58], but the time since then has not been long enough for evolution to compensate for extinctions.

At a minimum, these kinds of effects would be expected from a global-scale state shift forced by present drivers, not only in human-dominated regions but also in remote regions not now heavily occupied by humans (Fig. 1); indeed, such changes are already under way (see above[5,25,39,41–44]). Given that it takes hundreds of thousands to millions of years for evolution to build diversity back up to pre-crash levels after major extinction episodes[25], increased rates of extinction are of particular concern, especially because global and regional diversity today is generally lower than it was 20,000 yr ago as a result of the last planetary state shift[37,50,51,54,59]. This large-scale loss of diversity is not overridden by historical increases in plant species richness in many locales, owing to human-transported species homogenizing the world's biota[42]. Possible too are substantial losses of ecosystem services required to sustain the human population[60]. Still unknown is the extent to which human-caused increases in certain ecosystem services—such as growing food—balances the loss of 'natural' ecosystem services, many of which already are trending in dangerous directions as a result of overuse, pollutants and climate change[3,16]. Examples include the collapse of cod and other fisheries[45,61,62]; loss of millions of square kilometres of conifer forests due to climate-induced bark-beetle outbreaks;[63] loss of carbon sequestration by forest clearing[60]; and regional losses of agricultural productivity from desertification or detrimental land-use practices[1,35]. Although the ultimate effects of changing biodiversity and species compositions are still unknown, if critical thresholds of diminishing returns in ecosystem services were reached over large areas and at the same time global demands increased (as will happen if the population increases by 2,000,000,000 within about three decades), widespread social unrest, economic instability and loss of human life could result[64].

Towards improved biological forecasting and monitoring

In view of potential impacts on humanity, a key need in biological forecasting is the development of ways to anticipate a global critical transition, ideally in time to do something about it[65]. It is possible to imagine qualitative aspects of a planetary state shift given present human impacts (Fig. 1), but criteria that would indicate exactly how close we might be to a planetary-scale critical transition remain elusive. Three approaches should prove helpful in defining useful benchmarks and tracking progression towards them.

Tracking global-scale changes

The first approach acknowledges the fact that local-scale state changes—whether they result from sledgehammer or threshold effects—trigger critical transitions over regions larger than the directly affected area, as has been shown both empirically and theoretically[66–70]. On the landscape scale, tipping points in undisturbed patches are empirically evident when 50–90% of the surrounding patches are disturbed. Simulations indicate that critical transitions become much more likely when the probability of connection of any two nodes in a network (ecological or otherwise) drops below ~59% (refs 66–70). More generally, dense human populations, roads and infrastructure, and land transformation are known to cause ecological changes outside the areas that have actually undergone sledgehammer state changes[68]. Translating these principles to the planetary scale would imply that once a sufficient proportion of Earth's ecosystems have undergone transformation, the remainder can change rapidly (Fig. 2), especially because emergent, larger-scale forcings (for instance changes in atmospheric and ocean chemistry, nutrient and energy cycling, pollution and so on) multiply and interact to exacerbate local forcings[21] (Fig. 1). It is still unknown, however, what percentage of Earth's ecosystems actually have to be transformed to new states by the direct action of humans for rapid state changes to be triggered in remaining 'natural' systems. That percentage may be knowable only in retrospect, but, judging from landscape-scale observations and simulations[66–70], it can reasonably be expected to be as low as 50% (ref. [68]), or even lower if the interaction effects of many local ecosystem transformations cause sufficiently large global-scale forcings to emerge.

In that context, continued efforts to track global-scale changes by remote sensing and other techniques will be essential in assessing how close we are to tipping the balance towards an Earth where most ecosystems are directly altered by people. This is relatively straightforward for land and it has already been demonstrated that at least 43% of Earth's terrestrial ecosystems have undergone wholesale transformation[1,2,34,40], on average equating to ~2.27 transformed acres (0.92 ha) per capita for the present human population. Assuming that this average rate of land transformation per capita does not change, 50% of Earth's land will have undergone state shifts when the global population reaches 8,200,000,000, which is estimated to occur by the year 2025[31]. Under the same land-use assumption and according to only slightly less conservative population growth models, 70% of Earth's land could be shifted to human use (if the population reaches 11,500,000,000) by 2060[31].

Assessing the percentage change to new states in marine systems, and the direct human footprint on the oceans, is much more challenging, but available data suggest widespread effects[38,39]. More precise quantification of ecosystem state shifts in the oceans is an important task, to the extent that ocean ecosystems cover most of the planet.

Tracking local-scale changes caused by global forcings

The second approach is the direct monitoring of biological change in local study systems caused by external forcing. Such monitoring will be vital, particularly where the human footprint is thought to be small. Observing unusual changes in such areas, as has occurred recently in Yellowstone Park, USA, which has been protected since 1872[71], and in many remote watersheds[72], would indicate that larger-scale forcings[38,73] are influencing local ecological processes.

A key problem has been how to recognize 'unusual' change, because biological systems are dynamic and shifting baselines have given rise to many different definitions of 'normal', each of which can be specified as unusual within a given temporal context. However, identifying signals of a global-scale state shift in any local system demands a temporal context that includes at least a few centuries or millennia, to encompass the range of ecological variation that would be considered normal over the entire ~11,000-yr duration of the present interglacial period. Identifying unusual biotic changes on that scale has recently become possible through several different approaches, which are united by their focus on integrating spatial and temporal information (Box 2). Breakthroughs include characterizing ecosystems using taxon-independent metrics that can be tracked with palaeontological data through pre-anthropogenic times and then compared with present conditions and monitored into the future; recognizing macro-ecological patterns that indicate disturbed systems; combining phylochronologic and phylogeographic information to trace population dynamics over several millennia; and assessing the structure and stability of ecological networks using theoretical and empirical methods. Because all of these approaches benefit from time series data, long-term monitoring efforts and existing palaeontological and natural history museum collections will become particularly valuable[74].

BOX 2

Integrating spatio-temporal data on large scales to detect planetary state shifts

- Palaeontology uses historical, fossil and geological information to calibrate normal levels of fluctuation in biodiversity, species composition and abundance[80], food webs[82], ecomorphology[83], extinction[25] and so on. Recent work shows that some lightly populated ecosystems still operate within bounds that would be considered normal for the present interglacial period, but that others have been disturbed[80].
- Macroecology provides quantitative ways to identify when a particular ecosystem has unusual characteristics in such metrics as the species–area relationship, species abundance distributions, spatial aggregation patterns[84,85], the distribution of metabolic rates over individuals in a community[85,86], the inverse power-law relation between abundance and body size[87], and the distribution of linkages across species in a trophic network[88]. Recent advances in formalizing the maximum entropy (MaxEnt) theory of ecology[85,86] provide a theoretical means of accurately predicting such patterns in undisturbed ecosystems; significant departures from the predictions of MaxEnt probably indicate disturbed systems[85].
- Population biology uses life history, abundance, genetics and numerical modelling to assess population dynamics and viability. Recent advances in obtaining ancient DNA from samples several thousand years old, plus newly developed analytical models that take into account temporal (phylochronologic) as well as spatial (phylogeographic) patterning, increase power in testing whether genetic patterning on the modern landscape deviates significantly from patterns that arise on the scale of centuries to millennia[10,89].
- Ecological network theory regards ecosystems as complex networks of species connected by different interactions. Recent work identifies persistent and stabilizing characteristics of networks on different geographic and temporal scales[81,82] (both current and palaeontological), such as consumer–resource body size ratios[90], allometric scaling effects[91] and skewed distributions for connectivity[81,92,93] and interaction strengths[94–96]. Alteration in such characteristics signals perturbation of the normal network structure. Theoretical work also is revealing where information about species-specific traits such as body size[46,90,91], trophic generality[91], trophic uniqueness[97], non-trophic interactions[98] and phylogenetic information[99] may help predict when ecosystem services degrade as networks destabilize[46,100] and disassemble[97].

Synergy and feedbacks

Thresholds leading to critical transitions are often crossed when forcings are magnified by the synergistic interaction of seemingly independent processes or through feedback loops[3,16]. Given that several global-scale forcings are at work today, understanding how they may combine to magnify biological change is a key challenge[3,15–17]. For example, rapid climate change combined with highly fragmented species ranges can be expected to magnify the potential for ecosystem collapse, and wholesale landscape changes may in turn influence the biology of oceans.

Feedback loops also occur among seemingly discrete systems that operate at different levels of the biological hierarchy[6,8,37] (genotype, phenotype, populations, species distributions, species interactions and so on). The net effect is that a biological forcing applied on one scale can cause a critical transition to occur on another scale. Examples include inadvertent, anthropogenic selection for younger maturation of individual cod as a result of heavy fishing pressure[61]; population crashes due to decreased genetic diversity[75]; mismatch in the phenology of flowering and pollination resulting from interaction of genetic factors, temperature, photoperiod and/or precipitation[76]; and cascades of ecological changes triggered by the removal of top predators[62]. In most cases, these 'scale-jumping' effects, and the mechanisms that drive them, have become apparent only in hindsight, but even so they take on critical importance in revealing interaction effects that can now be incorporated into the next generation of biological forecasts.

Finally, because the global-scale ecosystem comprises many smaller-scale, spatially bounded complex systems (for instance the community within a given physiographic region), each of which overlaps and interacts with others, state shifts of the small-scale components can propagate to cause a state shift of the entire system[21]. Our understanding of complexity at this level can be increased by tracking changes within many different ecosystems in a parallel fashion, from landscape-scale studies of state-shifts[12,21] and from theoretical work that is under way[20]. Potential interactions between overlapping complex systems, however, are proving difficult to characterize mathematically, especially when the systems under study are not well known and are heterogeneous[20]. Nevertheless, one possibility emerging from such work is that long-term transient behaviours, where sudden changes in dynamics can occur after periods of relative stasis even in the absence of outside forces, may be pervasive at the ecosystem level[20], somewhat analogously to delayed metapopulation collapse as a result of extinction debt[77]. This potential 'lag-time' effect makes it all the more critical rapidly to address, where possible, global-scale forcings that can push the entire biosphere towards a critical transition.

Guiding the biotic future

Humans have already changed the biosphere substantially, so much so that some argue for recognizing the time in which we live as a new geologic epoch, the Anthropocene[3,16,78]. Comparison of the present extent of planetary change with that characterizing past global-scale state shifts, and the enormous global forcings we continue to exert, suggests that another global-scale state shift is highly plausible within decades to centuries, if it has not already been initiated.

As a result, the biological resources we take for granted at present may be subject to rapid and unpredictable transformations within a few human generations. Anticipating biological surprises on global as well as local scales, therefore, has become especially crucial to guiding the future of the global ecosystem and human societies. Guidance will require not only scientific work that foretells, and ideally helps to avoid[65], negative effects of critical transitions, but also society's willingness to incorporate expectations of biological instability[64] into strategies for maintaining human well-being.

Diminishing the range of biological surprises resulting from bottom-up (local-to-global) and top-down (global-to-local) forcings, postponing their effects and, in the optimal case, averting a planetary-scale critical transition demands global cooperation to stem current global-scale anthropogenic forcings[3,15–17,19]. This will require reducing world population growth[31] and per-capita resource use; rapidly increasing the proportion of the world's energy budget that is supplied by sources other than fossil fuels while also becoming more efficient in using fossil fuels when they provide the only option[79]; increasing the efficiency of existing means of food production and distribution instead of converting new areas[34] or relying on wild species[39] to feed people; and enhancing efforts to manage as reservoirs of biodiversity and ecosystem services, both in the terrestrial[80] and marine realms[39], the parts of Earth's surface that are not already dominated by humans. These are admittedly huge tasks, but are vital if the goal of science and society is to steer the biosphere towards conditions we desire, rather than those that are thrust upon us unwittingly.

1. Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. *Science* **277,** 494–499 (1997).
2. Haberl, H. *et al.* Quantifying and mapping the human appropriation of net primary production in Earth's terrestrial ecosystems. *Proc. Natl Acad. Sci. USA* **104,** 12942–12947 (2007).
3. Steffen, W. *et al.* The Anthropocene: from global change to planetary stewardship. *AMBIO* **40,** 739–761 (2011).
This paper summarizes the many ways in which humans are changing the planet, argues that the combined effect is as strong as geological forces and points to the likelihood of planetary tipping points.
4. Convention on Biological Diversity. Strategic Plan for Biodiversity 2011–2020, http://www.cbd.int/sp/ (2011).
5. Pereira, H. M. *et al.* Scenarios for global biodiversity in the 21st century. *Science* **330,** 1496–1501 (2010).
6. Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C. & Mace, G. M. Beyond predictions: biodiversity conservation in a changing climate. *Science* **332,** 53–58 (2011).
7. *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* www.ipbes.net (2011).
8. Lavergne, S., Mouquet, N., Thuiller, W. & Ronce, O. Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annu. Rev. Ecol. Evol. Syst.* **41,** 321–350 (2010).
9. Jackson, S. T., Betancourt, J. L., Booth, R. K. & Gray, S. T. Ecology and the ratchet of events: climate variability, niche dimensions, and species distributions. *Proc. Natl Acad. Sci. USA* **106,** 19685–19692 (2009).
10. Ramakrishnan, U. & Hadly, E. A. Using phylochronology to reveal cryptic population histories: review and synthesis of four ancient DNA studies. *Mol. Ecol.* **18,** 1310–1330 (2009).
11. Gilman, S. E., Urban, M. C., Tewksbury, J., Gilchrist, G. W. & Holt, R. D. A framework for community interactions under climate change. *Trends Ecol. Evol.* **25,** 325–331 (2010).
12. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).
This paper presents a general approach to the detection of critical transitions and outlines the possibility of there being general indicators.
13. Carpenter, S. R. *et al.* Early warnings of regime shifts: a whole-ecosystem experiment. *Science* **332,** 1079–1082 (2011).
14. Drake, J. M. & Griffen, B. D. Early warning signals of extinction in deteriorating environments. *Nature* **467,** 456–459 (2010).
15. Folke, C. *et al.* Reconnecting to the biosphere. *AMBIO* **40,** 719–738 (2011).
16. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).
This paper specifies important planetary boundaries and explains why exceeding them would be detrimental to humanity.
17. Westley, F. *et al.* Tipping toward sustainability: emerging pathways of transformation. *AMBIO* **40,** 762–780 (2011).
18. Lenton, T. M. Early warning of climate tipping points. *Nature Clim. Change* **1,** 201–209 (2011).
19. Galaz, V. *et al.* 'Planetary boundaries' — exploring the challenges for global environmental governance. *Curr. Opin. Environ. Sustain.* **4,** 80–87 (2012).
20. Hastings, A. & Wysham, D. Regime shifts in ecological systems can occur with no warning. *Ecol. Lett.* **13,** 464–472 (2010).
This paper points out that regime shifts in complex systems need not result from saddle-node bifurcations and thus may not show the typical early warning signals.
21. Peters, D. P. C. *et al.* in *Real World Ecology* (eds Miao, S. L., Carstenn, S. & Nungesser, M. K.) 47–71 (Springer, 2009).
22. Getz, W. M. Disease and the dynamics of foodwebs. *PLoS Biol.* **7,** e1000209 (2009).
23. Getz, W. M. Biomass transformation webs provide a unified approach to consumer–resource modeling. *Ecol. Lett.* **14,** 113–124 (2011).
24. Hoek, W. Z. The last glacial-interglacial transition. *Episodes* **31,** 226–229 (2008).
25. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471,** 51–57 (2011).
26. Marshall, C. R. Explaining the Cambrian "Explosion" of animals. *Annu. Rev. Earth Planet. Sci.* **34,** 355–384 (2006).
27. Barnosky, A. D. Megafauna biomass tradeoff as a driver of Quaternary and future extinctions. *Proc. Natl Acad. Sci. USA* **105,** 11543–11548 (2008).
28. Brown, J. H. *et al.* Energetic limits to economic growth. *Bioscience* **61,** 19–26 (2011).
29. McDaniel, C. N. & Borton, D. N. Increased human energy use causes biological diversity loss and undermines prospects for sustainability. *Bioscience* **52,** 929–936 (2002).
30. Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. *Annu. Rev. Ecol. Evol. Syst.* **37,** 215–250 (2006).
31. United Nations, Department of Economic and Social Affairs. World Population Prospects, the 2010 Revision, http://esa.un.org/unpd/wpp/Analytical-Figures/htm/fig_1.htm (2011).
32. Population Reference Bureau. Population Projections 2050, http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=15 (2012).
33. United Nations. *World Population to 2300* 1–254 (United Nations, Department of Economic and Social Affairs Population Division, 2004).
34. Foley, J. A. *et al.* Solutions for a cultivated planet. *Nature* **478,** 337–342 (2011).
This paper provides estimates for the amount of land that has been transformed by agricultural activities and summarizes steps required to feed 9,000,000,000 people.
35. Vitousek, P. M., Ehrlich, P. R., Ehrlich, A. H. & Matson, P. A. Human appropriation of the products of photosynthesis. *Bioscience* **36,** 368–373 (1986).
36. Maurer, B. A. Relating human population growth to the loss of biodiversity. *Biodivers. Lett.* **3,** 1–5 (1996).
37. Blois, J. L. & Hadly, E. A. Mammalian response to Cenozoic climatic change. *Annu. Rev. Earth Planet. Sci.* **37,** 181–208 (2009).

38. Doney, S. C. The growing human footprint on coastal and open-ocean biogeochemistry. *Science* **328,** 1512–1516 (2010).
39. Jackson, J. B. C. Ecological extinction and evolution in the brave new ocean. *Proc. Natl Acad. Sci. USA* **105,** 11458–11465 (2008).
40. Ellis, E. C. Anthropogenic transformation of the terrestrial biosphere. *Phil. Trans. R. Soc. A* **369,** 1010–1035 (2011).
41. Parmesan, C. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* **37,** 637–669 (2006).
42. Ellis, E. C., Antill, E. C. & Kref, H. Plant biodiversity in the Anthropocene. *PLoS ONE* **7,** e30535 (2012).
43. Vié, J.-C., Hilton-Taylor, C. & Stuart, S. N. (eds) *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species* 180 (IUCN, 2009).
44. Hoffmann, M. *et al.* The impact of conservation on the status of the world's vertebrates. *Science* **330,** 1503–1509 (2010).
45. Jackson, J. B. C. *et al.* Historical overfishing and the recent collapse of coastal ecosystems. *Science* **293,** 629–637 (2001).
46. Bascompte, J., Melián, C. J. & Sala, E. Interaction strength combinations and the overfishing of a marine food web. *Proc. Natl Acad. Sci. USA* **102,** 5443–5447 (2005).
47. Loarie, S. R. *et al.* The velocity of climate change. *Nature* **462,** 1052–1055 (2009).
48. Williams, J. W., Jackson, S. T. & Kutzbach, J. E. Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl Acad. Sci. USA* **104,** 5738–5742 (2007).
49. Graham, R. W, *et al.* Spatial response of mammals to late Quaternary environmental fluctuations. *Science* **272,** 1601–1606 (1996).
50. Blois, J. L., McGuire, J. L. & Hadly, E. A. Small mammal diversity loss in response to late-Pleistocene climatic change. *Nature* **465,** 771–774 (2010).
51. Carrasco, M. A., Barnosky, A. D. & Graham, R. W. Quantifying the extent of North American mammal extinction relative to the pre-anthropogenic baseline. *PLoS ONE* **4,** e8331 (2009).
52. Williams, J. W. & Jackson, S. T. Novel climates, no-analog communities, and ecological surprises. *Front. Ecol. Environ* **5,** 475–482 (2007).
53. Williams, J. W., Shuman, B. N. & Webb, T. III. Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. *Ecology* **82,** 3346–3362 (2001).
54. Williams, J. W., Shuman, B. N., Webb, T. III, Bartlein, P. J. & Leduc, P. L. Late Quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecol. Monogr.* **74,** 309–334 (2004).
55. Hadly, E. A. *et al.* Genetic response to climatic change: insights from ancient DNA and phylochronology. *PLoS Biol.* **2,** e290 (2004).
56. Shapiro, B. *et al.* Rise and fall of the Beringian steppe bison. *Science* **306,** 1561–1565 (2004).
57. Hewitt, G. M. Genetic consequences of climatic oscillations in the Quaternary. *Phil. Trans. R. Soc. Lond. B* **359,** 183–195 (2004).
58. Lister, A. M. The impact of Quaternary Ice Ages on mammalian evolution. *Phil. Trans. R. Soc. Lond. B* **359,** 221–241 (2004).
59. Barnosky, A. D., Carrasco, M. A. & Graham, R. W. in *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies* (eds McGowan, A. J. & Smith, A. B.) 179–189 (Geological Society, 2011).
60. Foley, J. A. *et al.* Global consequences of land use. *Science* **309,** 570–574 (2005).
61. Olsen, E. M. *et al.* Maturation trends indicative of rapid evolution preceded the collapse of northern cod. **428,** 932–935 (2004).
62. Estes, J. A. *et al.* Trophic downgrading of planet Earth. *Science* **333,** 301–306 (2011).
63. Kurz, W. A. *et al.* Mountain pine beetle and forest carbon feedback to climate change. *Nature* **452,** 987–990 (2008).
64. Shearer, A. W. Whether the weather: comments on 'An abrupt climate change scenario and its implications for United States national security'. *Futures* **37,** 445–463 (2005).
65. Biggs, R., Carpenter, S. R. & Brock, W. A. Turning back from the brink: detecting an impending regime shift in time to avert it. *Proc. Natl Acad. Sci. USA* **106,** 826–831 (2009).
66. Bascompte, J. & Solé, R. V. Habitat fragmentation and extinction thresholds in spatially explicit models. *J. Anim. Ecol.* **65,** 465–473 (1996).
67. Swift, T. L. & Hannon, S. J. Critical thresholds associated with habitat loss: a review of the concepts, evidence, and applications. *Biol. Rev. Camb. Philos. Soc.* **85,** 35–53 (2010).
This paper synthesizes studies that quantify thresholds of habitat disturbance above which regime shifts can propagate to undisturbed patches.
68. Noss, R. F. *et al.* Bolder thinking for conservation. *Conserv. Biol.* **26,** 1–4 (2012).
69. Pardini, R., Bueno, A. A., Gardner, T. A., Prado, P. I. & Metzger, J. P. Beyond the fragmentation threshold hypothesis: regime shifts in biodiversity across fragmented landscapes. *PLoS ONE* **5,** e13666 (2010).
70. Bradonjić, M., Hagberg, A., &. Percus, A. G. in *Algorithms and Models for the Web-Graph (WAW 2007)* (eds Bonato, A. & Chung, F.) 209–216 (Springer, 2007).
71. McMenamin, S. K., Hadly, E. A. & Wright, C. K. Climatic change and wetland desiccation cause amphibian decline in Yellowstone National Park. *Proc. Natl Acad. Sci. USA* **105,** 16988–16993 (2008).
72. Holtgrieve, G. W. *et al.* A coherent signature of anthropogenic nitrogen deposition to remote watersheds of the northern hemisphere. *Science* **334,** 1545–1548 (2011).
This paper documents how human impacts are reaching into remote ecosystems.
73. Peñuelas, J., Sardans, J., Rivas-Ubach, A. & Janssens, I. A. The human-induced imbalance between C, N and P in Earth's life system. *Glob. Change Biol.* **18,** 3–6 (2012).
74. Johnson, K. G. *et al.* Climate change and biosphere response: unlocking the collections vault. *Bioscience* **61,** 147–153 (2011).
75. Ramakrishnan, U., Hadly, E. A. & Mountain, J. L. Detecting past population bottlenecks using temporal genetic data. *Mol. Ecol.* **14,** 2915–2922 (2005).
76. Forrest, J. & Miller-Rushing, A. J. Toward a synthetic understanding of the role of phenology in ecology and evolution. *Phil. Trans. R. Soc. B* **365,** 3101–3112 (2010).
77. Hanski, I. & Ovaskainen, O. Extinction debt at extinction threshold. *Conserv. Biol.* **16,** 666–673 (2002).
78. Zalasiewicz, J., Williams, M., Haywood, A. & Ellis, M. The Anthropocene: a new epoch of geological time? *Phil. Trans. R. Soc. A* **369,** 835–841 (2011).
79. Pacala, S. & Socolow, R. Stabilization wedges: solving the climate problem for the next 50 years with current technologies. *Science* **305,** 968–972 (2004).
80. Hadly, E. A. & Barnosky, A. D. in *Conservation Paleobiology: Using the Past to Manage for the Future* (eds Dietl, G. P. & Flessa, K. W.) 39–59 (Paleontological Society, 2009).
This paper summarized metrics that can be tracked through millennia and into the future to assess when ecosystems are perturbed from the Holocene baseline, and discusses conservation strategies that will be needed in the future.
81. Dunne, J. A., Williams, R. J., Martinez, N. D., Wood, R. A. & Erwin, D. H. Compilation and network analysis of Cambrian food webs. *PLoS Biol.* **6,** e102 (2008).
82. Roopnarine, P. D. in *Quantitative Methods in Paleobiology* (eds Alroy, J. & Hunt, G.) 143–161 (Paleontological Society, 2010).
83. Polly, P. D. *et al.* History matters: ecometrics and integrative climate change biology. *Proc. R. Soc. B* **278,** 1131–1140 (2011).
84. Brown, J. H. *Macroecology* (Univ. Chicago Press, 1995).
85. Harte, J. *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics* (Oxford Univ. Press, 2011).
This book presents comprehensive evidence that prevailing patterns in the spatial distribution, abundance and energetics of species in relatively undisturbed ecosystems are predicted by the maximum-information-entropy inference procedure, and that systematic departures from theory arise in highly disturbed ecosystems.
86. Harte, J., Smith, A. B. & Storch, D. Biodiversity scales from plots to biomes with a universal species-area curve. *Ecol. Lett.* **12,** 789–797 (2009).
87. White, E., Ernest, S., Kerkhoff, A. & Enquist, B. Relationships between body size and abundance in ecology. *Trends Ecol. Evol.* **22,** 323–330 (2007).
88. Williams, R. J. Simple MaxEnt models explain foodweb degree distributions. *Theor. Ecol.* **3,** 45–52 (2010).
89. Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L. & Hadly, E. A. Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21,** 1733–1734 (2005).
90. Brose, U., Williams, W. J. & Martinez, N. D. Allometric scaling enhances stability in complex food webs. *Ecol. Lett.* **9,** 1228–1236 (2006).
91. Otto, S. B., Rall, B. C. & Brose, U. Allometric degree distributions facilitate food-web stability. *Nature* **450,** 1226–1229 (2007).
92. Jordano, P., Bascompte, J. & Olesen, J. M. Invariant properties in coevolutionary networks of plant-animal interactions. *Ecol. Lett.* **6,** 69–81 (2003).
93. Solé, R. V. & Montoya, J. M. Complexity and fragility in ecological networks. *Proc. R. Soc. Lond. B* **268,** 2039–2045 (2001).
94. Kokkoris, G. D., Troumbis, A. Y. & Lawton, J. H. Patterns of species interaction strength in assembled theoretical competition communities. *Ecol. Lett.* **2,** 70–74 (1999).
95. McCann, K., Hastings, A., &. Huxel, G. R. Weak trophic interactions and the balance of nature. *Nature* **395,** 794–798 (1998).
96. Neutel, A.-M., Heesterbeek, J. A. P. & de Ruiter, P. C. Stability in real food webs: weak links in long loops. *Science* **296,** 1120–1123 (2002).
97. Sahasrabudhe, S. & Motter, A. E. Rescuing ecosystems from extinction cascades through compensatory perturbations. *Nature Commun.* **2,** 170 (2011).
98. Kéfi, S. *et al.* More than a meal: integrating non-feeding interactions into food webs. *Ecol. Lett.* **15,** 291–300 (2012).
99. Rezende, E. L., Lavabre, J. E., Guimarães, P. R. Jr, Jordano, P. & Bascompte, J. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448,** 925–928 (2007).
100. Berlow, E. L. *et al.* Simple prediction of interaction strengths in complex food webs. *Proc. Natl Acad. Sci. USA* **106,** 187–191 (2009).
This computational exploration of complex network structure and dynamics successfully predicts the quantitative effect of a species loss on other species within its community and therefore demonstrates the potential of ecological network theory to predict state changes following species loss.

Acknowledgements This research grew out of a workshop funded by The University of California at Berkeley Office of the Vice Chancellor for Research under the auspices of the Berkeley Initiative for Global Change Biology. We thank J. Jackson for discussions and Paul Ehrlich for comments.

Author Contributions All authors participated in the workshop and discussions that resulted in this paper, and provided key insights from their respective research specialties. A.D.B. and E.A.H. were the lead writers and synthesizers. J.B., E.L.B., J.H.B., M.F., W.M.G., J.H., A.H., A.M., P.A.M, N.D.M., P.R., G.V. and J.W.W. compiled data and/or figures and wrote parts of the text. R.G., J.K., C.M., N.M., D.P.M., E.R. and A.B.S. contributed to finalizing the text.

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 A.D.B. (barnosky@berkeley.edu).

doi:10.1038/nature11148

Biodiversity loss and its impact on humanity

Bradley J. Cardinale[1], J. Emmett Duffy[2], Andrew Gonzalez[3], David U. Hooper[4], Charles Perrings[5], Patrick Venail[1], Anita Narwani[1], Georgina M. Mace[6], David Tilman[7], David A. Wardle[8], Ann P. Kinzig[5], Gretchen C. Daily[9], Michel Loreau[10], James B. Grace[11], Anne Larigauderie[12], Diane S. Srivastava[13] & Shahid Naeem[14]

The most unique feature of Earth is the existence of life, and the most extraordinary feature of life is its diversity. Approximately 9 million types of plants, animals, protists and fungi inhabit the Earth. So, too, do 7 billion people. Two decades ago, at the first Earth Summit, the vast majority of the world's nations declared that human actions were dismantling the Earth's ecosystems, eliminating genes, species and biological traits at an alarming rate. This observation led to the question of how such loss of biological diversity will alter the functioning of ecosystems and their ability to provide society with the goods and services needed to prosper.

In the past 20 years remarkable progress has been made towards understanding how the loss of biodiversity affects the functioning of ecosystems and thus affects society. Soon after the 1992 Earth Summit in Rio de Janeiro, interest in understanding how biodiversity loss might affect the dynamics and functioning of ecosystems, and the supply of goods and services, grew dramatically. Major international research initiatives formed; hundreds of experiments were performed in ecosystems all over the globe; new ecological theories were developed and tested against experimental results.

Here we review two decades of research that has examined how biodiversity loss influences ecosystem functions, and the impacts that this can have on the goods and services ecosystems provide (Box 1). We begin with a brief historical introduction. We then summarize the major results from research that has provided increasingly rigorous answers to the question of how and why the Earth's biological diversity influences the functioning of ecosystems. After this, we consider the closely related issue of how biodiversity provides specific ecosystem services of value to humanity. We close by considering how the next generation of biodiversity science can reduce our uncertainties and better serve policy and management initiatives.

A brief history

During the 1980s, concern about the rate at which species were being lost from ecosystems led to research showing that organisms can influence the physical formation of habitats (ecosystem engineering[1]), fluxes of elements in biogeochemical cycles (for example, ecological stoichiometry[2]), and the productivity of ecosystems (for example, via trophic cascades and keystone species[3]). Such research suggested that loss of certain life forms could substantially alter the structure and functioning of whole ecosystems.

By the 1990s, several international initiatives were focused on the more specific question of how the diversity of life forms impacts upon ecosystems. The Scientific Committee on Problems of the Environment (SCOPE) produced an influential book reviewing the state of knowledge on biodiversity and ecosystem functioning (BEF)[4]. The United Nations Environment Program commissioned the Global Biodiversity Assessment to evaluate the state of knowledge on biodiversity, including its role in ecosystem and landscape processes[5]. Building on early studies of the effects of biodiversity on ecosystem processes, DIVERSITAS, the international programme dedicated to biodiversity science, produced a global research agenda[6].

By the mid-1990s, BEF studies had manipulated the species richness of plants in laboratory and field experiments and suggested that ecosystem functions, like biomass production and nutrient cycling, respond strongly to changes in biological diversity[7–10]. Interpretation of these studies was initially controversial, and by the late 1990s BEF researchers were involved in a debate over the validity of experimental designs, the mechanisms responsible for diversity effects, and the relevance of results to non-experimental systems[11]. This controversy helped to create a decade of research that, by 2009, generated several hundred papers reporting results of >600 experiments that manipulated more than 500 types of organisms in freshwater, marine and terrestrial ecosystems[11,12].

As the field of BEF developed, a related body of research began to form an agenda for biodiversity and ecosystem services (BES) research built on the idea that ecosystems provide essential benefits to humanity[13,14]. Although BES did not evolve separately from BEF, it took a distinctly different direction. The main focus of BES was on large-scale patterns across landscapes more relevant to economic or cultural evaluation. For many BES applications, biodiversity was considered an ecosystem service in-and-of itself[15]. When biodiversity was viewed as an underlying factor driving ecosystem services, the term was often used loosely to mean the presence/absence of entire habitats or groups of organisms (for example, impact of mangrove forests on flood protection, or of all native pollinators on pollination).

The 2005 Millennium Ecosystem Assessment[16] appraised, for the first time, the condition and trends in the world's ecosystems and the services they provide, and highlighted two distinct foci of BEF and BES research. Research on BEF had developed a large body of experiments and mathematical theory describing how genetic, species and functional diversity of organisms control basic ecological processes (functions) in ecosystems (Box 1). Studies on BES were, in contrast, mostly correlative, conducted at the landscape scale and often focused on how major habitat modifications influenced 'provisioning' and 'regulating' services of ecosystems.

[1]School of Natural Resources and Environment, University of Michigan, Ann Arbor, Michigan 48109, USA. [2]Virginia Institute of Marine Science, The College of William and Mary, Gloucester Point, Virginia 23062, USA. [3]McGill University, Department of Biology, Montreal, Quebec H3A 1B1, Canada. [4]Western Washington University, Department of Biology, Bellingham, Washington 98225, USA. [5]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [6]Centre for Population Biology, Imperial College London, Silwood Park SL5 7PY, UK. [7]Department of Ecology, Evolution & Behavior, University of Minnesota, Saint Paul, Minnesota 55108, USA. [8]Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, S- 901 83 Umeå, Sweden. [9]Department of Biology and Woods Institute, Stanford University, Stanford, California 94305, USA. [10]Station d'Ecologie Expérimentale, Centre National de la Recherche Scientifique, 09200 Moulis, France. [11]US Geological Survey, National Wetlands Research Center, Lafayette, Louisiana 70506, USA. [12]Museum National d'Histoire Naturelle, 57, Rue Cuvier, CP 41 75231, Paris Cedex 05, France. [13]Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [14]Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, New York 10027, USA.

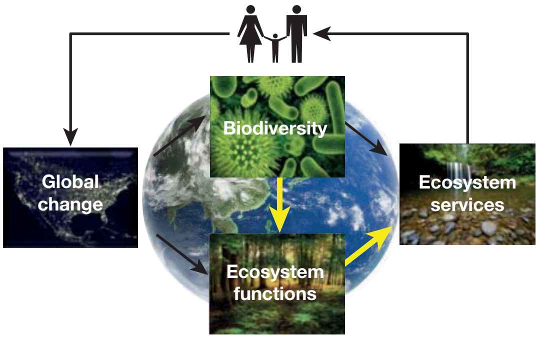
BOX 1
The scope of our review

In this Review we ask how biodiversity per se—that is, the variety of genes, species, or functional traits in an ecosystem—has an impact on the functioning of that ecosystem and, in turn, the services that the ecosystem provides to humanity (yellow arrows, Box 1 Fig. 1 below). This encompasses questions such as can a forest store more carbon if it has a greater variety of tree species? Can a stream clean up more pollution if it has a greater variety of microbial genotypes? Can natural enemies better control agricultural pests if they are composed of a variety of predators, parasites and pathogens?

Biodiversity is the variety of life, including variation among genes, species and functional traits. It is often measured as: richness is a measure of the number of unique life forms; evenness is a measure of the equitability among life forms; and heterogeneity is the dissimilarity among life forms.

Ecosystem functions are ecological processes that control the fluxes of energy, nutrients and organic matter through an environment. Examples include: primary production, which is the process by which plants use sunlight to convert inorganic matter into new biological tissue; nutrient cycling, which is the process by which biologically essential nutrients are captured, released and then recaptured; and decomposition, which is the process by which organic waste, such as dead plants and animals, is broken down and recycled.

Ecosystem services are the suite of benefits that ecosystems provide to humanity. Here we focus on two types of ecosystem services—provisioning and regulating. Provisioning services involve the production of renewable resources (for example, food, wood, fresh water). Regulating services are those that lessen environmental change (for example, climate regulation, pest/disease control).



Images from NASA and Shutterstock.com; used with permission.

The 20th anniversary of the 1992 Earth Summit is an opportune time to review what has been learned from both fields, and to continue their synthesis towards a data-driven consensus. In the sections that follow, we summarize how biological variation per se acts as an independent variable to affect the functions and services of ecosystems.

20 years of research on BEF

In addition to the proliferation of experiments (>600 since 1990)[12], BEF research has developed a substantial body of mathematical theory[17–19], and expanded its scope to include global patterns in natural ecosystems[20–22]. More than half of all work has been published since the last consensus paper in 2005 (ref. 23), and since that time, several milestones have been crossed: the field has coalesced around a series of key findings and themes that have been fostered by the publication of 13 quantitative data syntheses[12,24–35]; many of the early scientific debates have subsided as data have amassed to resolve key controversies; a new consensus is emerging concerning the field's unanswered questions and how to address them. These milestones provide a unique opportunity to re-evaluate earlier conclusions and to identify emerging trends.

Six consensus statements

We conclude that the balance of evidence that has accrued over the last two decades justifies the following statements about how biodiversity loss has an impact on the functioning of ecosystems.

Consensus statement one

There is now unequivocal evidence that biodiversity loss reduces the efficiency by which ecological communities capture biologically essential resources, produce biomass, decompose and recycle biologically essential nutrients.

Meta-analyses published since 2005 have shown that, as a general rule, reductions in the number of genes, species and functional groups of organisms reduce the efficiency by which whole communities capture biologically essential resources (nutrients, water, light, prey), and convert those resources into biomass[12,24–28,30–35] (Fig. 1). Recent meta-analyses further suggest that plant litter diversity enhances decomposition and recycling of elements after organisms die[12], although the effects tend to be weaker than for other processes. Biodiversity effects seem to be remarkably consistent across different groups of organisms, among trophic levels and across the various ecosystems that have been studied[12,24,25,31]. This consistency indicates that there are general underlying principles that dictate how the organization of communities influences the functioning of ecosystems. There are exceptions to this statement for some ecosystems and processes[12,32,36], and these offer opportunities to explore the boundaries that constrain biodiversity effects.

Consensus statement two

There is mounting evidence that biodiversity increases the stability of ecosystem functions through time.

Numerous forms of 'stability' have been described, and there is no theoretical reason to believe that biodiversity should enhance all forms of stability[37]. But theory and data both support greater temporal stability of a community property like total biomass at higher levels of diversity. Five syntheses have summarized how diversity has an impact on variation of ecosystem functions through time[38–42], and these have

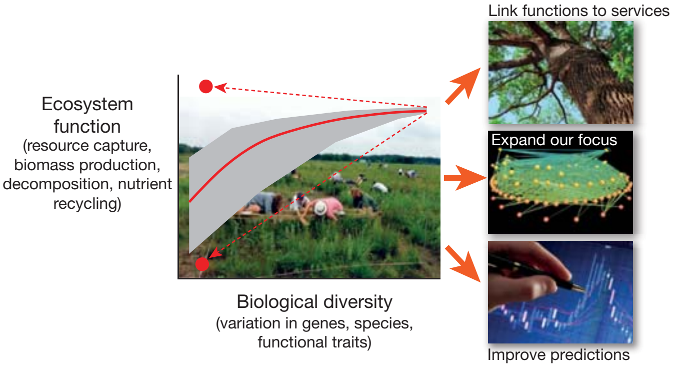


Figure 1 | The form of a typical diversity–function relationship. This conceptual diagram summarizes what we know about the shape of the biodiversity–ecosystem functioning (BEF) relationship based on summaries of several hundred experiments[12,24–35]. The red line shows the average change across all combinations of genes, species, or traits. The grey polygon represents the 95% confidence interval, whereas red dots give maximum and minimum values of the most or least productive species grown alone in monoculture (see main text about uncertainties associated with the upper bound). To improve our predictions of how diversity loss influences the goods and services of ecosystems, we must now take this experimental relationship and (1) link the ecosystem functions measured in experiments to the provisioning and regulating services of ecosystems; (2) expand the focus of research to better mimic realistic extinction scenarios and trophic structures of natural ecosystems; and (3) develop mathematical models that can scale experimental results to whole landscapes. Images from D.T., N. Martinez and Shutterstock.com; used with permission.

shown that total resource capture and biomass production are generally more stable in more diverse communities. The mechanisms by which diversity confers stability include over-yielding, statistical averaging and compensatory dynamics. Over-yielding enhances stability when mean biomass production increases with diversity more rapidly than its standard deviation. Statistical averaging occurs when random variation in the population abundances of different species reduces the variability of aggregate ecosystem variables[43]. Compensatory dynamics are driven by competitive interactions and/or differential responses to environmental fluctuations among different life forms, both of which lead to asynchrony in their environmental responses[18,44]. We have yet to quantify the relative importance of these mechanisms and the conditions under which they operate.

Consensus statement three

The impact of biodiversity on any single ecosystem process is nonlinear and saturating, such that change accelerates as biodiversity loss increases.

The form of BEF relationships in most experimental studies indicates that initial losses of biodiversity in diverse ecosystems have relatively small impacts on ecosystem functions, but increasing losses lead to accelerating rates of change[12,25,31](Fig. 1). We do not yet have quantitative estimates of the level of biodiversity at which change in ecosystem functions become significant for different processes or ecosystems, and this is an active area of research[12,31]. Although our statement is an empirical generality, some researchers question whether saturating curves are an artefact of overly simplified experiments[45]. Saturation could be imposed by the spatial homogeneity, short timescales, or limited species pools of experiments that minimize opportunities for expression of niche differences. In support of this hypothesis, select case studies suggest that as experiments run longer, saturating curves become more monotonically increasing[46]. In addition, biodiversity–ecosystem function relationships in natural ecosystems sometimes differ from saturating curves[22], and future research needs to assess when and why these differences occur.

Consensus statement four

Diverse communities are more productive because they contain key species that have a large influence on productivity, and differences in functional traits among organisms increase total resource capture.

Much of the historical controversy in BEF research involved the extent to which diversity effects are driven by single, highly productive species versus some form of 'complementarity' among species[47,48]. Research and syntheses over the past 10 years have made it clear that both the identity and the diversity of organisms jointly control the functioning of ecosystems. Quantification of the variance explained by species identity versus diversity in $>$200 experiments found that, on average across many ecosystems, each contributes roughly 50% to the net biodiversity effect[12]. Complementarity may represent niche partitioning or positive species interactions[48], but the extent to which these mechanisms broadly contribute to ecosystem functioning has yet to be confirmed[12,49].

Consensus statement five

Loss of diversity across trophic levels has the potential to influence ecosystem functions even more strongly than diversity loss within trophic levels.

Much work has shown that food web interactions are key mediators of ecosystem functioning, and that loss of higher consumers can cascade through a food web to influence plant biomass[50,51]. Loss of one or a few top predator species can reduce plant biomass by at least as much[52] as does the transformation of a diverse plant assemblage into a species monoculture[12]. Loss of consumers can also alter vegetation structure, fire frequency, and even disease epidemics in a range of ecosystems[51].

Consensus statement six

Functional traits of organisms have large impacts on the magnitude of ecosystem functions, which give rise to a wide range of plausible impacts of extinction on ecosystem function.

The extent to which ecological functions change after extinction depends greatly on which biological traits are extirpated[23,53]. Depending on the traits lost, scenarios of change vary from large reductions in ecological processes (for example, if the surviving life form is highly unproductive) to the opposite where the efficiency, productivity and stability of an ecosystem increase. To illustrate this latter possibility, a summary of BEF experiments showed that 65% of 1,019 experimental plots containing plant polycultures produced less biomass than that achieved by their most productive species grown alone[27]. This result has been questioned on statistical grounds[54], and because the short duration of experiments may limit the opportunity for diverse polycultures to out-perform productive species[27]. Even so, the key point is that although diversity clearly has an impact on ecosystem functions when averaged across all genes, species and traits, considerable variation surrounds this mean effect, stemming from differences in the identity of the organisms and their functional traits (Fig. 1). To predict accurately the consequences of any particular scenario of extinction, we must know which life forms have greatest extinction risk, and how the traits of those organisms influence function[55]. Quantifying functional trait diversity and linking this to both extinction risk and ecosystem processes is a rapidly expanding area of research[53,55].

Four emerging trends

In addition to the consensus statements above, data published in the past few years have revealed four emerging trends that are changing the way we view the functional consequences of biodiversity loss.

Emerging trend one

The impacts of diversity loss on ecological processes might be sufficiently large to rival the impacts of many other global drivers of environmental change.

Although biodiversity has a significant impact on most ecosystem functions, there have been questions about whether these effects are large enough to rank among the major drivers of global change. One recent study[56] compared 11 long-term experiments performed at one research site, and another[57] used a suite of meta-analyses from published data to show that the impacts of species loss on primary productivity are of comparable magnitude to the impacts of drought, ultraviolet radiation, climate warming, ozone, acidification, elevated CO_2, herbivory, fire and certain forms of nutrient pollution. Because the BEF relationship is nonlinear (see above), the exact ranking of diversity relative to other drivers will depend on the magnitude of biodiversity loss, as well as magnitudes of other environmental changes. Nevertheless, these two studies indicate that diversity loss may have as quantitatively significant an impact on ecosystem functions as other global change stressors (for example, climate change) that have already received substantial policy attention[58].

Emerging trend two

Diversity effects grow stronger with time, and may increase at larger spatial scales.

Diversity effects in small-scale, short-term experiments may underestimate the impacts of diversity loss on the functioning of more natural ecosystems[45]. At larger spatial scales and with greater temporal fluctuations, more environmental heterogeneity may increase opportunities for species to exploit more niches. Consistent with this argument, a growing body of research now shows that the net effects of biodiversity on ecosystem functions grow stronger as experiments run longer[27,46,59]. Limited data also support the notion that diversity effects grow stronger at larger spatial scales[12,60,61] and with greater resource heterogeneity[62–64]. Thus, BEF research so far may have underestimated the minimum levels of biodiversity required for ecosystem processes.

Emerging trend three

Maintaining multiple ecosystem processes at multiple places and times requires higher levels of biodiversity than does a single process at a single place and time.

Most BEF research has focused on one diversity–function relationship at a time. An emerging body of work suggests that the number of species needed to sustain any single process is lower than the number of species needed to sustain multiple processes simultaneously[21,65–67]. Moreover, organisms that control ecological processes at any single location, or in any particular year, often differ from those that control processes in other locations or years[67]. As such, more biodiversity is required to maintain the 'multi-functionality' of ecosystems at multiple places and times.

Emerging trend four

The ecological consequences of biodiversity loss can be predicted from evolutionary history.

BEF research has been dominated by studies that have used species richness as their primary measure of biodiversity. But species represent 'packages' for all the genetic and trait variation that influences the efficiency and metabolism of an organism, and these differences are shaped by patterns of common ancestry[68]. Recent meta-analyses suggest that phylogenetic distances among species (that is, a measure of genetic divergence) may explain more variation in biomass production than taxonomic diversity[34,35]. This suggests that evolutionary processes that generate trait variation among organisms are, in part, responsible for the ecosystem consequences of biodiversity loss.

20 years of research on BES

Over the past 20 years, researchers have developed a rigorous understanding of the services that natural and modified ecosystems provide to society[16]. We have learned that (1) optimizing ecosystems for certain provisioning services, especially food, fibre and biofuel production, has greatly simplified their structure, composition and functioning across scales; (2) simplification has enhanced certain provisioning services, but reduced others, particularly regulating services; and (3) simplification has led to major losses of biodiversity[16]. However, critical questions remain about whether biodiversity loss per se is the cause of impaired ecosystem services in simplified landscapes.

The BES field has resulted in fewer syntheses than has the BEF field, in part because many services cannot be measured directly or manipulated experimentally. We have, therefore, summarized the balance of evidence with our own literature review (Box 2). We began by collating lists of ecosystem services that have been used in recent summaries[15,24,33,69]. We did not include cultural services in our review, which would describe people's non-consumptive uses of biodiversity such as recreation, tourism, education, science and cultural identity. Whether people are motivated by an interest in particular species (for example, totemic or charismatic megafauna) or particular landscapes (for example, wilderness areas or national parks), their demand for cultural services implies a demand for the biodiversity and ecosystem functions required to support the species or communities of interest. Even so, cultural services have rarely been investigated with respect to diversity per se. Here we focused our efforts on the provisioning and regulating services of ecosystems (Box 1), as these are the services that biodiversity studies have most often measured, and that are most frequently related to ecosystem functions.

We began our review by identifying data syntheses that have used either 'vote-counting' (in which the authors tallied the number of studies showing positive, negative, or nonsignificant relationships) or formal statistical meta-analyses (in which authors analysed previously published data to measure standardized correlation coefficients, regression slopes or effect sizes) to quantify relationships between biodiversity and each ecosystem service. For any service for which a data synthesis was not found, we performed our own summary of peer-reviewed articles using search terms in Supplementary Table 1. Papers were sorted by relevance to maximize the match to search terms, after which, we reviewed the top 100 papers for each ecosystem service (leading to a review of >1,700 titles and abstracts). For papers with data, we categorized the diversity–service relationship as positive, negative, or nonsignificant according to the authors' own statistical tests.

Detailed results of our data synthesis are summarized in Supplementary Table 2, and the most salient points are given in Table 1. We believe the following statements are supported by this peer-reviewed literature.

Balance of evidence

Statement one

There is now sufficient evidence that biodiversity per se either directly influences (experimental evidence) or is strongly correlated with (observational evidence) certain provisioning and regulating services.

The green arrows in Table 1 show the ecosystem services for which there is sufficient evidence to conclude that biodiversity has an impact on the service as predicted. For provisioning services, data show that (1) intraspecific genetic diversity increases the yield of commercial crops; (2) tree species diversity enhances production of wood in plantations; (3) plant species diversity in grasslands enhances the production of fodder; and (4) increasing diversity of fish is associated with greater stability of fisheries yields. For regulating processes and services, (1) increasing plant biodiversity increases resistance to invasion by exotic plants; (2) plant pathogens, such as fungal and viral infections, are less prevalent in more diverse plant communities; (3) plant species diversity increases aboveground carbon sequestration through enhanced biomass production (but see statement 2 concerning long-term carbon storage); and (4) nutrient mineralization and soil organic matter increase with plant richness.

Most of these services are ones that can be directly linked to the ecosystem functions measured in BEF experiments. For example, experiments that test the effects of plant species richness on aboveground biomass production are also those that provide direct evidence for effects of diversity on aboveground carbon sequestration and on fodder production. For services less tightly linked to ecosystem functions (for example, services associated with specific populations rather than ecosystem-level properties), we often lack rigorous verification of the diversity–service relationship.

Statement two

For many of the ecosystem services reviewed, the evidence for effects of biodiversity is mixed, and the contribution of biodiversity per se to the service is less well defined.

The yellow arrows in Table 1 show ecosystem services for which the available evidence has revealed mixed effects of biodiversity on the service. For example, in one data synthesis, 39% of experiments in crop production systems reported that plant species diversity led to greater yield of the desired crop species, whereas 61% reported reduced yield[70]. Impacts of biodiversity on long-term carbon storage were similarly mixed, where carbon storage refers to carbon stocks that remained in the system (in plants or soils) for ≥10 years. Comparably few studies have examined storage rather than sequestration. Evidence on the effect of plant diversity on pest abundance is also mixed, with four available data syntheses showing different results. Evidence for an effect of animal diversity on the prevalence of animal disease is mixed, despite recent claims that biodiversity generally suppresses disease[71]. Important opportunities exist for new research to assess the factors that control variation in the response of these services to changes in biodiversity.

Statement three

For many services, there are insufficient data to evaluate the relationship between biodiversity and the service.

There were three ecosystem services for which we found no data, about one-third had less than five published relationships, and half had fewer than ten (see Supplementary Table 2, white cells). This included some noteworthy examples, such as the effect of fish diversity on fisheries yield (as opposed to stability), and the effect of biodiversity on flood regulation. Surprisingly, each of these services has been cited in the literature as being a direct product of biodiversity[16,26]. Some of this discrepancy may be attributable to different uses of the term biodiversity (Box 1). For example, the Millennium Ecosystem Assessment reported that biodiversity enhances flood protection[16], but examples were based on destruction of entire ecosystems (forests, mangroves, or wetlands) leading to increased flood risk. We did not consider complete habitat conversion in our analyses (see Box 2 for examples).

In addition, claims about biodiversity based on ancillary evidence are not reflected in our analyses. For example, we found little direct evidence that genetic diversity enhances the temporal stability of crop yield (as opposed to total yield); yet, most farmers and crop breeders recognize that genetic diversity provides the raw material for selection of desirable traits, and can facilitate rotations that minimize crop damage caused by pests, disease and the vagaries of weather[72]. Although in some instances the ancillary evidence provides rather convincing evidence for a role of biodiversity in providing the ecosystem service, other cases are less convincing. This emphasizes the need for stronger and more explicit evidence to back up claims for biodiversity effects on ecosystem services.

Statement four

For a small number of ecosystem services, current evidence for the impact of biodiversity runs counter to expectations.

The red arrows in Table 1 illustrate cases where the balance of evidence currently runs counter to claims about how biodiversity should affect the ecosystem service. For example, it has been argued that biodiversity could enhance the purity of water by removing nutrient and other chemical pollutants, or by reducing the loads of harmful pests (for example, faecal coliform bacteria, fungal pathogens)[16]. There are examples where genetic or species diversity of algae enhances removal of nutrient pollutants from fresh water[12], or where diversity of filter-feeding organisms reduces waterborne pathogens[73]. However, there are even more examples that show no relationship between biodiversity and water quality.

Finally, there are instances where increased biodiversity may be deleterious. For example, although diverse assemblages of natural enemies (predators, parasitoids and pathogens) are frequently more effective in reducing the density of herbivorous pests[74], diverse natural enemy communities sometimes inhibit biocontrol[75], often because enemies attack each other through intra-guild predation[76]. Another example relates to human health, where more diverse pathogen populations are likely to create higher risks of infectious disease, and strains of bacteria and viruses that evolve drug resistance pose health and economic burdens to people[77]. Such examples caution against making sweeping statements that biodiversity always brings benefits to society.

Outlook and directions

If we are to manage and mitigate for the consequences of diversity loss effectively, we need to build on the foundations laid down by BEF and BES research to expand its realism, relevance and predictive ability. At the same time, we need feedback from policy and management arenas to forge new avenues of research that will make the science even more useful. Here we consider how the next generation of biodiversity science can reduce our uncertainties and better serve policy and management initiatives for the global environment.

Integrating BEF and BES research

The fields of BEF and BES have close intellectual ties, but important distinctions are evident. We see at least two avenues that could facilitate better integration. First, an important frontier involves detailing the mechanistic links between ecosystem functions and services (Box 1). The BEF field has routinely measured functions without extending those to known services, whereas the BES field has routinely described services without understanding their underlying ecological functions. A challenge to linking these two perspectives is that services are often regulated by multiple functions, which do not necessarily respond to changes in biodiversity in the same way. For example, if we want to know how biodiversity influences the ability of ecosystems to remove CO_2 from the atmosphere and store carbon over long time frames, then we need to consider the net influence of biodiversity on photosynthesis (exchange of CO_2 for O_2), carbon sequestration (accumulation of carbon in live plant

BOX 2

Linking biodiversity to ecosystem services

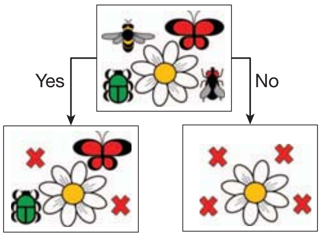
We reviewed >1,700 papers to summarize the balance of evidence linking biodiversity to the goods and services provided by ecosystems. We collated lists of provisioning and regulating services that have been the focus of recent summaries (Supplementary Table 1), and then searched the ISI Web of Knowledge to identify any previously published data syntheses that have summarized known relationships between biodiversity and each ecosystem service. When a data synthesis was not found, we completed our own summary of peer-reviewed articles and categorized the diversity–service relationship as positive, negative, or nonsignificant according to the authors' own statistical tests. Articles had to meet the following four criteria for inclusion.

Criterion 1: the study had to test explicitly for a relationship between biodiversity (defined in Box 1) and the focal ecosystem service using experimental (diversity manipulated) or observational (diversity not manipulated) data.

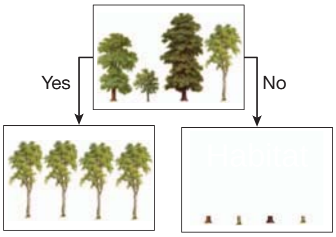
Criterion 2: the study had to quantify biodiversity and the focal service directly (that is, studies using proxies were excluded).

Criterion 3: if authors of the original study identified confounding variables, data were included only if the effects of those confounding variables were statistically controlled for before quantifying the diversity–service relationship.

Criterion 4: the study had to compare a more diverse to less diverse ecosystem containing at least one service providing unit. Any comparison to ecosystems with no service providing unit was excluded (see Box 2 Fig. 1 and Box 2 Fig. 2 for two examples).



Box 2 Figure 1 | Pollination is an ecosystem service provided by a wide variety of organisms, and is essential to the production of many of the world's food crops. We considered studies that compare services like pollination success (for example, fruit set) in a diverse system to a less diverse system (bottom left). But we excluded studies comparing services of a diverse system to one with no service providing organisms (bottom right). Although the latter can quantify the value of service providing organisms (for example, pollinators), it says nothing about the role of biodiversity.



Box 2 Figure 2 | Forests provide a wide array of ecosystem services such as carbon sequestration, wood production and water purification. We considered studies that compare diverse to less diverse habitats (bottom left). However, we did not consider studies that compare services in diverse habitats to those where the habitat was destroyed (for example, clear cut). Although the latter may show the value of the habitat for ecosystem services, it cannot tell us the role of biodiversity.

Table 1 | Balance of evidence linking biodiversity to ecosystem services

Category of service	Measure of service provision	SPU	Diversity level	Source	Study type	*N*	Relationship	
							Predicted	Actual
Provisioning								
Crops	Crop yield	Plants	Genetic	DS	Exp	575	🟩↗	🟩↗
			Species	DS	Exp	100	🟨↗	🟨⤧
Fisheries	Stability of fisheries yield	Fish	Species	PS	Obs	8	🟩↗	🟩↗
Wood	Wood production	Plants	Species	DS	Exp	53	🟩↗	🟩↗
Fodder	Fodder yield	Plants	Species	DS	Exp	271	🟩↗	🟩↗
Regulating								
Biocontrol	Control of herbivorous pests *(*bottom-up effect of plant diversity)	Plants	Species	DS*	Obs	40	🟩↘	🟩↘
		Plants	Species	DS†	Exp	100	🟩↘	🟩↘
		Plants	Species	DS‡	Exp	287	🟨↘	🟨⤧
		Plants	Species	DS§	Exp	100	🟥↘	🟥0
	Control of herbivorous pests (top-down effect of natural enemy diversity)	Natural enemies	Species/trait	DS*	Obs	18	🟩↘	🟩↘
		Natural enemies	Species	DS†	Exp/Obs	266	🟩↘	🟩↘
		Natural enemies	Species	DS‡	Exp	38	🟨↘	🟨⤧
	Resistance to plant invasion	Plants	Species	DS	Exp	120	🟩↗	🟩↗
	Disease prevalence (on plants)	Plants	Species	DS	Exp	107	🟩↘	🟩↘
	Disease prevalence (on animals)	Multiple	Species	DS	Exp/Obs	45	🟨↘	🟨⤧
Climate	Primary production	Plants	Species	DS	Exp	7	🟥↗	🟥0
	Carbon sequestration	Plants	Species	DS	Exp	479	🟩↗	🟩↗
	Carbon storage	Plants	Species/trait	PS	Obs	33	🟨↗	🟨⤧
Soil	Soil nutrient mineralization	Plants	Species	DS	Exp	103	🟩↗	🟩↗
	Soil organic matter	Plants	Species	DS	Exp	85	🟩↗	🟩↗
Water	Freshwater purification	Multiple	Genetic/species	PS	Exp	8	🟥↗	🟥0
Pollination	Pollination	Insects	Species	PS	Obs	7	🟨↗	🟨⤧

For each ecosystem service we searched the ISI Web of Knowledge for published data syntheses (DS). The footnote symbols in the 'Source' column refer to different syntheses. When a synthesis was not available, we completed our own primary search (PS, see Box 2). Detailed results are given in Supplementary Table 2. Data presented here are summarized as follows: green, actual data relationships agree with predictions; yellow, Data show mixed results; red, data conflict with predictions. Exp, experimental; *N*, number of data points; Obs, observed; SPU, service providing unit (where natural enemies include predators, parasitoids and pathogens). Note that 13 ecosystem services are not included in this table due to lack of data (<5 relationships, see Supplementary Table 2).

tissue), herbivory (plant carbon eaten by animals), and decomposition (carbon returned to atmosphere as plants die and decompose). Researchers in the BEF and BES fields will need to work more closely to quantify the networks of mechanistic links between ecosystem functions and services.

Second, the fields of BEF and BES could better exploit their complementary approaches to research. Research on BEF has focused mostly on smaller spatial scales conducive to controlled experiments, which has made it difficult to scale results to real ecosystems at larger scales where services are delivered. Studies on BES have relied heavily on observational data, and often failed to separate general biotic effects on ecosystem services (for example, biomass, habitats or entire groups of organisms) from effects of biodiversity per se (that is, variation in life forms). To better merge these two programmes, BEF and BES will need to expand their scopes of research and develop theoretical approaches that can link the small-scale, mechanistic focus of BEF research to large-scale patterns that are the focus of BES. We discuss each of these in turn.

Expanding our scope

The need to explore more realistic scenarios of diversity change that reflect how human activities are altering biodiversity is now urgent. Organisms are not lost from ecosystems at random, and traits that predispose species to extinction are often those that drive ecosystem processes[55,78]. So far this issue has mostly been explored through simulations[79,80], but food web theory[81] based on using environmental stressors to cause nonrandom extinctions may provide a basis for a new generation of BEF experiments. Furthermore, invasions and range expansions driven by anthropogenic change are homogenizing Earth's biota and, in several cases, increasing local taxonomic diversity[82]. Predicting the ecosystem consequences of simultaneous gains (invasion) and losses (extinction) requires that we first understand which biological traits predispose life forms to higher probabilities of extirpation or establishment (response traits), and detail how response traits covary with traits that drive ecosystem functioning (effect traits)[55]. For example, at local scales invasive plants often have functional traits that are associated with more rapid resource acquisition and growth than those of coexisting native species[83], although global meta-analyses suggest only modest differences between native and introduced plants in their effects on ecosystem processes[84]. Statistical models[85] have been developed that allow integration of invasion and extinction into a trait framework, and these models should now be extended to predict changes in ecosystem services.

Another challenge is to incorporate better the real complexity of food webs into BEF and BES research[30,52]. Most research so far has focused on simplified 'model' communities. Yet, in nature, food webs are complex networks with dozens to thousands of species, have reticulate webs of indirect and nonlinear interactions, and contain mismatches in the spatial and temporal dynamics of interacting organisms. This complexity can appear to preclude predictability. But recent theory[86,87] and experiments[88,89] suggest that food-web structure, interactions and stability can be predicted by a small subset of traits such as organismal body size, the degree of dietary generalism[88] and trophic level[89]. Simple trait-based approaches hold promise for simplifying the inherent complexity of natural food webs into a few key axes that strongly control ecosystem functions and services. We need to better identify these traits and food-web structures, and need better models to explain why certain food-web properties control ecosystem functions and services.

Improving predictions

Increasing the complexity and realism of experiments, however, will not be enough to move biodiversity research towards better forecasting. We also need sets of models and statistical tools that help us move from experiments that detail local biological processes to landscape-scale patterns where management and policy take place (Fig. 2). One fruitful approach may be to use data from BEF experiments to assign parameters to local models of species interactions that predict how biodiversity has an impact on ecosystem processes based on functional traits. These local models could then be embedded into spatially explicit meta-community and ecosystem models that incorporate habitat heterogeneity, dispersal and abiotic drivers to predict relationships between biodiversity and ecosystem services at the landscape level[18]. Statistical tools like structural equation modelling might then be used to assess whether predictions of these landscape models agree with observations from natural systems, and to disentangle effects of biodiversity from other covarying environmental factors[20].

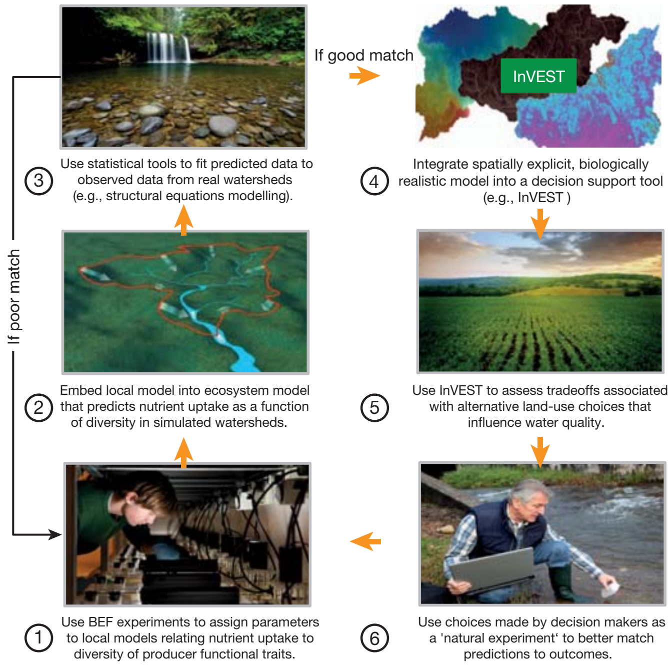

If good match
InVEST
If poor match
3 Use statistical tools to fit predicted data to observed data from real watersheds (e.g., structural equations modelling).
4 Integrate spatially explicit, biologically realistic model into a decision support tool (e.g., InVEST)
2 Embed local model into ecosystem model that predicts nutrient uptake as a function of diversity in simulated watersheds.
5 Use InVEST to assess tradeoffs associated with alternative land-use choices that influence water quality.
1 Use BEF experiments to assign parameters to local models relating nutrient uptake to diversity of producer functional traits.
6 Use choices made by decision makers as a 'natural experiment' to better match predictions to outcomes.

Figure 2 | Towards a better link between BEF and BES research. One of our greatest challenges now is to take what we have learned from 20 years of research and develop predictive models that are founded on empirically quantified mechanisms, and that forecast changes in ecosystem services at scales that are policy-relevant. We outline a hypothetical approach for linking biodiversity to the maintenance of water quality at landscape scales. Data from BEF experiments are used to parameterize competition or niche models that predict how biodiversity has an impact on nutrient assimilation and retention (step 1). Local models are then embedded in spatially explicit meta-community or ecosystem models that incorporate habitat heterogeneity, dispersal and abiotic drivers to predict relationships between biodiversity and water quality at landscape scales (step 2). Predictions of the landscape model are compared to observations from natural systems to assess fit, and statistical tools are used to disentangle effects of biodiversity from other environmental factors (step 3). Once a satisfactory fit is achieved, the model is integrated into a decision support tool (for example, InVEST (step 4)), which is used to simulate changes in ecosystem services at landscape scales where decision makers assess alternative land-use choices (step 5). Choices made by decision makers in real projects provide new data that allow biologists to refine their models and predictions (step 6). Images from B.J.C., G.C.D., US EPA and Shutterstock.com; used with permission.

Ideally, predictions arising from landscape-level models would be 'ground-truthed' by assessing their ability to predict the outcome of real restoration projects, or other management scenarios where policy actions are being taken to protect ecosystem services[90]. For example, given land-use pressure and climate change, freshwater supply is an ecosystem service in high demand, and water funds are becoming a common finance mechanism through which downstream water consumers pay for upstream changes in land use to achieve objectives like maintenance of water quality (nutrient, sediment and bacterial loads)[91]. Major initiatives are underway to standardize the design, implementation and monitoring of water funds, including a pilot programme supported by the World Bank, the Inter-American Development Bank, FEMSA, and The Nature Conservancy that spans 40 Latin American cities.

Initiatives like these represent opportunities to assess and refine our ability to predict biodiversity–ecosystem service relationships on realistic scales in situations where stake holders are expecting positive returns. For example, BEF and BES researchers have amassed substantial experimental evidence showing that species diversity of plants and algae increase uptake of nutrient pollutants from soil and water[12,24,25,33,63]. We have statistical models that quantify the functional form of these effects[12,31], and extensive data on the functional traits that influence such processes in different habitats[53,63,92]. One approach could involve developing spatially explicit predictions of how biodiversity influences water quality in a modelled watershed where local nutrient assimilation and retention are a function of the number and types of functional traits that locally co-occur (that is, traits of plants in a riparian zone, or of algae in a stream reach). One could then integrate this spatially explicit, biologically realistic model into a decision support tool (for example, InVEST (Integrated Valuation of Ecosystem Services and Tradeoffs))[93] to simulate changes in ecosystem services at landscape scales where decision makers can assess trade-offs associated with alternative land-use choices (Fig. 2). Choices made by decision makers in real projects could, in turn, serve as 'natural experiments' that provide biologists with an opportunity to test their predictions against outcomes.

Valuing biodiversity

Economists have developed a wide array of tools to estimate the value of natural and managed ecosystems and the market and non-marketed services that they provide[94]. Although there are good estimates of society's willingness to pay for a number of non-marketed ecosystem services, we still know little about the marginal value of biodiversity (that is, value associated with changes in the variation of genes, species and functional traits) in the production of those services. The economic value of biodiversity loss derives from the value of the affected services. Estimating this value requires calibration of ecosystem service 'production' functions that link biodiversity, ecosystem processes and ecosystem services. The derivative of such functions with respect to biodiversity defines the marginal physical product of biodiversity (for example, carbon sequestration or water purification), and when multiplied by the value of the service, yields the marginal value of biodiversity change.

Researchers in the BEF and BES fields need to work more closely to estimate the marginal value of biodiversity for ecosystem services. In doing so, at least three challenges require attention. First, ecosystems deliver multiple services, and many involve trade-offs in that increasing the supply of one reduces the supply of another. For example, carbon sequestration through afforestation or forest protection may enhance timber production but reduce water supplies[95]. The value of biodiversity change to society depends on the net marginal effect of the change on all ecosystem services[96]. Future work needs to quantify the marginal benefits of biodiversity (in terms of services gained) relative to marginal costs (in terms of services lost).

Furthermore, many trade-offs among services occur at very different spatial and temporal scales. The gains from simplifying ecosystems are often local and short term, whereas the costs are transmitted to people in other locations, or to future generations. For society to make informed choices about land uses that have mixed effects, the science linking biodiversity to ecosystem functioning and services must be extended to explore trade-offs between services at multiple temporal and spatial scales so that information can be incorporated into models of optimal land use.

Finally, there is increasing interest in developing incentives to encourage land holders to take full account of the ecosystem services that are affected by their actions. The concept of 'payments for ecosystem services' has emerged as one tool for bringing market value to ecosystems. Our Review has emphasized that many ecosystem services ultimately depend on the variety of life forms that comprise an ecosystem and that control the ecological processes that underlie all services. Therefore, successful plans to use payments for ecosystem services will need to be founded on a solid understanding of the linkages among biodiversity, ecosystem functioning and the production of ecosystem services[97]. This will require that such plans explicitly manage for biodiversity change.

Responding to the call of policy initiatives

The significance of biodiversity for human wellbeing was recognized 20 years ago with the formation of the Convention on Biological Diversity—an intergovernmental agreement among 193 countries to support the conservation of biological diversity, the sustainable use of its components, and the fair and equitable sharing of benefits. Despite this agreement, evidence gathered in 2010 indicated that biodiversity loss at the global scale was continuing, often at increasing rates[98]. This observation stimulated a set of new targets for 2020 (the Aichi targets) and, in parallel, governments have been negotiating the establishment of a new assessment body, the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). The IPBES will be charged with conducting regional, global and thematic assessments of biodiversity and ecosystem services, and will depend on the international scientific community to assess trends and evaluate risks associated with alternative patterns of development and changes in land use[99].

Significant gaps in both the science and policy need attention if the Aichi targets are to be met, and if future ecosystems are to provide the range of services required to support more people sustainably[99]. We have reported the scientific consensus that has emerged over 20 years of biodiversity research, to help orient the next generation of research on the links between biodiversity and the benefits ecosystems provide to humanity. One of the greatest challenges now is to use what we have learned to develop predictive models that are founded on empirically quantified ecological mechanisms; that forecast changes in ecosystem services at scales that are policy-relevant; and that link to social, economic and political systems. Without an understanding of the fundamental ecological processes that link biodiversity, ecosystem functions and services, attempts to forecast the societal consequences of diversity loss, and to meet policy objectives, are likely to fail[100]. But with that fundamental understanding in hand, we may yet bring the modern era of biodiversity loss to a safe end for humanity.

1. Jones, C. G., Lawton, J. H. & Shachak, M. Organisms as ecosystem engineers. *Oikos* **69,** 373–386 (1994).
2. Sterner, R. W. & Elser, J. J. *Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere* (Princeton Univ. Press, 2002).
3. Power, M. E. *et al.* Challenges in the quest for keystones. *Bioscience* **46,** 609–620 (1996).
4. Schulze, E. D. & Mooney, H. A. *Biodiversity and Ecosystem Function* (Springer, 1993).
   This influential book established many of the original hypotheses and ideas that laid the foundation for two decades of empirical work in BEF.
5. Heywood, V. H. (ed.) *Global Biodiversity Assessment* (Cambridge Univ. Press, 1995).
6. Loreau, M. *et al.* DIVERSITAS Report No. 1: DIVERSITAS Science Plan. (2002).
7. Tilman, D. & Downing, J. A. Biodiversity and stability in grasslands. *Nature* **367,** 363–365 (1994).
   This study, along with ref. 8, started a generation of research that examined how biodiversity influences the functioning of ecosystems.
8. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368,** 734–737 (1994).
9. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379,** 718–720 (1996).
10. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286,** 1123–1127 (1999).
11. Loreau, M., Naeem, S. & Inchausti, P. *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives* (Oxford Univ. Press, 2002).
    This book, which followed a 2000 conference in Paris, summarized the first decade of BEF research.
12. Cardinale, B. J. *et al.* The functional role of producer diversity in ecosystems. *Am. J. Bot.* **98,** 572–592 (2011).
13. Daily, G. C. *Nature's Services: Societal Dependence on Natural Ecosystems* (Island Press, 1997).
    This book cemented the notion that natural habitats provide essential goods services to society, and it helped to make ecosystem services a mainstream term.
14. Perrings, C., Folke, C. & Maler, K. G. The ecology and economics of biodiversity loss—The research agenda. *Ambio* **21,** 201–211 (1992).
15. Mace, G. M., Norris, K. & Fitter, A. H. Biodiversity and ecosystem services: a multilayered relationship. *Trends Ecol. Evol.* **27,** 19–26 (2012).
16. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis* (World Resources Institute, 2005).
17. Kinzig, A. P., Pacala, S. W. & Tilman, D. *The Functional Consequences of Biodiversity: Empirical Progress and Theoretical Extensions* (Princeton Univ. Press, 2002).
18. Loreau, M. *From Populations to Ecosystems: Theoretical Foundations for a New Ecological Synthesis* (Princeton Univ. Press, 2010).
19. Tilman, D., Lehman, D. & Thompson, K. Plant diversity and ecosystem productivity: Theoretical considerations. *Proc. Natl Acad. Sci. USA* **94,** 1857–1861 (1997).
20. Paquette, A. & Messier, C. The effect of biodiversity on tree productivity: from temperate to boreal forests. *Glob. Ecol. Biogeogr.* **20,** 170–180 (2011).
    This paper, along with ref. 21, exemplifies how to quantify biodiversity effects on ecosystem functions at large scales in real ecosystems.
21. Maestre, F. T. *et al.* Plant species richness and ecosystem multifunctionality in global drylands. *Science* **335,** 214–218 (2012).
22. Mora, C. *et al.* Global human footprint on the linkage between biodiversity and ecosystem functioning in reef fishes. *PLoS Biol.* **9,** e1000606 (2011).
23. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecol. Monogr.* **75,** 3–35 (2005).
    This paper was the last published scientific consensus statement on how biodiversity influences ecosystem functions and services.
24. Balvanera, P. *et al.* Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol. Lett.* **9,** 1146–1156 (2006).
    This paper, along with ref. 25, was the first to synthesize BEF research via statistical meta-analyses.
25. Cardinale, B. J. *et al.* Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature* **443,** 989–992 (2006).
26. Worm, B. *et al.* Impacts of biodiversity loss on ocean ecosystem services. *Science* **314,** 787–790 (2006).
27. Cardinale, B. J. *et al.* Impacts of plant diversity on biomass production increase through time due to complementary resource use: A meta-analysis. *Proc. Natl Acad. Sci. USA* **104,** 18123–18128 (2007).
28. Stachowicz, J., Bruno, J. F. & Duffy, J. E. Understanding the effects of marine biodiversity on communities and ecosystems. *Annu. Rev. Ecol. Evol. Syst.* **38,** 739–766 (2007).
29. Bruno, J. F. & Cardinale, B. J. Cascading effects of predator richness. *Front. Ecol. Environ* **6,** 539–546 (2008).
30. Cardinale, B. J. *et al.* in *Biodiversity and Human Impacts* (eds Naeem, S. *et al.*) 105–120 (Oxford Univ. Press, 2009).
31. Schmid, B. *et al.* in *Biodiversity and Human Impacts* (eds Naeem, S. *et al.*) 14–29 (Oxford Univ. Press, 2009).
32. Srivastava, D. S. *et al.* Diversity has stronger top-down than bottom-up effects on decomposition. *Ecology* **90,** 1073–1083 (2009).
33. Quijas, S., Schmid, B. & Balvanera, P. Plant diversity enhances provision of ecosystem services: A new synthesis. *Basic Appl. Ecol.* **11,** 582–593 (2010).
34. Cadotte, M. W., Cardinale, B. J. & Oakley, T. H. Evolutionary history and the effect of biodiversity on plant productivity. *Proc. Natl Acad. Sci. USA* **105,** 17012–17017 (2008).
35. Flynn, D. F. B., Mirotchnick, N., Jain, M., Palmer, M. I. & Naeem, S. Functional and phylogenetic diversity as predictors of biodiversity-ecosystem-function relationships. *Ecology* **92,** 1573–1581 (2011).
36. Wardle, D. A., Bonner, K. I. & Nicholson, K. S. Biodiversity and plant litter: Experimental evidence which does not support the view that enhanced species richness improves ecosystem function. *Oikos* **79,** 247–258 (1997).
37. Ives, A. R. & Carpenter, S. R. Stability and diversity of ecosystems. *Science* **317,** 58–62 (2008).
38. Cottingham, K. L., Brown, B. L. & Lennon, J. T. Biodiversity may regulate the temporal variability of ecological systems. *Ecol. Lett.* **4,** 72–85 (2001).
39. Jiang, L. & Pu, Z. C. Different effects of species diversity on temporal stability in single-trophic and multitrophic communities. *Am. Nat.* **174,** 651–659 (2009).
40. Hector, A. *et al.* General stabilizing effects of plant diversity on grassland productivity through population asynchrony and overyielding. *Ecology* **91,** 2213–2220 (2010).
41. Campbell, V., Murphy, G. & Romanuk, T. N. Experimental design and the outcome and interpretation of diversity-stability relations. *Oikos* **120,** 399–408 (2011).
42. Griffin, J. N. *et al.* in *Biodiversity and Human Impacts* (eds Naeem, S. *et al.*) 78–93 (Oxford Univ. Press, 2009).
43. Doak, D. F. *et al.* The statistical inevitability of stability-diversity relationships in community ecology. *Am. Nat.* **151,** 264–276 (1998).

44. Gonzalez, A. & Loreau, M. The causes and consequences of compensatory dynamics in ecological communities. *Annu. Rev. Ecol. Evol. Syst.* **40,** 393–414 (2009).
45. Duffy, J. E. Why biodiversity is important to the functioning of real-world ecosystems. *Front. Ecol. Environ* **7,** 437–444 (2009).
46. Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294,** 843–845 (2001).
This experiment continues to be one of the largest and longest running biodiversity studies ever conducted.
47. Huston, M. A. Hidden treatments in ecological experiments: Re-evaluating the ecosystem function of biodiversity. *Oecologia* **110,** 449–460 (1997).
This paper raised several criticisms against early BEF research, which forced the reconsideration of conclusions with better experiments and more rigorous data analyses.
48. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412,** 72–76 (2001).
49. Carroll, I. T., Cardinale, B. J. & Nisbet, R. M. Niche and fitness differences relate the maintenance of diversity to ecosystem function. *Ecology* **92,** 1157–1165 (2011).
50. Shurin, J. B. *et al.* A cross-ecosystem comparison of the strength of trophic cascades. *Ecol. Lett.* **5,** 785–791 (2002).
51. Estes, J. A. *et al.* Trophic downgrading of planet earth. *Science* **333,** 301–306 (2011).
This paper summarizes how the extinction of large carnivores has an impact on ecosystem processes, emphasizing the urgent need to integrate trophic interactions into BEF and BES research.
52. Duffy, J. E. *et al.* The functional role of biodiversity in ecosystems: Incorporating trophic complexity. *Ecol. Lett.* **10,** 522–538 (2007).
53. Diaz, S. *et al.* Incorporating plant functional diversity effects in ecosystem service assessments. *Proc. Natl Acad. Sci. USA* **104,** 20684–20689 (2007).
This paper outlined a framework for linking species functional traits to ecosystem services, which moves the field of BES research towards more predictive models.
54. Schmid, B., Hector, A., Saha, P. & Loreau, M. Biodiversity effects and transgressive overyielding. *J. Plant Ecol.* **1,** 95–102 (2008).
55. Suding, K. N. *et al.* Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. *Glob. Change Biol.* **14,** 1125–1140 (2008).
56. Tilman, D., Reich, P. & Isbell, F. Biodiversity impacts ecosystem productivity as much as resources, disturbance or herbivory. *Proc. Natl Acad. Sci. USA.* (in the press).
57. Hooper, D. U. *et al.* A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature* http://dx.doi.org/10.1038/nature11118 (2 May 2012).
58. Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, 2007).
59. Stachowicz, J. J., Graham, M., Bracken, M. E. S. & Szoboszlai, A. I. Diversity enhances cover and stability of seaweed assemblages: The role of heterogeneity and time. *Ecology* **89,** 3008–3019 (2008).
60. Dimitrakopoulos, P. G. & Schmid, B. Biodiversity effects increase linearly with biotope space. *Ecol. Lett.* **7,** 574–583 (2004).
61. Venail, P. A., Maclean, R. C., Meynard, C. N. & Mouquet, N. Dispersal scales up the biodiversity-productivity relationship in an experimental source-sink metacommunity. *Proc. R. Soc. Lond. B* **277,** 2339–2345 (2010).
62. Tylianakis, J. M. *et al.* Resource heterogeneity moderates the biodiversity-function relationship in real world ecosystems. *PLoS Biol.* **6,** e122 (2008).
63. Cardinale, B. J. Biodiversity improves water quality through niche partitioning. *Nature* **472,** 86–89 (2011).
64. Finke, D. L. & Snyder, W. E. Niche partitioning increases resource exploitation by diverse communities. *Science* **321,** 1488–1490 (2008).
65. Hector, A. & Bagchi, R. Biodiversity and ecosystem multifunctionality. *Nature* **448,** 188–190 (2007).
66. Zavaleta, E. S., Pasari, J. R., Hulvey, K. B. & Tilman, G. D. Sustaining multiple ecosystem functions in grassland communities requires higher biodiversity. *Proc. Natl Acad. Sci. USA* **107,** 1443–1446 (2010).
67. Isbell, F. *et al.* High plant diversity is needed to maintain ecosystem services. *Nature* **477,** 199–202 (2011).
68. Mace, G. M., Gittleman, J. L. & Purvis, A. Preserving the tree of life. *Science* **300,** 1707–1709 (2003).
69. Díaz, S., Fargione, J., Chapin, F. S. & Tilman, D. Biodiversity loss threatens human well-being. *PLoS Biol.* **4,** 1300–1305 (2006).
70. Letourneau, D. K. *et al.* Does plant diversity benefit agroecosystems? A synthetic review. *Ecol. Appl.* **21,** 9–21 (2011).
71. Keesing, F. *et al.* Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468,** 647–652 (2010).
72. Zhang, W., Ricketts, T. H., Kremen, C., Carney, K. & Swinton, S. M. Ecosystem services and dis-services to agriculture. *Ecol. Econ.* **64,** 253–260 (2007).
73. Latta, L. C. *et al.* Species and genotype diversity drive community and ecosystem properties in experimental microcosms. *Evol. Ecol.* **25,** 1107–1125 (2011).
74. Denoth, M., Frid, L. & Myers, J. H. Multiple agents in biological control: improving the odds? *Biol. Control* **24,** 20–30 (2002).
75. Letourneau, D. K., Jedlicka, J. A., Bothwell, S. G. & Moreno, C. R. Effects of natural enemy biodiversity on the suppression of arthropod herbivores in terrestrial ecosystems. *Annu. Rev. Ecol. Evol. Syst.* **40,** 573–592 (2009).
76. Vance-Chalcraft, H. D., Rosenheim, J. A., Vonesh, J. R., Osenberg, C. W. & Sih, A. The influence of intraguild predation on prey suppression and prey release: A meta-analysis. *Ecology* **88,** 2689–2696 (2007).
77. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. J. Risk factors for human disease emergence. *Phil. Trans. R. Soc. Lond. B* **356,** 983–989 (2001).
78. Wardle, D. A., Bardgett, R. D., Callaway, R. M. & Van der Putten, W. H. Terrestrial ecosystem responses to species gains and losses. *Science* **332,** 1273–1277 (2011).
79. Solan, M. *et al.* Extinction and ecosystem function in the marine benthos. *Science* **306,** 1177–1180 (2004).
80. Bunker, D. E. *et al.* Species loss and aboveground carbon storage in a tropical forest. *Science* **310,** 1029–1031 (2005).
81. Ives, A. R. & Cardinale, B. J. Food-web interactions govern the resistance of communities after non-random extinctions. *Nature* **429,** 174–177 (2004).
82. Sax, D. F. & Gaines, S. D. Species diversity: from global decreases to local increases. *Trends Ecol. Evol.* **18,** 561–566 (2003).
83. Bardgett, R. D. & Wardle, D. A. *Aboveground-belowground Linkages: Biotic Interactions, Ecosystem Processes, and Global Change* (Oxford Univ. Press, 2010).
84. Vilà, M. *et al.* Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. *Ecol. Lett.* **14,** 702–708 (2011).
85. Fox, J. W. & Kerr, B. Analyzing the effects of species gain and loss on ecosystem function using the extended Price equation partition. *Oikos* **121,** 290–298 (2012).
86. Loeuille, N. & Loreau, M. Evolutionary emergence of size-structured food webs. *Proc. Natl Acad. Sci. USA* **102,** 5761–5766 (2005).
87. Berlow, E. L. *et al.* Simple prediction of interaction strengths in complex food webs. *Proc. Natl Acad. Sci. USA* **106,** 187–191 (2009).
88. O'Gorman, E. J., Jacob, U., Jonsson, T. & Emmerson, M. C. Interaction strength, food web topology and the relative importance of species in food webs. *J. Anim. Ecol.* **79,** 682–692 (2010).
89. Wood, S. A., Lilley, S. A., Schiel, D. R. & Shurin, J. B. Organismal traits are more important than environment for species interactions in the intertidal zone. *Ecol. Lett.* **13,** 1160–1171 (2010).
90. Kinzig, A. P. *et al.* Paying for ecosystem services-promise and peril. *Science* **334,** 603–604 (2011).
91. Goldman-Benner, R. *et al.* Water funds and PES: Practice learns from theory and theory can learn from practice. *Oryx* **46,** 55–63 (2012).
92. Kattge, J. *et al.* TRY—a global database of plant traits. *Glob. Change Biol.* **17,** 2905–2935 (2011).
93. Kareiva, P., Tallis, H., Ricketts, T., Daily, G. & Polasky, S. *Natural Capital: Theory & Practice of Mapping Ecosystem Services* (Oxford Univ. Press, 2011).
This book summarizes the state-of-the-art in modelling ecosystem services.
94. Heal, G. M. *et al. Valuing Ecosystem Services: Toward Better Environmental Decision Making* (The National Academies Press, 2005).
95. Jackson, R. B. *et al.* Trading water for carbon with biological carbon sequestration. *Science* **310,** 1944–1947 (2005).
96. Perrings, C. *et al.* Ecosystem services, targets, and indicators for the conservation and sustainable use of biodiversity. *Front. Ecol. Environ* **9,** 512–520 (2011).
97. Kinzig, A. P. *et al.* Ecosystem services: Free lunch no more response. *Science* **335,** 656–657 (2012).
98. Butchart, S. H. M. *et al.* Global biodiversity: Indicators of recent declines. *Science* **328,** 1164–1168 (2010).
99. Perrings, C., Duraiappah, A., Larigauderie, A. & Mooney, H. The biodiversity and ecosystem services science-policy interface. *Science* **331,** 1139–1140 (2011).
100. Larigauderie, A. *et al.* Biodiversity and ecosystem services science for a sustainable planet: The DIVERSITAS vision for 2012–20. *Curr. Opin. Environ. Sust.* **4,** 101–105 (2012).

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

Acknowledgements This work was conceived as a part of the working group, Biodiversity and the Functioning of Ecosystems: Translating Model Experiments into Functional Reality, supported by the National Center for Ecological Analysis and Synthesis, a Center funded by the National Science Foundation (NSF Grant EF-0553768), the University of California, Santa Barbara, and the State of California. Additional funds were provided by NFS' DIMENSIONS of Biodiversity program to BJC (DEB-104612), and by the Biodiversity and Ecosystem Services Research Training Network (BESTNet) (NSF Grant 0639252). The use of trade names is for descriptive purposes only and does not imply endorsement by the US Government.

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 B.C. (bradcard@umich.edu).

doi:10.1038/nature11157

Securing natural capital and expanding equity to rescale civilization

Paul R. Ehrlich[1], Peter M. Kareiva[2] & Gretchen C. Daily[3]

In biophysical terms, humanity has never been moving faster nor further from sustainability than it is now. Our increasing population size and per capita impacts are severely testing the ability of Earth to provide for peoples' most basic needs. Awareness of these circumstances has grown tremendously, as has the sophistication of efforts to address them. But the complexity of the challenge remains daunting. We explore prospects for transformative change in three critical areas of sustainable development: achieving a sustainable population size and securing vital natural capital, both in part through reducing inequity, and strengthening the societal leadership of academia.

"For most of the last century, economic growth was fuelled by what seemed to be a certain truth: the abundance of natural resources. We mined our way to growth. We burned our way to prosperity. We believed in consumption without consequences. Those days are gone … Over time, that model is a recipe for national disaster. It is a global suicide pact." United Nations (UN) Secretary General Ban Ki-Moon addressing The World Economic Forum, 29 January 2011.

'Sustainability' has become a remarkably popular word, now featured on over 100 million websites, yet there are huge challenges in putting it into successful practice. Dictionaries define sustainability in terms of maintaining valued qualities without interruption, weakening or loss. The most widely used form of the concept can be traced to the Brundtland Report[1], where sustainability is described as human development that "meets the needs and aspirations of the present without compromising the ability of future generations to meet their own needs." Although there has been some dissatisfaction with this definition (for example, see refs 2–4), it has moved rapidly into the mainstream. The 2005 World Summit recognized that sustainability requires reconciling environmental, social equity and economic demands—the three 'pillars' or 'Es' of sustainability, or the 'triple bottom line'[5].

Given the contention associated with fertility decisions, it is amazing to see that reductions in fertility, and hence population pressure, have actually proven tractable. Many countries have achieved rapid fertility declines, with attendant social changes in equity that could potentially be spread much more widely[6,7]. By contrast, ever-rising consumption rates are proving extremely difficult to check. One barrier is that the consumer culture of developed countries represents the "needs and aspirations" of their populations today, but rarely considers the needs and aspirations of future people. A second pernicious obstacle is the inexorable spread of the developed world's consumption patterns to developing nations as they gain wealth[8]. If consumption is not brought in line with reduced population pressure, there is little hope of major gains in sustainability.

The urgency for the Rio+20 Earth Summit concerns whether we can find ways to change our economic and social systems, such that we reduce strains on the environment, that are durable and timely; that is, before severe impairment of Earth's 'life-support systems'. In some senses, the environmental demand pillar of sustainability is non-negotiable—we cannot change the physics of climate and other laws of nature—but we can change human social and economic systems. Here we briefly sketch some essential features of conceptual thought and quantitative analysis on sustainability, and explore promising pathways for deep and lasting transformation.

The fundamentals of sustainability

Numerous approaches have been developed to estimate the numbers of people and lifestyles that can be sustained. We highlight both the founding principles of sustainability and some key ideas that have emerged more recently.

Multiplicative drivers

IPAT—which posits that society's impact (I) on Earth's life-support systems is a product of population size (P), per capita consumption ('affluence'; A), and a 'technology' factor (T) that reflects the environmental impact caused by technologies, cultural practices, and institutions through which each unit of consumption is generated—was introduced just before the first Earth Summit, convened in Stockholm in 1972[9]. It's purpose was to counter claims on the one hand that global environmental problems were the result of runaway population growth in poor countries and, on the other, that the sole problem was environmentally damaging technologies[10].

Then, and still today, although the developing world contains the majority (approximately 80%) of the global population, the impact of the developed world is far greater. Using energy consumption as a surrogate for per capita impacts, the current average impact of each inhabitant of developed countries exceeds that of inhabitants of developing countries 2–14 fold[11]. The importance of the A and T factors in reducing impacts is even greater today than it was when the concept was first presented, and has stimulated innovative research, in the social sciences for example[12].

Limits of Earth's life-support capabilities

Carrying capacity is a foundational concept for characterizing the dynamics of populations and limiting resources. A famous, early analysis of humanity's interaction with limits received harsh criticism, but recent evaluation suggests that "the values predicted by the limits-to-growth model and actual data for 2008 are very close."[13]. Similarly, although carrying capacity is difficult to measure, it is clear that the human population's size and consumption patterns are well above what Earth could support without impairment of vital life-support systems[14,15],

[1]Department of Biology, 371 Serra Mall, Stanford University, Stanford, California 94305, USA. [2]Peter M. Kareiva, The Nature Conservancy, 4722 Latona Ave. NE, Seattle, Washington 98105, USA. [3]Department of Biology and Woods Institute, 371 Serra Mall, Stanford University, Stanford, California 94305, USA.

exceeding "planetary boundaries"[16]. According to ecological footprint analysis, rich countries use 2–5 times their per capita equitable 'Earth shares', were the carrying capacity of the planet divided into 7 billion equal parts[17].

An interesting new development is the extent to which the business world has started to worry about the implications of resource scarcity for their own enterprises[18]. The private sector's awakening to resource limitation has created recognition that improved management of natural systems could be key to addressing resource scarcity[18]. But their interest is broader, increasingly exploring what people would like to sustain and how to reach agreement on this, constrained by estimates of what is feasible. The question is whether societies can develop a widely supported vision of the targets for rescaling, in terms of population size, prosperity, equity and risk of reducing Earth's carrying capacity.

Sustainable futures and how to get to them

Backcasting is a way of exploring alternative desired futures (for 2050, say) and then determining in fairly specific terms the 'must haves' decade by decade in order to arrive at a given future. A procedural model is the Vision 2050 effort of the World Business Council for Sustainable Development[19]. Backcasting exercises could help people, governments and corporations to recognize the values of natural capital in supporting human well-being, and more routinely incorporate protecting these values, demographics, patterns of consumption and environmental constraints into planning and decision-making. Other approaches are needed to address the multitudinous problems of global governance[20].

Buffering as a goal

Resilience is defined as the ability of a system to absorb disturbance and recover from it or withstand it while maintaining the same basic structure and functioning. Through the lens of coupled human and natural systems, declines in resilience increase the risk that societies and their supporting ecosystems will not recover from a certain magnitude shock[21]. Resilience expands attention on economic growth and efficiency to encompass flexibility and capacity for recovery from inevitable shocks[22,23].

An emerging challenge to resilience is the global connectedness of ecosystems and economies, so that shocks in one place can rapidly spread across continents and around the world. Although the banking crisis is the most dramatic example of global propagation of shocks, demand for biofuels and food is now leading to massive conversion of forests[24], including by relatively rich nations purchasing and converting vast tracts of land in poor nations[25].

Equity and justice

The distribution of wealth and power permeates social, economic and ecological thinking[6,26–29]. If the biomass of an ecosystem is concentrated in a few species, the wealth of a nation amassed in one bank, or the environmental hazard of a place directed at one subpopulation, then there is little opportunity to absorb and mitigate disasters when they come[30]. The uneven distribution of climate risk, for example, poses especially severe challenges to political and social stability, as it falls most heavily on the world's poorest nations (Fig. 1). In particular, nations with the lowest Human Development Index (HDI) tend to be nations that in the near term will face the greatest risk of climate stress from impaired agricultural production, sea level rise, and extreme weather events. The combination of enormous inequity in wealth and environmental hazard means that there will be flashpoints for ecological and human disaster.

Natural capital and ecosystem services

Earth's lands and waters and their biodiversity can be seen as a capital stock from which people derive vital ecosystem services. These include the production of goods (such as food, timber and industrial products), regulating services (such as water purification, crop pollination and coastal protection), cultural benefits (such as inspiration and recreation), and preservation of options (such as genetic diversity for future use). This basic framework, dating back at least to the 1970s[31–33], has recently been greatly enhanced with tools and approaches for quantifying the tradeoffs associated with alternative scenarios or choices, to inform business and policy decisions[34,35]. The Millennium Ecosystem Assessment evaluated for the first time the status of the world's ecosystem services, concluding that two-thirds were declining[36].

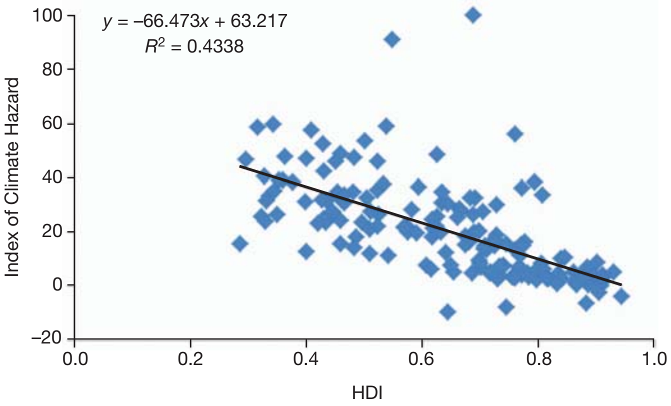


Figure 1 | The relationship between national-level poverty (as measured by HDI) and vulnerability (as measured by the Index of Climate Hazard). The HDI combines indicators of life expectancy, educational attainment and income into a composite index that ranges between 0 and 1 (data taken from the UN Development Programme Human Development Report; http://hdr.undp.org/en/statistics/hdi/). The Index of Climate Hazard combines three dimensions of climate risk: sea level rise and storm surge, extreme weather events, and reduced agricultural productivity, taken from D. Wheeler[99]; this climate hazard represents the expected near-term increase in risk (that is, from 2008 to 2015). The two outliers are China (0.687, 100) and India (0.547, 90.8), probably owing to their very large populations and large, climatically diverse land areas, serious water problems, and long coast lines.

Taken together, these conceptual and analytical approaches assess the implications of human population and consumption trajectories, and help define the challenges for sustainable development. They reveal the great disparity between rich and poor nations in per capita appropriation of Earth's capacity to support human activity and in vulnerability to natural catastrophes. They indicate that increasing well-being in poor countries will require significant reduction in the deleterious environmental impacts by rich and poor countries alike, and that greatly narrowing the rich–poor gap will be key to achieving sustainability advances in many arenas. We next explore powerful options for achieving sustainable development, looking first at what population sizes are likely and how socially beneficial trajectories can be promoted.

Population futures

Median population projections from the UN Population Division and other sources[37] suggest that the world is most likely to have about 9.5 billion people in 2050 (range of projections: 8.1–10.6 billion), and slightly over 10 billion in 2100 (range: 6–16 billion). Demographer Ronald Lee points out that they implicitly assume no negative economic or environmental feedbacks, although "it is possible that desertification, global warming, shortage of fresh water, extinctions of species, and other man-made degradations of the natural resource base will lead to catastrophic effects on the population and its growth"[38]. He believes that the occurrence of those feedbacks will be determined in part by what policy measures are taken to ameliorate the environmental impacts of population and economic growth[38] (Fig. 2). We know that, in aggregate, each person added to the population will derive food and other resources from poorer sources, generally involving more energy and disproportionate environmental impact[9,39].

One important win–win way to reduce fertility rates is by meeting the 'unmet need' for contraception; that is, by supplying safe, modern means to those who do not want a child in the next two years of their lives but are not using any means of birth control[40]. On the basis of data

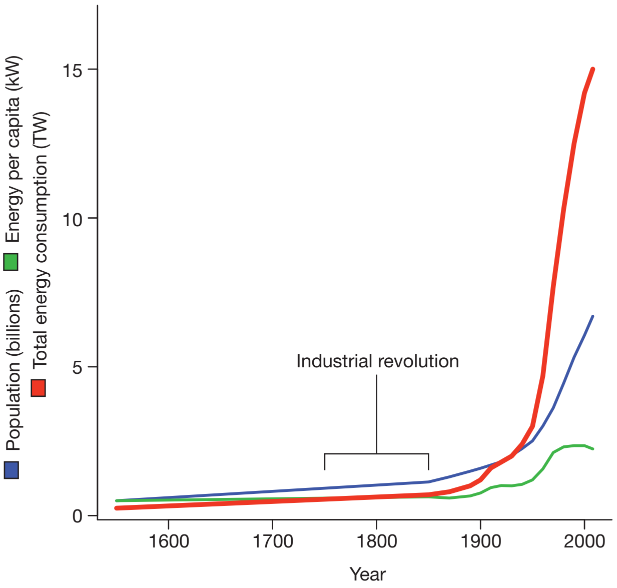


Figure 2 | History of growth in world population and environmental impact of *Homo sapiens*, indicated by its surrogates, per capita and total human energy use. Note the more than 20-fold increase in total energy use since the industrial revolution, with the growth caused slightly more by population increase than by expansion of per capita consumption[100] (Population Reference Bureau, UN, World Population Projections to 2100 (1998), and US Energy Information Administration).

from demographic and health surveys in 53 Asian, African and Latin American countries between 1995 and 2005, an estimated 7–15% of women have an unmet need for contraception. In sub-Saharan Africa, the region where unmet need is greatest, the estimate is about 25% of married women[41].

There are roughly 75 million unintended pregnancies in the world annually and almost half of them end in abortion[42]. Making reproduction education and family planning universally available in the developing world could theoretically avert 20 million or more births annually (estimates vary), avoid over 25 million abortions[43], reduce maternal mortality by 25–40% (ref. 44), and greatly reduce the population growth rate.

A second win–win way to reduce fertility rates is to raise levels of education, especially of young women. Lutz and Samir[45] considered four different educational scenarios. In the most pessimistic scenario, no new schools are built in response to population growth, and educational conditions and enrolment rates deteriorate. In the second, the education system expands just rapidly enough to accommodate the additional students produced by growth. In a third, educational investments permit school expansion rates that were experienced by countries further along in this process. Finally, in the fourth scenario, all countries are presumed to initiate ambitious but practical programs to maximize the rate at which the educational system is expanded and improved.

The results of the analysis are impressive. If there were a crash program of education globally, there would be roughly a billion fewer people in 2050 than if there were no effort to keep educational investment commensurate with population size. Education and subsequent empowerment of women lowers infant and childhood mortality, an effect that is more than offset demographically by the associated growing desire and ability to have fewer children. For example, for Kenya, Lutz and Samir show that the baseline population of 30 million in 2000 would, under the first scenario, "increase to an incredible 114 million." By contrast, adopting the third and fourth scenarios, the population would grow to 84 and 85 million by 2050.

Many potential benefits are associated with these demographic impacts of education. No nation has successfully developed without providing substantial education to women, and educating women leads to better health and nutrition for children, and higher productivity in agriculture and other sectors of society[46,47]. It is also a key step towards reaching the ideal, nowhere yet achieved, of a society in which women are fully equal to men.

Securing natural capital and human well-being

Given likely population trajectories, what natural capital is most vital for sustaining human well-being? This is a key question in many important sub-systems, including in food[48,49], fresh water[50], energy[11], climate[51] and health[52]. Over the past two decades, great advances have been made in each of these areas, with a growing effort at synthesis in the development of ecosystem service science, tools and decision-making approaches[36,53–55].

In theory, if institutions recognize the values of ecosystem services, then we can greatly enhance investments in the natural capital that generates them and foster human well-being at the same time. In practice, we are still in the early stages of developing the scientific basis, and the policy and finance mechanisms, for integrating natural capital into land use and other resource decisions on large scales. Relative to other forms of capital, assets embodied in ecosystems have been poorly understood and scarcely monitored, and are undergoing rapid, unchecked degradation[56–58]. Natural capital and the ecosystem services that flow from it are typically undervalued—by governments, businesses and the public—and recognized only upon their loss[59–62].

The urgent challenge today is to move from theory to real-world honing and implementation of ecosystem service tools and approaches to resource decisions taken by individuals, communities, corporations, governments and other organizations[63]. A great diversity of efforts to implement the ecosystem services framework has emerged worldwide (Box 1). Collectively, they represent a promising shift towards a more inclusive, integrated and effective set of strategies[64–66]. Taken together, these efforts span the globe and target a full suite of ecosystem services, such as carbon sequestration, water supply, flood control, coastal protection and enhancement of scenic beauty (and associated recreation/tourism values)[67,68].

The ecosystem service investments being made in China today stand out in their ambitious goals, scale and duration. Prompted by massive droughts and flooding in 1997–1998, China implemented several national forestry and conservation initiatives to address the nation's growing environmental crises, involving approximately 120 million households and investing approximately 700 billion yuan over 2000–2010[69]. China is currently also undertaking a first national assessment of ecosystem services, spanning a wide range of ecosystems, services and scales. Perhaps most ambitiously, China is establishing a new network of 'ecosystem function conservation areas' (EFCAs; Fig. 3). EFCAs are being zoned so as to focus conservation and restoration in places with high return on investment for public benefit; at the same time, high-impact human activities are being zoned to sustain or enhance natural capital values.

These initiatives have dual goals: to harmonize people and nature by securing critical natural capital, and to alleviate poverty. Specifically, the government aims to protect ecosystems and their biodiversity for flood control, hydropower production efficiency, irrigation supply, more productive agriculture and tourism. In addition, it aims to open non-farm sectors, increase household income and make land-use practices more sustainable in rural areas[70]. Although these initiatives represent a massive scientific and policy undertaking, there is still little understanding of the local costs of implementation, or their effects on poor and vulnerable populations in or near the target areas. The EFCA model represents a new paradigm for integrating conservation and human development, but for this policy innovation to have wide applicability, it will be important to assess and improve local livelihoods[71].

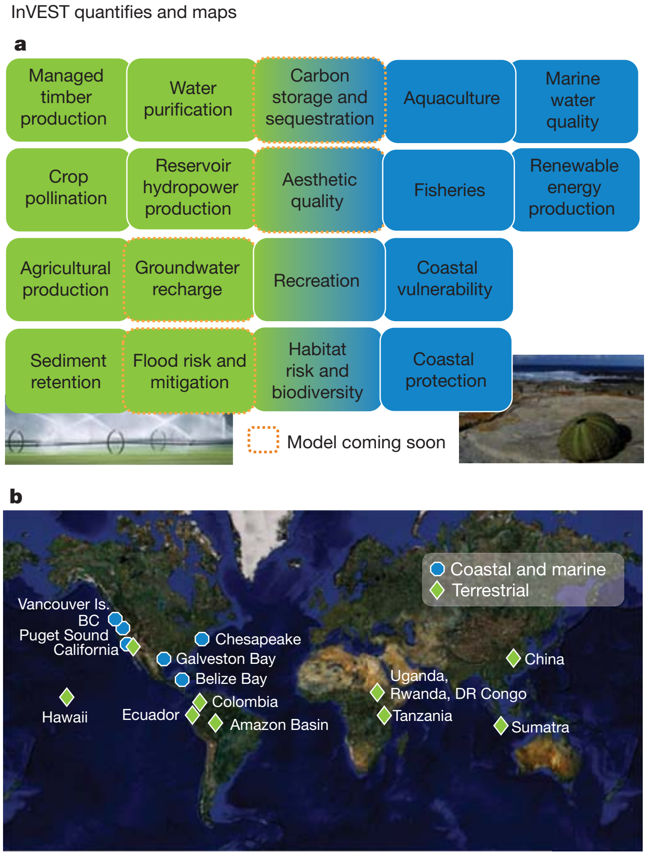
Gender and gender equity in sustainability

As evidence for the value of community-based resource management has accumulated[72], a special role for women in environmental sustainability has begun to emerge. Cross-national studies reveal a strong association between high levels of deforestation and impaired

BOX 1

Quantifying the values of natural capital under future scenarios

The Natural Capital Project, an international partnership, is developing tools for the Integrated Valuation of Ecosystem Services and Tradeoffs (InVEST). These software-based models help decision makers visualize the impacts of potential policies by quantifying and mapping the generation, distribution and economic value of ecosystem services under alternative scenarios[35]. The models span a range of terrestrial and marine services (Box 1 Fig. 1a). They are designed for use in an iterative decision-making process, in which stakeholders identify critical management decisions and explore scenarios of change (for example, demographic, climate, technological). The outputs identify tradeoffs and compatibilities between environmental, economic and social benefits. The models are being applied in a wide range of decision contexts and scales (Box 1 Fig. 1b).



Box 1 Figure | Mapping ecosystem services. a, The suite of InVEST models, created and being improved through an open-source process. **b**, Applications of InVEST models in major policy decisions so far. Many new applications are now being initiated.

health for women, increased household labour and reduced income[73]. Conversely, studies of community-based conservation reveal that the more women are involved in local governance, the more effective forest protection and compliance with regulations[74]. Throughout the world there are differences in the roles of genders in terms of daily time use and activities. For example, in Africa, women are primarily responsible for providing food and childcare[75]. This means that women need to be explicitly targeted when designing strategies for promoting conservation and reducing environmental degradation. There is an interesting parallel between the empowerment of women in cities and reduced fertility, and the empowerment of women in rural poor communities and more sustainable resource use. Equally suggestive is the finding that CO_2 emissions per capita are lower in nations where women have higher political status, so that working to increase gender equity everywhere could interact positively with other steps to achieve sustainability[76].

Academic leadership for rescaling

For at least two decades, the scientific community has largely agreed that humanity is in the midst of an unprecedented slow-motion global emergency[77,78]. The question is then how the academic community can be more effective in stimulating innovation, and in testing promising new approaches in major demonstrations that integrate the biophysical, economic and social pillars of sustainability effectively. At the interface of science and policy, analysis and real decisions, we need ambitious scaling of efforts that embody three core elements, developed recently in the sustainability realm.

First, there is tremendous potential for innovation and real-world implementation in university partnerships with non-governmental organizations (NGOs), community organizations, government agencies and, increasingly, human development organizations and businesses, internationally. In many cases what is needed are boundary institutions that serve to link science and public policy[79].

Second, we need to test and hone ideas in compelling models of success, in the context of live policy opportunities and decisions. Such demonstrations would ideally involve (1) a major policy opportunity with clear and relevant objectives (for example, mainstreaming natural capital into decisions); (2) the strength and commitment among partners necessary for success; (3) a diversity of places and sectors; (4) potential for adapting, replicating and scaling up the approach; and (5) the opportunity for learning, involving input, evaluation and synthesis across academic disciplines, economic sectors and social classes, well beyond the realm of academia. Such demonstrations would involve an iterative process for engaging stake-holders in an 'end-to-end' way, so that the work is conducted jointly throughout[80,81].

Third, the traditional distinction between basic and applied research does not fit our modern sustainability challenge. Instead what is needed is best thought of as use-inspired research[82,83]. This type of research is conducted in a collaborative setting, takes advantage of existing and well-functioning institutions, and is aimed at widely recognized decision points or tradeoffs and conflicts.

By approaches such as these, academics can play major, collaborative roles in transforming the dominant social paradigm globally by opening new options and incentives to change[84,85]. Science indicates that continuing on our current course, or making just incremental changes, are great gambles, and gambles whose odds can be seriously miscalculated[86]. Biophysical problems interact tightly with human governance systems, institutions and civil societies that are inadequate to deal with them[25,87].

Meanwhile, there are aspects of society that require greater development and promotion—such as knowledge, education, health, security, equity and population stability[88]. Development in these sectors implies evolving new sets of norms that make redistribution of power and wealth more acceptable to the rich, and forgoing a full repeat of the Victorian industrial revolution more acceptable to the poor. In the past, the technical community has generally been good at alerting society to threats such as loss of biodiversity and climate disruption, but we know that simply describing an environmental situation scientifically does not necessarily change human behaviour, and that human beings cannot be counted on to behave rationally[89].

The Millennium Alliance for Humanity and the Biosphere is bringing social scientists and humanities scholars into the effort to understand and spark cultural evolution for rescaling. There is already much knowledge that could be used to accelerate the movement[90]. Culture change can involve the dissemination of provincial norms[91], the use of deliberative polling[92], applying lessons from history[93,94], using classroom exercises to change attitudes[95], decreasing the chance of large catastrophes through decentralization[96], and perhaps foremost developing new narratives to

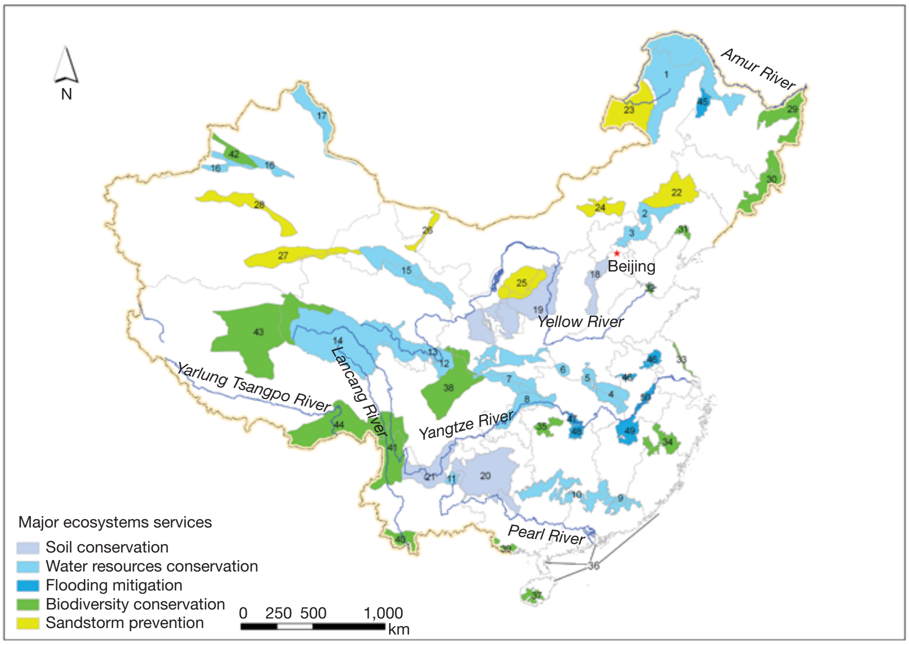


Figure 3 | China's new system of EFCAs. As delineated by the Ministry of Environmental Protection and the Chinese Academy of Sciences in 2007, EFCAs span 24% of China's land area and 25% (708) of its counties. EFCAs are designed to secure biodiversity, soils and water resources, and to mitigate floods and prevent sandstorms. The implementation of EFCAs also serves a major social goal of alleviating poverty. Figure is a modified version of a map provided courtesy of Z. Ouyang, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences.

communicate needs, goals and desired futures, for example see ref. 97. The embedding of environmental concerns into public discourse and policy less than a quarter of a century after 1960 is in itself a fine example of the speed with which cultural change towards sustainability can occur. Now we need to see if we can go much further, in far less time.

A crucial time

There will be a critical opportunity at Rio+20 to persuade leaders—individuals and institutions—to push for the high-level policy changes that are both practical (within reach) and that would greatly accelerate the rescaling of society[98]. The connectivity of the world is increasing and many of the most important environmental problems are global. As all individuals inevitably have an impact on Earth's life-support systems (although in different ways and to varying degrees), it is in everybody's interest to reduce ethically both the size of the population and our per capita impacts. The main uncertainty in the human future is tied to how equitable the pathways chosen for that rescaling are, the degree of overshoot that occurs, and the amount of irreversible damage that overshoot inflicts on Earth's life-support systems.

Rapid change has occurred enough times in human history in relation to fundamental aspects of culture to give hope that such change can be triggered now. One need only consider the advances in women's rights of the past century, the transformation of the racial situation in the United States such that an African American can be elected President, and the collapse of the Soviet Union, to see that cultural change does not necessarily proceed at a glacial pace. And, just as climate change is speeding the flow of glaciers, it should speed the transition of the human enterprise towards a sustainable scale—at which care for all human beings and the natural capital upon which they depend is at the top of the political agenda. The choice seems stark and clear enough: rescaling or global bust.

1. World Commission on Environment and Development. *Our Common Future* (Oxford Univ. Press, 1987).
2. Levin, S. A. Science and sustainability. *Ecol. Appl.* **3,** 545–546 (1993).
3. Fuentes, R. E. Scientific research and sustainable development. *Ecol. Appl.* **3,** 576–577 (1993).
4. Johnston, P., Everard, M., Santillo, D. & Robèrt, K. Reclaiming the definition of sustainability. *Environ. Sci. Pollut. Res. Int.* **14,** 60–66 (2007).
5. Holdren, J. P. Science and technology for sustainable well-being. *Science* **319,** 424–434 (2008).
6. Daily, G. C. & Ehrlich, P. R. Socioeconomic equity, sustainability, and Earth's carrying capacity. *Ecol. Appl.* **6,** 991–1001 (1996).
7. Campbell, M., Cleland, J., Ezeh, A. & Prata, N. Return of the population growth factor. *Science* **315,** 1501–1502 (2007).
8. Spence, M. *The Next Convergence: The Future of Economic Growth in a Multispeed World* (Farrar, Straus, and Giroux, 2011).
9. Ehrlich, P. R. & Holdren, J. Impact of population growth. *Science* **171,** 1212–1217 (1971).
   Foundational discussion of population–consumption–environment interactions.
10. Commoner, B. *The Closing Circle* (Knopf, 1971).
11. International Energy Agency. *Key World Energy Statistics* (International Energy Agency, 2011).
12. Rosa, E. A., York, R. & Dietz, T. Tracking the anthropogenic drivers of ecological impacts. *Ambio* **333,** 509–512 (2004).
    Social scientists expand IPAT to permit hypothesis testing.
13. Hall, C. A. S. & Day, J. W. Jr. Revisiting the limits to growth after peak oil. *Am. Sci.* **97,** 230–237 (2009).
    Re-examination shows original study to be essentially correct.
14. Daily, G. C. & Ehrlich, P. R. Population, sustainability, and Earth's carrying capacity. *Bioscience* **42,** 761–771 (1992).
15. Arrow, K. *et al.* Economic growth, carrying capacity, and the environment. *Science* **268,** 520–521 (1995).
16. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).
    A framework for defining preconditions for human development to avoid critical thresholds of human impact on the environment.
17. Rees, W. E. in *Encyclopedia of Biodiversity* 2nd edn (ed. Levin. S.) (Academic, 2011).
18. Dobbs, R., Oppenheim, J., Thompson, F., Brinkman, M. & Zornes, M. *Resource Revolution: Meeting the World's Energy, Materials, Food, and Water Needs* (McKinsey Global Institute, 2011).
19. World Business Council for Sustainable Development. *Vision 2050: The New Agenda for Business* (WBCSD, 2010).
20. Barrett, S. *Environment and Statecraft: The Strategy of Environmental Treaty-Making* (Oxford Univ. Press, 2003).
21. Liu, J. *et al.* Complexity of coupled human and natural systems. *Science* **317,** 1513–1516 (2007).
    A framework and case studies for understanding complex patterns and dynamics of coupled human and natural systems.
22. Levin, S. A. *et al.* Resilience in natural and socioeconomic systems. *Environ. Dev. Econ.* **3,** 221–262 (1998).
23. Folke, C. *et al.* Reconnecting to the biosphere. *Ambio* **40,** 719–738 (2011).
24. UN Department of Economic and Social Affairs. *World Economic and Social Survey 2011: The Great Green Technological Transformation* (United Nations, 2011).

25. Klare, M. T. *The Race for What's Left: The Global Scramble for the World's Last Resources* (Metropolitan Books, 2012).
Excellent overview of declining marginal returns in resource extraction.
26. Thurow, L. *Generating Inequaltiy: Mechanisms of Distribution in the U.S. Economy* (Basic Books, 1975).
27. Walzer, M. *Spheres of Justice: A Defense of Pluralism and Equality* (Basic Books, 1984).
28. Frank, R. H. *Falling Behind: How Rising Inequality Harms the Middle Class* (Univ. of California Press, 2007).
29. Wilkinson, R. & Pickett, K. *The Spirit Level: Why More Equal Societies Almost Always Do Better* (Penguin, 2009).
30. May, R. M., Levin, S. & Sugihara, G. Ecology for bankers. *Nature* **451,** 893–895 (2008).
31. Study of Critical Environmental Problems. *Man's Impact on the Global Environment* (MIT Press, 1970).
32. Holdren, J. P. & Ehrlich, P. R. Human population and the global environment. *Am. Sci.* **62,** 282–292 (1974).
33. Ehrlich, P. R. & Mooney, H. M. Extinction, substitution and ecosystem services. *Bioscience* **33,** 248–254 (1983).
34. Polasky, S. *et al.* Where to put things? Spatial land management to sustain biodiversity and economic returns. *Biol. Conserv.* **141,** 1505–1524 (2008).
35. Kareiva, P., Tallis, H., Ricketts, T. H., Daily, G. C. & Polasky, S. *Natural Capital: Theory and Practice of Mapping Ecosystem Services* (Oxford Univ. Press, 2011).
36. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Synthesis* (Island, 2005).
First global assessment of natural capital and ecosystem services.
37. Population Reference Bureau. *2011 World Population Data Sheet* (Population Reference Bureau, 2011).
38. Lee, R. The outlook for population growth. *Science* **333,** 569–573 (2011).
39. Harte, J. Human population as a dynamic factor in environmental degradation. *Popul. Environ.* **28,** 223–236 (2007).
40. Potts, M. Where next? *Phil. Trans. R. Soc. B* **364,** 3115–3124 (2009).
41. Sedgh, G., Hussain, R., Bankole, A. & Singh, S. Women with an unmet need for contraception in developing countries and their reasons for not using a method. Report No. 37 (Guttmacher Institute, 2007).
42. Bongaarts, J. Human population growth and the demographic transition. *Phil. Trans. R. Soc. B* **364,** 2985–2990 (2009).
43. Singh, S., Sedgh, G. & Hussain, R. Unintended pregnancy: worldwide levels, trends, and outcomes. *Stud. Fam. Plann.* **41,** 241–250 (2010).
44. Campbell, O. M. R. & Graham, W. J. Strategies for reducing maternal mortality: getting on with what works. *Lancet* **368,** 1284–1299 (2006).
45. Lutz, W. & Samir, K. C. Global human capital: integrating education and population. *Science* **333,** 587–592 (2011).
46. Hill, M. A. & King, E. M. Women's education and economic well-being. *Fem. Econ.* **1,** 21–46 (1995).
47. Brown, L. R. *Plan B 4.0: Mobilizing to Save Civilization* (Norton, 2009).
48. Foley, J. A. *et al.* Solutions for a cultivated planet. *Nature* **478,** 337–342 (2011).
49. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl Acad. Sci. USA* **108,** 20260–20264 (2011).
50. Gleick, P. *The World's Water* Vol. 7 (Island, 2012).
51. Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2001: Synthesis Report* (Cambridge Univ. Press, 2001).
52. World Health Organization. *World Health Statistics 2011* (World Health Organization, 2011).
53. Daily, G. & Matson, P. Ecosystem services: from theory to implementation. Special feature. *Proc. Natl Acad. Sci. USA* **105,** 9455–9456 (2008).
54. Sukhdev, P. Costing the Earth. *Nature* **462,** 277 (2009).
55. Perrings, C. *et al.* Ecosystem services for 2020. *Science* **330,** 323–324 (2010).
56. Heal, G. *Nature and the Marketplace: Capturing the Value of Ecosystem Services* (Island, 2000).
57. Mäler, K.-G., Aniyar, S. & Jansson, A. Accounting for ecosystem services as a way to understand the requirements for sustainable development. *Proc. Natl Acad. Sci. USA* **105,** 9501–9506 (2008).
58. Dasgupta, P. *The Place of Nature in Economic Development* 4977–5046 (Elsevier, 2010).
Integrating natural capital into mainstream economics.
59. Daily, G. C. *et al.* The value of nature and the nature of value. *Science* **289,** 395–396 (2000).
60. Balmford, A. *et al.* Economic reasons for preserving biodiversity. *Science* **297,** 950–953 (2002).
61. NRC (National Research Council). *Valuing Ecosystem Services: Toward Better Environmental Decision-Making* (National Academy Press, 2005).
62. Kinzig, A. P. *et al.* Paying for ecosystem services: promise and peril. *Science* **334,** 603–604 (2011).
63. Chapin, F. S. III *et al.* Ecosystem stewardship: sustainability strategies for a rapidly changing planet. *Trends Ecol. Evol.* **25,** 241–249 (2010).
64. Daily, G. C. & Ellison, K. *The New Economy of Nature: The Quest to Make Conservation Profitable* (Island, 2002).
65. Pagiola, S., Arcenas, A. & Platais, G. Can payments for environmental services help reduce poverty? An exploration of the issues and the evidence to date from Latin America. *World Dev.* **33,** 237–253 (2005).
66. Mace, G., Norris, K. & Fitter, A. Biodiversity and ecosystem services: a multilayered relationship. *Trends Ecol. Evol.* **27,** 19–26 (2012).
67. Tallis, H., Goldman, R., Uhl, M. & Brosi, B. Integrating conservation and development in the field: implementing ecosystem service projects. *Frontiers (Boulder)* **7,** 12–20 (2009).
68. Goldman-Benner, R. *et al.* Water funds and PES: practice learns from theory and theory can learn from practice. *Oryx* **46,** 55–63 (2012).
69. Liu, J., Li, S., Ouyang, Z., Tam, C. & Chen, X. Ecological and socioeconomic effects of China's policies for ecosystem services. *Proc. Natl Acad. Sci. USA* **105,** 9489–9494 (2008).
70. Li, J., Feldman, M., Li, S. & Daily, G. C. Rural household income and inequality under payment for ecosystem services: The Sloping Land Conversion Program in Western China. *Proc. Natl Acad. Sci. USA* **108,** 7721–7726 (2011).
71. Li, J. *et al. Rural Household Livelihoods and Environmental Sustainability in the West of China* (Social Sciences Academic Press, 2012).
72. Ostrom, E. A general framework for analyzing sustainability of social-ecological systems. *Science* **325,** 419–422 (2009).
Foundation for integrating social and ecological dimensions of sustainability.
73. Shandra, J. M., Shandra, C. L. & London, B. Women, non-governmental organizations and deforestation: a cross-national study. *Popul. Environ.* **30,** 48–72 (2008).
74. Agarwal, B. Gender and forest conservation: the impact of women's participation in community forest governance. *Ecol. Econ.* **68,** 2785–2799 (2009).
75. Blackden, C. M. & Wodon, Q. *Gender, Time-Use, and Poverty in Sub-Saharan Africa* (World Bank, 2006).
76. Ergas, C. & York, R. Women's status and carbon dioxide emissions: a quantitative cross-national analysis. *Soc. Sci. Res.* http://dx.doi.org/10.1016/j.ssresearch.2012.03.008 (17 March 2012).
77. National Academy of Sciences USA. in *Population Summit of the World's Scientific Academies* (National Academy Press, 1993).
78. Union of Concerned Scientists. *World Scientists' Warning to Humanity* (Union of Concerned Scientists, 1993).
79. Miller, C. Hybrid management: boundary organizations, science policy, and environmental governance in the climate regime. *Sci. Technol. Human Values* **26,** 478–500 (2001).
80. Kindon, S., Pain, R. & Kesby, M. *Participatory Action Research: Approaches and Methods* (Routledge, 2007).
81. Cottam, H. Participatory systems. http://hir.harvard.edu/big-ideas/participatory-systems (Harvard International Review, 2010).
82. Cash, D. *et al.* Knowledge systems for sustainable development. *Proc. Natl Acad. Sci. USA* **100,** 8086–8091 (2003).
83. Clark, W., William, C., Mitchell, R. & Cash, D. in *Global Environmental Assessments: Information and Influence* (eds Mitchell, R., Clark, W., Cash, D. & Dickson, N.) Ch. 1 15–23 (MIT Press, 2006).
84. Clark, W. C. & Dickson, N. M. Sustainability science: the emerging research program. *Proc. Natl Acad. Sci. USA* **100,** 8059–8061 (2003).
85. Matson, P. A. The sustainability transition. *Issues Sci. Technol.* **25,** 39–42 (2009).
86. Weitzman, M. L. On modeling and interpreting the economics of catastrophic climate change. *Rev. Econ. Stat.* **91,** 1–19 (2009).
87. Turner, G. M. A comparison of The Limits to Growth with 30 years of reality. *Glob. Environ. Change* **18,** 397–411 (2008).
88. Kates, R. W., Parris, T. M. & Leiserowitz, A. A. What is sustainable development? Goals, indicator, values, and practice. *Environ. Sci. Pol. Sustain. Dev.* **47,** 8–21 (2005).
89. Ariely, D. *Predictably Irrational, Revised and Expanded Edition* (HarperCollins, 2009).
90. Ehrlich, P. R. & Ornstein, R. E. *Humanity on a Tightrope: Thoughts on Empathy, Family, and Big Changes for a Viable Future* (Rowman & Littlefield, 2010).
91. Cialdini, R. B. *Influence: Science and Practice* (Allyn & Bacon, 2008).
92. Fishkin, J. S. The televised deliberative poll: an experiment in democracy. *Ann. Am. Acad. Pol. Soc. Sci.* **546,** 132–140 (1996).
93. Harff, B. No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. *Am. Polit. Sci. Rev.* **97,** 57–73 (2003).
94. Diamond, J. & Robinson, J. A. *Natural Experiments of History* (Harvard Univ. Press, 2010).
95. Aronson, E. & Patnoe, S. *Cooperation in the Classroom: The Jigsaw Method* 3rd edn (Pinter & Martin, 2010).
96. Perrow, C. *The Next Catastrophe: Reducing Our Vulnerabilities to Natural, Industrial, and Terrorist Disasters* (Princeton Univ. Press, 2007).
97. Lakoff, G. *Don't Think of an Elephant! Know Your Values and Frame the Debate* (Chelsea Green, 2004).
98. Westley, F. *et al.* Tipping towards sustainability: emerging pathways of transformation. *Ambio* **40,** 762–780 (2011).
99. Wheeler, D. *Quantifying Vulnerability to Climate Change: Implications for Adaptation Assistance* (Center for Global Development, 2011).
100. Holdren, J. P. Population and the energy problem. *Popul. Environ.* **12,** 231–255 (1991).

Acknowledgements We are indebted to the following colleagues who were kind enough to give useful criticism: K. Arrow, P. Dasgupta, L. Goulder, R. Horn, D. Karp, S. Levin, C. Mendenhall, H. Mooney, P. Raven, E. Rosa, M. Ruckelshaus and H. Tallis. We are grateful for support from the LuEsther T Mertz Charitable Trust, the Moore Foundation, the Winslow Foundation, and Peter and Helen Bing.

Author Contributions P.E., P.K. and G.C.D. wrote the article jointly.

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 P.E. (pre@stanford.edu).

doi:10.1038/nature11094

Cis-regulatory control of corticospinal system development and evolution

Sungbo Shim[1], Kenneth Y. Kwan[1], Mingfeng Li[1], Veronique Lefebvre[2] & Nenad Šestan[1]

The co-emergence of a six-layered cerebral neocortex and its corticospinal output system is one of the evolutionary hallmarks of mammals. However, the genetic programs that underlie their development and evolution remain poorly understood. Here we identify a conserved non-exonic element (E4) that acts as a cortex-specific enhancer for the nearby gene *Fezf2* (also known as *Fezl* and *Zfp312*), which is required for the specification of corticospinal neuron identity and connectivity. We find that SOX4 and SOX11 functionally compete with the repressor SOX5 in the transactivation of E4. Cortex-specific double deletion of *Sox4* and *Sox11* leads to the loss of *Fezf2* expression, failed specification of corticospinal neurons and, independent of *Fezf2*, a *reeler*-like inversion of layers. We show evidence supporting the emergence of functional SOX-binding sites in E4 during tetrapod evolution, and their subsequent stabilization in mammals and possibly amniotes. These findings reveal that SOX transcription factors converge onto a *cis*-acting element of *Fezf2* and form critical components of a regulatory network controlling the identity and connectivity of corticospinal neurons.

The emergence and expansion of the neocortex in mammals has been crucial to the evolution of complex perceptual, cognitive, emotional and motor abilities[1–3]. The neocortex is organized into six layers based largely on the distinct subtypes of excitatory projection (or pyramidal) neurons and their patterns of connectivity[4–9]. Upper-layer (L2, L3 and L4) projection neurons form synaptic connections solely with other cortical neurons. By contrast, the majority of neurons in the deeper layers (L5 and L6) project to subcortical regions. Studies of laminar inversion in *reeler* mice lacking the reelin (RELN) protein[10–14] have shown that neuron identity and connectivity are determined by birth order rather than by laminar position, suggesting that neuronal specification and positioning are largely separately encoded.

The layer-specific pattern of connectivity is dependent on cortical areas. The long-range projections of L5 neurons in somatosensory-motor areas form the corticospinal (CS) system that directly connects the neocortex with various subcortical regions[15–22]. A major component of the system, the CS (or pyramidal) tract, descends through the brainstem and into the spinal cord to provide a high degree of direct control over the precise motor functions affected in many clinical conditions[21,22]. Despite these important functional implications, the genetic programs controlling CS system development and evolution remain unclear.

Phenotypic specification and evolution of neural circuits depend on precise regulation of the timing, location and level of gene expression[23,24]. Transcriptional control via *cis*-regulatory elements has emerged as a crucial mechanism[25–29]. The *cis*-regulatory mechanisms underlying the specification of distinct neuronal cell types and circuits, however, remain poorly understood. Specification of CS neurons and the formation of the CS tract critically depend on *Fezf2*, which encodes a zinc-finger transcription factor highly enriched in early cortical progenitor cells and their deep-layer neuron progenies[30–35]. Inactivation of *Fezf2* disrupts the molecular specification of deep-layer neurons and the formation of corticofugal projections, including the CS tract, without affecting the inside-out pattern of neurogenesis and lamination[33–35]. Misexpression of *Fezf2* in L2 or L3 cortico-cortical projection neurons[35,36] or striatal interneurons[37] alters their molecular profile and induces ectopic subcerebral projections. These findings indicate that the precisely regulated transcription of *Fezf2* is probably critical to the proper specification of distinct types of cortical projection neurons.

In this study, we used bacterial artificial chromosome (BAC) engineering and genetic inactivation in mice to identify and characterize a cortex-specific *Fezf2* enhancer and its *trans*-regulators. We show that three SOX transcription factors converge onto the *Fezf2* enhancer to control CS system development via functional binding sites that emerged in tetrapods. We also found that *Sox4* and *Sox11* are required for *Fezf2*-independent regulation of cortical RELN expression and laminar organization. Thus, these findings reveal novel developmental genetic programs that control layer formation and CS neuron identity, and the regulatory mechanisms by which they may have evolved.

E4 enhancer controls cortical *Fezf2* expression

Previously, it has been shown that the BAC transgenic mouse harbouring 200 kb of the mouse *Fezf2* locus and the *Gfp* reporter gene (*Fezf2-Gfp*; ref. 38) recapitulates the spatio-temporal expression of endogenous *Fezf2* (refs 39, 40). This indicated that the *cis*-regulatory elements required for cortical *Fezf2* expression are located within the BAC sequence. On the basis of the remarkable similarity in cortical expression pattern of *Fezf2* between mouse and human[39,41], we proposed that the regulatory elements are also highly evolutionarily conserved. Comparative sequence analysis revealed several conserved non-exonic elements (CNEEs) within the *Fezf2-Gfp* BAC (Fig. 1a). To test whether the selected CNEEs regulate *Fezf2-Gfp* expression, we generated multiple lines of BAC transgenic mice in which one of the CNEEs within the *Fezf2-Gfp* BAC was deleted (ΔE1 to ΔE4; Fig. 1b and Methods). We analysed GFP expression in whole-mount brain preparations and tissue sections of transgenic mice at embryonic and postnatal ages (Fig. 1c–g and Supplementary Figs 1 and 2). We found no pronounced change in ΔE1, ΔE2 and ΔE3 mutants, which, like

[1]Department of Neurobiology and Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, Connecticut 06510, USA. [2]Department of Cell Biology and Orthopaedic and Rheumatologic Research Center, Cleveland Clinic Lerner Research Institute, Cleveland, Ohio 44195, USA.

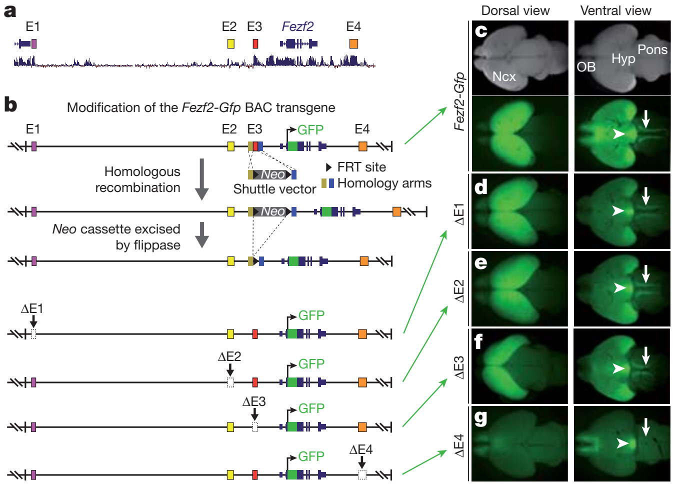


Figure 1 | Identification of a cortex-specific *Fezf2* enhancer. **a**, The locations of CNEEs (E1 to E4) analysed in this study are indicated. **b**, Deletion of each CNEE from the *Fezf2-Gfp* BAC and transgenesis. A positive-selection neomycin cassette (*Neo*) flanked by homology arms was inserted by homologous recombination, resulting in the deletion of each CNEE. After the removal of *Neo* by flippase, the modified BACs (ΔE1 to ΔE4) were used for transgenesis. **c**–**g**, Whole-mount brains of P0 control *Fezf2-Gfp* (**c**) and founder-mutant (**d**–**g**) transgenic mice. $n \geq 3$ founders per mutant line. In *Fezf2-Gfp* mice (**c**), GFP was expressed in the neocortex (Ncx), olfactory bulb (OB), hypothalamus (Hyp) (arrowhead) and CS axons in the pons (arrow). The deletion of E4 (**g**), but not E1 to E3 (**d**–**f**), led to a specific loss of GFP expression in the neocortex and CS axons, but not in the olfactory bulb or hypothalamus.

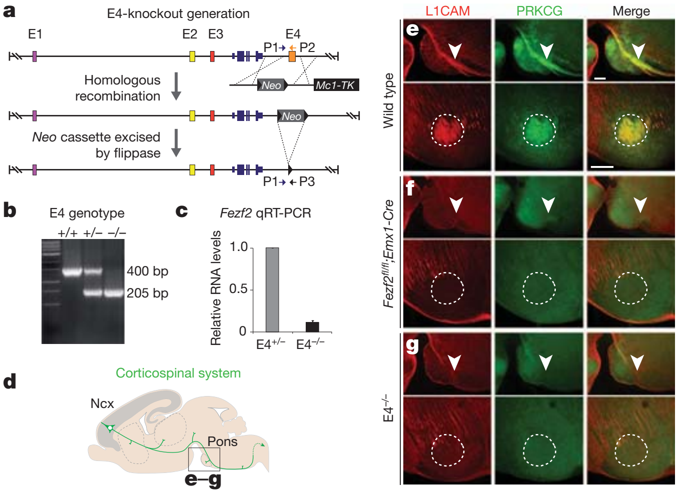


Figure 2 | Loss of neocortical *Fezf2* expression and CS axons in E4-knockout mice. **a**, Generation of E4-knockout mice ($E4^{-/-}$) using the *pGK-Neo/Mc1-TK* selection cassette. **b**, Duplex PCR genotyping of wild-type and mutant alleles using primers P1, P2 and P3. **c**, Analysis of neocortical *Fezf2* expression by quantitative RT–PCR (qRT–PCR). Normalized to *Gapdh*, *Fezf2* messenger RNA levels were significantly reduced in the $E4^{-/-}$ mice compared with heterozygous littermate controls ($E4^{+/-}$). $P = 2.1 \times 10^{-6}$; one-tailed Student's *t*-test; $n = 3$ per genotype. Error bars; s.e.m. **d**, Schematic depiction of the CS system. **e**–**g**, Sagittal (top row) and coronal (bottom row) sections of the pons from wild-type (**e**), cortex-specific *Fezf2*-knockout (**f**) and $E4^{-/-}$ (**g**) mice immunostained for axon marker L1CAM (red) and PRKCG (green). The near-complete loss of CS axons (arrowheads and dashed outlines) in the $E4^{-/-}$ mice is a phenocopy of cortex-specific *Fezf2* deletion. Scale bars, 200 μm.

wild-type *Fezf2-Gfp* mice, expressed GFP in neocortex and pontine CS axons (Fig. 1d–f). By contrast, the deletion of CNEE E4 (ΔE4), which is located 7.3 kb downstream of the *Fezf2* transcription start site, resulted in a drastic loss of GFP expression in the cortex but not in the hypothalamus or the olfactory bulb (Fig. 1g). Moreover, in ΔE4 mice, GFP-positive CS axon fascicles were lost from the ventral surface of the pons (Fig. 1g). Consistent with this finding, E4 was identified as an *in vivo* binding site of the enhancer-associated protein EP300 in mouse embryonic forebrain[27]. Together, these results demonstrate that E4 is a *cis*-regulatory module acting as a cortex-specific enhancer of *Fezf2*.

Loss of *Fezf2* expression and CS axons in $E4^{-/-}$ mice

To test the function of the E4 enhancer directly, we deleted the E4 region, while leaving intact the entire *Fezf2* coding and proximal promoter regions, using homologous recombination in embryonic stem cells (Fig. 2a, b). Mice lacking E4 ($E4^{-/-}$) were viable, fertile and without overt behavioural or motor phenotypes. Quantitative reverse-transcription PCR (RT–PCR) analysis revealed that *Fezf2* expression is drastically downregulated in neocortex of these mice (Fig. 2c). Immunostaining of tissue sections of E4-null mice for PRKCG and L1CAM, two proteins expressed by CS neurons and their axons[11] (Fig. 2d, e), revealed that CS axons were absent from the pons (Fig. 2g), similar to mutant mice with a cortex-specific deletion of *Fezf2* (*Fezf2*$^{fl/fl}$;*Emx1-Cre*) (Fig. 2f). Furthermore, the expression of *Bcl11b* (*Ctip2*), a gene functioning downstream of *Fezf2* (refs 33, 36), was also downregulated in the E4-null neocortex (Supplementary Fig. 3). Taken together, these results indicate that the E4 enhancer is required for neocortical *Fezf2* expression, molecular specification of L5/6 neurons and the formation of CS tract.

SOX4 and SOX11 bind and activate the E4 enhancer

We have previously shown that SOX5, a SOXD member of the large family of SOX transcription factors, binds to and represses the transcriptional activity of the E4 enhancer[39]. Unexpectedly, SOX5 itself is also required for CS tract formation independent of its repression of *Fezf2* (ref. 39). Because different SOX transcription factors are known to compete for a common motif to mediate both activation and repression of regulatory elements[42–44], we proposed that other SOX members may activate the E4 enhancer, perhaps acting competitively against SOX5-mediated repression. Consistent with this hypothesis, our sequence analysis using MATINSPECTOR (Genomatix) revealed eight putative SOX-binding sites in E4. To prioritize which SOX transcription factors may be good candidates for potential *trans*-regulators of the E4 enhancer, we searched for those that are most highly correlated in their spatio-temporal expression pattern with *FEZF2*, using the Human Brain Transcriptome database[45] (http://www.humanbrain-transcriptome.org). The highest-correlated *SOX* genes were three members of the *SOXC* group (*SOX4*, *SOX11* and *SOX12*) and *SOX5* (Supplementary Fig. 4). The transcription factors encoded by these *SOX* genes are crucial in regulating cell fate and differentiation[42–44,46], and a *de novo* deletion of *SOX11* was described in a patient with autism and intellectual disability[47]. Moreover, *Sox4* and *Sox11* act as transcriptional activators[45] and their expression patterns overlap with that of *Fezf2* in developing cortex (Supplementary Fig. 5).

To determine whether SOX4 or SOX11 binds E4, we performed chromatin immunoprecipitation (ChIP)-PCR assays in Neuro-2a cells transiently expressing V5-tagged SOX4 or SOX11 (Methods). Anti-V5 antibodies precipitated E4 DNA, but not E1, E2 or E3 DNA, confirming binding of SOX4 and SOX11 to the E4 enhancer (Fig. 3a). Moreover, recruitment of RNA polymerase II to E4 occurred in the presence of SOX4 or SOX11, which is consistent with increased transcriptional activity. To test the functional consequence of *SoxC*-gene expression on E4, we expressed a luciferase reporter under the control of E4 (pGL4-E4) in Neuro-2a cells (Fig. 3b). Luciferase activity driven by the E4 enhancer was significantly increased by co-transfected *Sox4* or *Sox11*, but not *Sox12*.

To dissect which sequences within E4 drive *Fezf2* expression, we generated four truncated versions of the E4 sequence (E4F1to E4F4; Fig. 3c). The luciferase activity of the E4F2 fragment was increased by SOX4 and SOX11, but not by SOX12 (Fig. 3d). Using

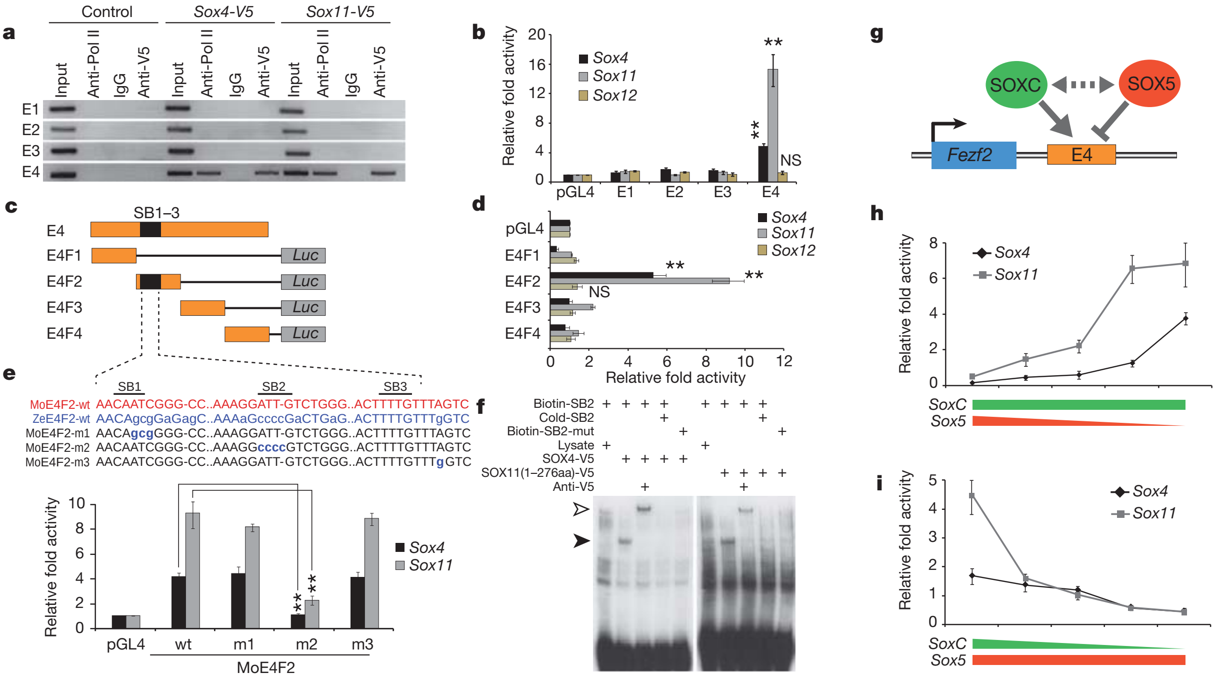


Figure 3 | SOX4 and SOX11 bind to and activate E4 via competition with SOX5. a, Chromatin immunoprecipitation (ChIP) from Neuro-2a cells expressing V5-tagged SOX4 and SOX11. Captured DNA was analysed by PCR using primers specific for E1–E4. SOX4 and SOX11 bound and recruited RNA polymerase II (Pol II) to E4 but not E1–E3. **b**, Analysis of SOXC transactivation using empty luciferase vectors (pGL4) or luciferase vectors containing E1–E3. The activity of E4, but not E1–E3, was significantly increased by *Sox4* (≥4-fold) and *Sox11* (≥13-fold), but not by *Sox12* (≤1.5-fold). **c**, **d**, Analysis of functional elements within E4 using luciferase vectors containing deletion fragments of E4 (E4F1 to E4F4) (**c**). E4F2, but not the other fragments, was significantly activated by co-transfection of *Sox4* (≥5-fold) and *Sox11* (≥9-fold), but not of *Sox12* (≤1.5 fold) (**d**). NS, not significant. **e**, Analysis of putative SOX-binding sites (SB1 to SB3) using mouse (Mo) E4F2 luciferase vectors mutagenized by replacement (blue lowercase nucleotides) with zebrafish (Ze) sequence. Targeted mutation of SB2 (MoE4F2-m2) significantly diminished the transactivating ability of SOX4 and SOX11. **f**, Electrophoretic mobility shift assay with biotin-labelled SB2 DNA. SOX4-V5 and SOX11(1–276aa)-V5 shifted wild-type SB2 DNA (arrowhead) but not mutated SB2 DNA. SOXC–SB2 complexes were supershifted (open arrowhead) by an anti-V5 antibody. **g**, Schematic model of E4 regulation by SOX5 and SOXC proteins. **h**, **i**, Analysis of competition between SOX5 and SOXC proteins using the E4-containing luciferase vector. Decreasing concentrations of co-transfected *Sox5* led to a dose-dependent increase in E4 activation in response to *Sox4* and *Sox11* (**h**), whereas decreasing concentrations of co-transfected *Sox4* and *Sox11* led to a dose-dependent increase in *Sox5* repression of E4 (**i**). One-tailed Student's *t*-test; $^{**}P < 0.01$; $n = 4$ per condition. Error bars; s.e.m.

MATINSPECTOR, six potential SOX-binding sites within the E4F2 sequence were predicted. To define the precise base pairs with which SOXC proteins interact, we first tested whether the zebrafish E4F2 sequence, which has 74.2% sequence identity with mouse E4F2, is activated by SOXC proteins. Remarkably, the zebrafish E4F2 sequence was not activated by SOX11 (Supplementary Fig. 6), indicating that the SOXC-interacting sequences of mouse E4F2 lie within the sequences that are divergent in zebrafish. Of the predicted SOX-binding sites in mouse E4F2, three putative SOX-binding sites (SB1 to SB3) are absent from zebrafish E4F2. To test the function and specificity of these sites, we generated mutant versions of E4F2 by substituting SB1, SB2 or SB3 with the zebrafish sequence (Fig. 3e). Only mutations of the mouse SB2 site significantly attenuated the ability of SOX4 and SOX11 to activate the luciferase reporter gene, indicating that this site is crucial to species differences in SOXC-mediated transactivation of E4. Next we assessed the ability of SOX4 and SOX11 to bind to the SB2 site *in vitro*, using an electrophoretic mobility shift assay. We synthesized V5-tagged SOX4 and a higher-affinity, but equally specific, truncated form of SOX11 (SOX11(1–276aa)), because the binding affinity of native SOX11 is weak[43]. Biotinylated SB2 DNA was shifted in the presence of SOX4-V5 or SOX11(1–276aa)-V5 and supershifted to a higher molecular weight by an anti-V5 antibody (Fig. 3f), but not when excess unlabelled or mutated SB2 DNA was used. Therefore, SOX4 and SOX11 directly bind to and activate the transcription activity of E4 via SB2.

To test whether SOX4 and SOX11 functionally compete with the repressor SOX5 in the transactivation of E4 (Fig. 3g), we used the pGL4-E4 luciferase assay. Luciferase activity was significantly increased with increasing amounts of co-transfected *Sox11* and, to a lesser extent, *Sox4*, whereas increased amounts of *Sox5* significantly decreased luciferase activity (Fig. 3h, i). Taken together, our sequence analysis and assays, both *in vitro* and *in vivo*, demonstrate that SOX4 and SOX11 functionally compete with SOX5 repression to activate *Fezf2* transcription via direct binding to sites within E4.

Loss of *Fezf2* expression and CS axons in cdKO mice

To test the phenotypic consequences of *Sox4* and *Sox11* inactivation, we generated cortex-specific deletions of *Sox4* and *Sox11* using the *Emx1-Cre* line (Methods) to circumvent the prenatal lethality associated with constitutive deletions of the two genes[44]. Single conditional knockout mice (cKO: *Sox4*$^{fl/fl}$;*Sox11*$^{fl/+}$;*Emx1-Cre* and *Sox4*$^{fl/+}$;*Sox11*$^{fl/fl}$;*Emx1-Cre*) were both viable and fertile, whereas the conditional double knockout mice (cdKO: *Sox4*$^{fl/fl}$;*Sox11* $^{fl/fl}$;*Emx1-Cre*) died within the first postnatal week (data not shown). Inspection of the cerebrum during the first postnatal week revealed no obvious gross defects in single *Sox4* or *Sox11* cKO animals (Supplementary Fig. 7). However, in the absence of both *Sox4* and *Sox11*, we observed a reduction in the size of the cerebral hemispheres and the olfactory bulb.

To assess the requirement of *Sox4* and *Sox11* in E4-mediated activity, we transfected primary cortical neurons cultured from heterozygous littermate control or *Sox4;Sox11* cdKO embryos with control or E4-containing luciferase constructs (Fig. 4a). In control neurons, the presence of E4 increased luciferase activity by a factor of 2.5 ± 0.3 ($P = 1.9 \times 10^{-4}$; one-tailed Student's *t*-test). This increase was abolished in neurons from *Sox4;Sox11* cdKO (factor of 1.1 ± 0.1; $P = 0.866$), indicating that the two SOXC transcription factors are major activators of the enhancer in cortical neurons.

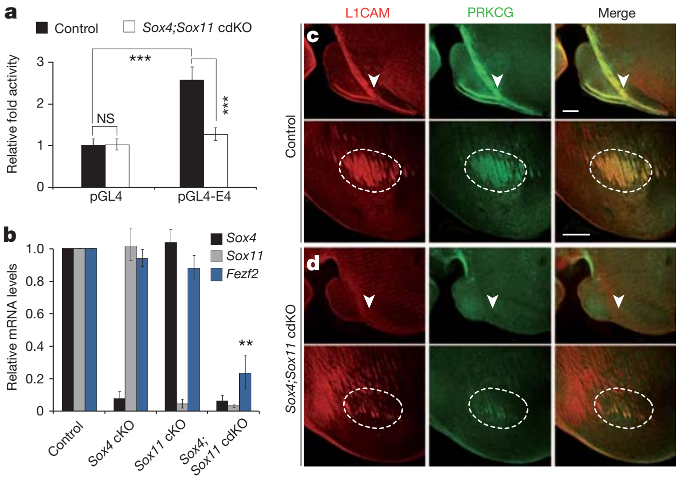


Figure 4 | *Sox4* and *Sox11* are required for *Fezf2* expression and CS tract formation. a, The requirement of SOX4 and SOX11 for E4 transactivation was determined in neurons cultured from E14.5 heterozygous littermate control mice and *Sox4*;*Sox11* cdKO mice, and transfected with the E4 luciferase construct. The activity of E4 in double-knockout neurons was not significantly above background (1.1-fold). **b**, Analysis of neocortical *Fezf2* expression by quantitative RT–PCR in cortex-specific *Sox4* and/or *Sox11* mutants. Normalized to *Gapdh*, *Fezf2* messenger RNA (mRNA) levels were drastically reduced in cdKO mice but not in *Sox4* or *Sox11* mutants. **c**, **d**, Sagittal (top row) and coronal (bottom row) sections of the control and cdKO pons immunostained for L1CAM (red) and PRKCG (green). The drastic loss of L1CAM- and PRKCG-positive CS tract axons in the cdKO mice (arrowhead and dashed outline) is similar to that in the cortex-specific *Fezf2* mutant. One-tailed Student's *t*-test; $**P < 0.01$, $***P < 0.001$; $n = 4$ per genotype. Errors bars; s.e.m.; scale bars, 200 μm.

Next, using quantitative RT–PCR, we found a significant reduction in neocortical *Fezf2* expression in cdKO littermates compared with control or single cKO littermates at postnatal day 0 (P0) (Fig. 4b). In addition, immunostaining for L1CAM and PRKCG revealed a drastic loss of CS axons in *Sox4*;*Sox11* cdKO mice but not in single cKO littermates at P0 and P6 (Fig. 4c, d and Supplementary Fig. 8), whereas the organization of other brainstem tracts was not affected. Because the absence of L1CAM and PRKCG labelling could reflect their down-regulated expression, as opposed to loss of CS axons, in the cdKO mice, we confirmed the absence of the CS tract using the CRE-responsive *CAG-Cat-Gfp* transgenic line to express GFP in all cortical projection neurons and their axons[39,40] (Supplementary Fig. 9).

Two additional observations indicate that the loss of *Fezf2* and CS axons in these mice was not due to an absence of L5 neurons. First, only a moderate increase in cell death was detected in the somatosensory-motor areas, in which CS axons originate (Supplementary Fig. 10). Second, many BCL11B-immunopositive L5 neurons were present in the somatosensory-motor areas of cdKO mice, although lightly immunostained (Supplementary Fig. 11), indicating that neurons that would normally give rise to the CS tract were present. Analysis of additional laminar markers further revealed an inversion of cortical layers similar to that of the *reeler* mutant mouse[10–12] (Supplementary Fig. 11), indicating that *Sox4* and *Sox11* are also required for proper laminar positioning of neurons. Previous studies have shown that the CS tract is present in *reeler* mice[12], indicating that this laminar phenotype occurs independently of *Fezf2* and probably in response to defects in the RELN signalling pathway[13,14]. In support of this possibility, RELN was absent from cdKO mouse neocortex (Supplementary Fig. 12). Thus, the combined deletion of *Sox4* and *Sox11* leads to defects in the laminar position of neurons and the molecular specification and connectivity of CS neurons.

Functional and evolutionary implications of E4 sequence variations

The differences in SOXC-mediated transcription between mouse E4F2 and zebrafish E4F2 suggest that species differences in the E4F2 sequence have functional implications. Analysis of SB1, SB2 and SB3 sites within the E4F2 sequence in 23 species revealed that SB2 is conserved in all analysed mammals and the two available non-mammalian amniotes (chicken and lizard) (Supplementary Fig. 13). To investigate the functional consequences of species differences, we used luciferase assays in Neuro-2a cells to analyse the activity of E4F2 sequence from different vertebrates (Fig. 5a, b) in response to SOX11, a more powerful transactivator than SOX4 (Fig. 3). Of the constructs containing an E4F2 sequence of a mammal (human, chimpanzee, macaque, or mouse) or non-mammalian amniote (chicken), which have conserved SB1-3 sequences, the luciferase reporter activity was strongly increased following co-transfection with *Sox11*. The reporter activity of a non-amniote tetrapod (*Xenopus*) E4F2 construct was moderately increased compared with the empty plasmid, but was not as high as that of the mammalian constructs ($P = 3.9 \times 10^{-16}$; one-tailed Student's *t*-test; Fig. 5b). By contrast, the reporter construct containing the zebrafish E4F2 sequence, the SB2 sequence of which is highly divergent from those of mammals, exhibited only basal level activity similar to the empty control luciferase plasmid. To confirm that this is dependent on sequence variations between species, we mutated zebrafish E4F2 by 'murinizing' its SB1, SB2 or SB3 sequence (ZeE4F2-m1, -m2 or -m3, respectively). Reporter activity of the murinized zebrafish SB2 (ZeE4F2-m2), but not SB1 or SB3, was robustly induced by SOX11 (Fig. 5c). ZeE4F2-m2 murinization, however, was insufficient to restore the level of expression observed in the wild-type mouse sequence, suggesting the potential contribution of additional sequences. Thus, evolutionary differences in the E4F2 sequence, especially within SB2, are directly related to functional differences in the ability of SOX11 to activate this regulatory element.

Next, to investigate the functional consequences of species differences in the FEZF2 coding sequence, we tested the ability of zebrafish *fezf2* to rescue the effects of mouse *Fezf2* deficiency. We electroporated *in utero* the neocortical wall of the floxed *Fezf2* ($Fezf2^{fl/fl}$) mice[40] with *Cre* and CRE-responsive *Gfp*-expressing constructs with or without *fezf2* at embryonic day 12.5 (E12.5) (Fig. 5d). Analysis of electroporated mice at P0 revealed that *fezf2* was sufficient to rescue the formation of the CS tract in *Fezf2*-null mouse neurons (Fig. 5e). Taken together, these results suggest that sequence variations in the E4 enhancer, but not the *Fezf2* coding region, gave *Fezf2* an essential role in CS tract evolution.

Discussion

Our data provide critical insight into the genes and regulatory components controlling CS system development, centring on the *cis*-regulation of *Fezf2* by SOX4 and SOX11. We show that genetic inactivation of the E4 *cis*-regulatory module results in compartmentalized phenotypic effects largely limited to the loss of cortical *Fezf2* expression and CS system. The spatio-temporal dynamics of cortical *Fezf2* is further controlled by SOX5 and TBR1, two transcription factors expressed post-mitotically in projection neurons[39,40,46,48,49]. On its initial activation in early cortical progenitors, *Fezf2* is highly expressed in deep-layer neurons during early corticogenesis but subsequently repressed in L6 neurons by SOX5 and TBR1 to create a postnatal L5-enriched pattern. The earlier expression of SOX4 and SOX11 is consistent with their role in *Fezf2* activation before the L6 upregulation of SOX5, which is a functional competitor. Despite what is known about the function and regulation of *Fezf2*, a number of key questions have yet to be addressed. First, additional, as yet untested, *trans*-regulators and regulatory elements probably contribute to *Fezf2* regulation in a context-dependent manner. These include the mediators of *Fezf2* repression in L2, L3 and L4 projection neurons. Second, the distinct, and seemingly incongruous, effects of SOX5 in repressing *Fezf2* but being independently required for CS tract formation[39] have not been resolved. Third, the direct transcriptional targets of FEZF2 in CS neurons remain largely unknown.

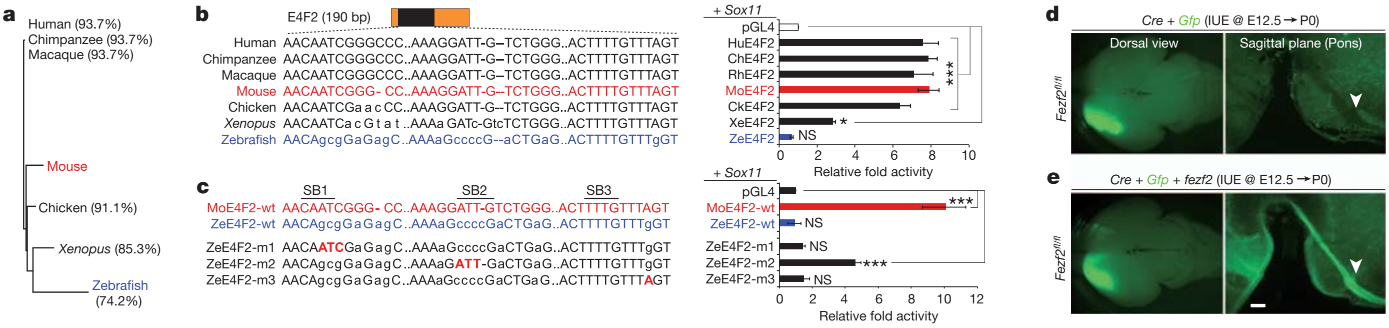


Figure 5 | Functional analysis of species differences in E4 sequence. **a**, Hierarchical clustering of E4F2 sequences from seven vertebrates. Percentage nucleotide identity relative to mouse E4F2 is indicated in parentheses. **b**, Analysis of species differences in *Sox11* activation of E4F2. *Sox11* transactivated the E4F2 sequence of four mammals and chick (≥6.3-fold) and *Xenopus* (2.8-fold). By contrast, the activity of zebrafish E4F2, which is divergent in SB1 and SB2, was not activated by *Sox11*. **c**, Murinization (red uppercase nucleotides) of zebrafish SB2 (ZeE4F2-m2) partly rescued the loss of transactivation of wild-type zebrafish E4F2 (ZeE4F2-wt) by *Sox11*. **d**, **e**, Cell-autonomous rescue of mouse *Fezf2* loss of function by zebrafish *fezf2*. *In utero* electroporation (IUE) of *Fezf2*$^{fl/fl}$ neocortical wall at E12.5 with *Cre* and CRE-responsive *Gfp* plasmids. *Fezf2*-deficient L5 neurons do not form CS tract at P0 (**d**). Co-electroporation of an *fezf2* plasmid cell-autonomously rescued the formation of CS tract by *Fezf2*-deficient neurons (**e**). One-tailed Student's *t*-test; $*P < 0.05$, $***P < 0.001$; $n = 4$ per condition. Errors bars, s.e.m.; scale bar, 200 μm.

The laminar position and radial organization of projection neurons are key features of cortical development, which rely on RELN signalling[10–14]. We also show that SOX4 and SOX11 are crucial in regulating RELN expression and the inside-out pattern of cortical layer formation, independent of E4 or *Fezf2* and probably involving interactions with distinct regulatory elements. Moreover, SOX4 and SOX11 have additional roles, as in mice lacking both genes, the cortex and olfactory bulb are smaller and cell death is increased. Thus, SOX4 and SOX11 have pleiotropic functions, which are probably mediated by distinct regulatory elements and downstream target genes that are involved in multiple developmental processes.

Our results indicate that, following their emergence in tetrapods, functional SOX-binding sites have retained high conservation through purifying selection in mammals and some amniotes, thus directly linking species variations in regulatory sequences to functional outcomes. E4 sequence substitutions in SB2 may constitute an evolutionary turning point for *Fezf2* function during forebrain development, possibly facilitating the formation of descending telencephalic pathways including the CS system. Whereas minor projections from the ventral (subpallial) telencephalon to the spinal cord are present in amphibians[2], direct dorsal telencephalo-spinal projections resembling the CS tract have been reported only in mammals and some birds[2,50]. We propose that the concurrent emergence of the described regulatory mechanisms and direct telencephalo-spinal projections in early amniotes, together with subsequent changes in genetic programs driving the patterning and expansion of a six-layered dorsal pallium, made possible the evolution of the CS system in mammals.

METHODS SUMMARY

Selected CNEE sequences were deleted from a *Fezf2-Gfp* BAC by recombination-mediated genetic engineering, and GFP expression from modified BACs was analysed by transgenesis. To confirm the requirement of E4 enhancer for *Fezf2* expression, a germline E4 deletion mutant was generated. Putative E4 transactivators SOX4 and SOX11, identified by E4 sequence and co-expression analyses, were tested using ChIP and luciferase reporter assays. To analyse the requirement of *Sox4* and *Sox11* for *Fezf2* expression and cortical development, we generated cortex-specific single and double *Sox4* and *Sox11* null mice. Complete materials and methods are described in Supplementary Information.

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

Received 7 November 2011; accepted 28 March 2012.

Published online 30 May 2012.

1. Northcutt, R. G. & Kaas, J. H. The emergence and evolution of mammalian neocortex. *Trends Neurosci.* **18,** 373–379 (1995).
2. Nieuwenhuys, R., ten Donkelaar, H. J. & Nicholson, C. *The Central Nervous System of Vertebrates* (Springer, 1998).
3. Butler, A. B., Reiner, A. & Karten, H. J. Evolution of the amniote pallium and the origins of mammalian neocortex. *Ann. NY Acad. Sci.* **1225,** 14–27 (2011).
4. O'Leary, D. D. M. & Koester, S. E. Development of projection neuron types, axon pathways, and patterned connections of the mammalian cortex. *Neuron* **10,** 991–1006 (1993).
5. Rash, B. G. & Grove, E. A. Area and layer patterning in the developing cerebral cortex. *Curr. Opin. Neurobiol.* **16,** 25–34 (2006).
6. Molyneaux, B. J., Arlotta, P., Menezes, J. R. L. & Macklis, J. D. Neuronal subtype specification in the cerebral cortex. *Nature Rev. Neurosci.* **8,** 427–437 (2007).
7. Leone, D. P., Srinivasan, K., Chen, B., Alcamo, E. & McConnell, S. K. The determination of projection neuron identity in the developing cerebral cortex. *Curr. Opin. Neurobiol.* **18,** 28–35 (2008).
8. Hansen, D. V., Rubenstein, J. L. & Kriegstein, A. R. Deriving excitatory neurons of the neocortex from pluripotent stem cells. *Neuron* **70,** 645–660 (2011).
9. Kwan, K. Y., Sestan, N. & Anton, E. S. Transcriptional co-regulation of neuronal migration and laminar identity in the neocortex. *Development* **139,** 1535–1546 (2012).
10. Caviness, V. S. & Sidman, R. L. Time of origin of corresponding cell classes in cerebral-cortex of normal and *reeler* mutant mice: autoradiographic analysis. *J. Comp. Neurol.* **148,** 141–151 (1973).
11. Steindler, D. A. & Colwell, S. A. *Reeler* mutant mouse: maintenance of appropriate and reciprocal connections in the cerebral cortex and thalamus. *Brain Res.* **113,** 386–393 (1976).
12. Terashima, T. Anatomy, development and lesion-induced plasticity of rodent corticospinal tract. *Neurosci. Res.* **22,** 139–161 (1995).
13. Bar, I., de Rouvroit, C. L. & Goffinet, A. M. The evolution of cortical development. An hypothesis based on the role of the Reelin signaling pathway. *Trends Neurosci.* **23,** 633–638 (2000).
14. Rice, D. S. & Curran, T. Role of the Reelin signaling pathway in central nervous system development. *Annu. Rev. Neurosci.* **24,** 1005–1039 (2001).
15. Joosten, E. A. J. & Bar, D. P. R. Axon guidance of outgrowing corticospinal fibres in the rat. *J. Anat.* **194,** 15–32 (1999).
16. Martin, J. H. The corticospinal system: from development to motor control. *Neuroscientist* **11,** 161–173 (2005).
17. Canty, A. J. & Murphy, M. Molecular mechanisms of axon guidance in the developing corticospinal tract. *Prog. Neurobiol.* **85,** 214–235 (2008).
18. Lemon, R. N. Descending pathways in motor control. *Annu. Rev. Neurosci.* **31,** 195–218 (2008).
19. Rathelot, J.-A. & Strick, P. L. Subdivisions of primary motor cortex based on cortico-motoneuronal cells. *Proc. Natl Acad. Sci. USA* **106,** 918–923 (2009).
20. Nudo, R. J. & Masterton, R. B. Descending pathways to the spinal-cord. IV. Some factors related to the amount of cortex devoted to the corticospinal tract. *J. Comp. Neurol.* **296,** 584–597 (1990).
21. ten Donkelaar, H. J. *et al.* Development and malformations of the human pyramidal tract. *J. Neurol.* **251,** 1429–1442 (2004).
22. Eyre, J. A. Corticospinal tract development and its plasticity after perinatal injury. *Neurosci. Biobehav. Rev.* **31,** 1136–1149 (2007).
23. Jessell, T. M. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nature Rev. Genet.* **1,** 20–29 (2000).
24. Hobert, O., Carrera, I. & Stefanakis, N. The molecular and gene regulatory signature of a neuron. *Trends Neurosci.* **33,** 435–445 (2010).
25. Wray, G. A. The evolutionary significance of *cis*-regulatory mutations. *Nature Rev. Genet.* **8,** 206–216 (2007).
26. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
27. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
28. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468,** 911–920 (2010).

29. Williamson, I., Hill, R. E. & Bickmore, W. A. Enhancers: from developmental genetics to the genetics of common human disease. *Dev. Cell* **21,** 17–19 (2011).
30. Hashimoto, H. *et al.* Expression of the zinc finger gene fez-like in zebrafish forebrain. *Mech. Dev.* **97,** 191–195 (2000).
31. Matsuo-Takasaki, M., Lim, J. H., Beanan, M. J., Sato, S. M. & Sargent, T. D. Cloning and expression of a novel zinc finger gene, *Fez*, transcribed in the forebrain of *Xenopus* and mouse embryos. *Mech. Dev.* **93,** 201–204 (2000).
32. Inoue, K., Terashima, T., Nishikawa, T. & Takumi, T. *Fez1* is layer-specifically expressed in the adult mouse neocortex. *Eur. J. Neurosci.* **20,** 2909–2916 (2004).
33. Molyneaux, B. J., Arlotta, P., Hirata, T., Hibi, M. & Macklis, J. D. *Fezl* is required for the birth and specification of corticospinal motor neurons. *Neuron* **47,** 817–831 (2005).
34. Chen, B., Schaevitz, L. R. & McConnell, S. K. *Fezl* regulates the differentiation and axon targeting of layer 5 subcortical projection neurons in cerebral cortex. *Proc. Natl Acad. Sci. USA* **102,** 17184–17189 (2005).
35. Chen, J. G., Rasin, M. R., Kwan, K. Y. & Sestan, N. *Zfp312* is required for subcortical axonal projections and dendritic morphology of deep-layer pyramidal neurons of the cerebral cortex. *Proc. Natl Acad. Sci. USA* **102,** 17792–17797 (2005).
36. Chen, B. *et al.* The *Fezf2-Ctip2* genetic pathway regulates the fate choice of subcortical projection neurons in the developing cerebral cortex. *Proc. Natl Acad. Sci. USA* **105,** 11382–11387 (2008).
37. Rouaux, C. & Arlotta, P. *Fezf2* directs the differentiation of corticofugal neurons from striatal progenitors *in vivo. Nature Neurosci.* **13,** 1345–1347 (2010).
38. Gong, S. C. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425,** 917–925 (2003).
39. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl Acad. Sci. USA* **105,** 16021–16026 (2008).
40. Han, W. Q. *et al.* TBR1 directly represses *Fezf2* to control the laminar origin and development of the corticospinal tract. *Proc. Natl Acad. Sci. USA* **108,** 3041–3046 (2011).
41. Fertuzinhos, S. *et al.* Selective depletion of molecularly defined cortical interneurons in human holoprosencephaly with severe striatal hypoplasia. *Cereb. Cortex* **19,** 2196–2207 (2009).
42. Bergsland, M., Werme, M., Malewicz, M., Perlmann, T. & Muhr, J. The establishment of neuronal properties is controlled by Sox4 and Sox11. *Genes Dev.* **20,** 3475–3486 (2006).
43. Dy, P. *et al.* The three SoxC proteins-Sox4, Sox11 and Sox12-exhibit overlapping expression patterns and molecular properties. *Nucleic Acids Res.* **36,** 3101–3117 (2008).
44. Bhattaram, P. *et al.* Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nature Commun.* **1,** 9 (2010).
45. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).
46. Lai, T. *et al.* SOX5 controls the sequential generation of distinct corticofugal neuron subtypes. *Neuron* **57,** 232–247 (2008).
47. Lo-Castro, A. *et al.* Deletion 2p25.2: a cryptic chromosome abnormality in a patient with autism and mental retardation detected using aCGH. *Eur. J. Med. Genet.* **52,** 67–70 (2009).
48. Bedogni, F. *et al.* *Tbr1* regulates regional and laminar identity of postmitotic neurons in developing neocortex. *Proc. Natl Acad. Sci. USA* **107,** 13129–13134 (2010).
49. McKenna, W. L. *et al.* *Tbr1* and *Fezf2* regulate alternate corticofugal neuronal identities during neocortical development. *J. Neurosci.* **31,** 549–564 (2011).
50. Wild, J. M. & Williams, M. N. Rostral wulst in passerine birds. I. Origin, course, and terminations of an avian pyramidal tract. *J. Comp. Neurol.* **416,** 429–450 (2000).

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

Acknowledgements We thank W. Han, Y. Imamura Kawasawa, D. Liu and T. Nottoli for technical help; A. Giraldez and A. M. M. Sousa for reagents; and the Sestan laboratory for discussions. This work was supported by the National Institutes of Health (NS054273, MH081896, AR54153), the March of Dimes Foundation and a McDonnell Scholar Award (N.S.).

Author Contributions S.S., K.Y.K. and N.S. designed the research; S.S. performed the experiments; S.S. and K.Y.K. performed the confocal imaging, M.L. analysed coexpression and deep sequencing data; S.S., K.Y.K. and N.S. analysed the other data; V.L. generated mice with floxed *Sox4* and *Sox11* alleles; N.S. conceived the study; and S.S., K.Y.K. and N.S. wrote the manuscript. All authors discussed and commented on the data.

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

METHODS

Animals. All experiments were performed in accordance with a protocol approved by Yale University's Committee on Animal Research. The generation of mice with the floxed *Sox4* and *Sox11* alleles (*Sox4*$^{fl/fl}$ and *Sox11*$^{fl/fl}$, respectively) was described elsewhere[44,51]. The *CAG-Cat-Gfp* and *Emx1-Cre* P1 artificial chromosome transgenic mice were a gift from Melissa Colbert[52] and Takuji Iwasato[53], respectively. The *Fezf2-Gfp* BAC transgenic mouse was generated by the GENSAT project[38].

Generation of BAC transgenic mouse. To generate BAC reporter constructs harbouring deletions of putative enhancer elements (E1–E4) in the mouse *Fezf2-Gfp* BAC (RP23-141E17-*Gfp*; the GENSAT project[38]), homology arms A and B flanking the sequences targeted for deletion (E1, Chr14: 13,204,475–13,204,944, 470 bp; E2, Chr14: 13,183,314–13,183,991, 678 bp; E3, Chr14: 13,180,797–13,181,283, 487 bp; E4, Chr14: 13,170,100–13,170,960, 861 bp) were amplified by PCR and cloned into the PL451 shuttle vector. SW102 *Escherichia coli* transformed with the *Fezf2-Gfp* BAC was induced for recombination-mediated genetic engineering (recombineering) at 42 °C and electroporated with the linearized shuttle vector. After neomycin selection, *Flpase* expression was induced by arabinose to excise the neomycin cassette[54]. Founder lines were confirmed by PCR using primers corresponding to sequences from E1–E4 (Supplementary Table 1). *Fezf2-Gfp* lines were maintained on a C57BL6/J background and all animals were housed under identical conditions. The intactness of the integrated BACs was confirmed by PCR and deep sequencing of genomic DNA.

Generation of E4-null mice. The targeting vector for the deletion of the E4 enhancer was constructed by a recombineering-based method with an unmodified BAC clone (RP23-141E17). A 6,627-bp fragment containing the E4 enhancer was retrieved from the BAC and inserted into the PL253 vector by recombination in SW102 bacteria. The removal of the E4 sequence was achieved using the strategy described above for the generation of the ΔE4 BAC transgenic mouse. ES cells (C57BL6/J) were electroporated with the resulting targeting construct and expanded under positive (*pGK-Neo*, from the PL451 vector) and negative (*Mc1-TK*, from the PL253 vector) selection. Correctly recombined clones were identified by long-distance PCR, confirmed by sequencing, and used for blastocyst injection and subsequent generation of mice with the targeted allele. The neomycin cassette was subsequently removed by flippase-mediated recombination by breeding with *Flp* mice. The mice were genotyped using the P1/P2 primer set (400 bp) for the wild-type allele and the P1/P3 primer set (205 bp) the E4 deletion allele (Supplementary Table 1).

Immunohistochemistry. P0 and P6 brains were fixed by immersion in 4% paraformaldehyde overnight at 4 °C and sectioned using a vibratome (Leica). Immunostaining was performed as previously described[20]. The following primary antibodies were used: anti-L1CAM (rat, 1:300; Millipore), anti-PRKCG (rabbit, 1:300; Santa Cruz), anti-GFP (chicken, 1:3,000; Abcam), anti-CUX1 (rabbit, 1:150; Santa Cruz), anti-SATB2 (mouse, 1:200; Genway), anti-BCL11B (rat, 1:250; Santa Cruz), anti-ZFPM2 (rabbit, 1:250; Santa Cruz), anti-RELN (mouse, 1:300; Millipore) and anti-CASP3, active (rabbit, 1:200; Millipore). We tested a total of 12 commercially available anti-SOX4 antibodies (Abcam, ab52043, ab86809, ab90696; Abgent, AP2045a, AP2045c; LS Bioscience, LS-B3520; Millipore, AB5803, AB10537; Pierce, PA1-38638, PA1-38639; Santa Cruz Biotechnology, sc-17326; Sigma, HPA029901) and 10 commercially available anti-SOX11 antibodies (Abcam, ab42853; Aviva Systems Biology, ARP33328, ARP38235; LS Bioscience, LS-B1567, LS-C10306; Millipore, AB9090; Pierce, PA5-19728; Santa Cruz Biotechnology, sc-20096, sc-17347; Sigma, HPA000536) on western blots and immunostaining. Three antibodies for each of SOX4 and SOX11 recognized a single band of the expected size on western blots but none of them were suitable for immunostaining or ChIP assay. In addition, a few antibodies showed nuclear immunostaining but also had strong background staining of tissue sections of the knockout brain, so we concluded that they are not specific.

Plasmids. For expression of *SoxC* genes and *Fezf2*, full-length complementary DNAs (mouse *Sox4*, BC052736; mouse *Sox11*, BC062095; mouse *Sox12*, BC067019, zebrafish *fezf2*, BC085677) were inserted into pCAGEN. For ChIP and electrophoretic mobility shift assay, PCR-amplified products (for mouse *Sox4* and *Sox11* full-length complementary DNA and *Sox11(1–276aa)*) were inserted into pcDNA3.1/V5-His TOPO vector (Invitrogen). For luciferase reporter plasmids, PCR-amplified products (for E1 to E4, E4F1 to E4F4 of mouse and E4F2 of additional species) and annealed, 190-bp complementary synthetic oligonucleotides (for MoE4F2-m1-3 and ZeE4F2-m1-3) were inserted into pGL4 (Promega). The sequences of the PCR primers and synthetic oligonucleotides are listed in Supplementary Table 1.

Gene expression analysis. To show which genes in the SOX family are highly correlated with human *FEZF2*, we used the previously generated data[45] for pairwise Pearson comparisons. This data set is generated by exon arrays and is available from the Human Brain Transcriptome database (http://www.humanbraintranscriptome.org) and the NCBI Gene Expression Omnibus under the accession number GSE25219. It covers 16 brain cortical regions over 15 periods, ranging in stage from embryonic development to late adulthood. The evaluation of gene expression in each region and each period was detailed in the study[45]. In this study, we averaged gene expression values of 11 neocortical regions in each of 15 developmental periods represented in the data set, and then performed pairwise Pearson correlation analysis for *FEZF2* and the members of the SOX family. In total, 18 *SOX* genes were represented in the database. The correlation analysis was shown by heatmap and the correlation coefficients were indicated in each cell (Supplementary Fig. 4). Mouse *Fezf2*, *Sox4*, *Sox5*, *Sox11*, and *Sox12 in situ* hybridization images were obtained from the Genepaint database[55] (http://www.genepaint.org).

Chromatin immunoprecipitation. After testing available anti-SOX4 and anti-SOX11 antibodies and finding none to be specific and suitable for ChIP assay, we performed ChIP assay with Neuro-2a cells that has been transfected with pcDNA3-Sox4-V5 or pcDNA3-Sox11-V5 plasmid using Lipofectamine 2000 (Invitrogen). At 36 h after transfection, cells were crosslinked in 1% formaldehyde for 10 min at 37 °C and processed for ChIP assay using the EZ-ChIP kit (Millipore) according to the manufacturer's instructions. Capture of the DNA fragments was tested by PCR using primers 5′-TGGAGAGAAGGCCAACAAAC-3′ (E1, forward); 5′- GCTGGGGATGGAGAAGAATA-3′ (E1, reverse); 5′-TCACCAAAGCGCCTTTTTAT-3′ (E2, forward); 5′-GTAAGCGGACATGCCATTTT-3′ (E2, reverse); 5′-TGACTTTCCCCAGCCTTCTA-3′ (E3, forward); 5′-CAACAGCTCACCCACACAAT-3′ (E3, reverse); 5′-ATGCCTAGCCCCAAAGAAAT-3′ (E4, forward) and 5′-TTAACTCCCCCTTTGGCTCT-3′ (E4, reverse).

Electrophoretic mobility shift assay. Double-stranded DNA probes were generated by annealing with complementary single-stranded oligonucleotides. Each DNA probe was biotin-end-labelled with Klenow enzyme and then purified using the G-50 Sephadex Column (Roche). The labelled and unlabelled (cold probe; 100-fold) probes were incubated with SOX4-V5 and SOX11(1–276aa)-V5 fusion proteins, which were produced using a TNT Quick coupled transcription/translation kit (Promega) in a DNA-binding buffer (75 mM NaCl, 1 mM EDTA, 1 mM DTT, 10 mM Tris-HCl (pH 7.5), 6% glycerol, 2 μg BSA, 16 ng poly (dI-dC) and 0.1 μg salmon sperm DNA) at 30 °C for 30 min. For supershift assay, anti-V5 antibody (Invitrogen) was added into the binding reaction and incubated for an additional 10 min. The shift assay was carried out in 6% polyacrylamide gel in ×0.5 TBE buffer. Subsequently, the labelled DNA was transferred on a nylon membrane (Amersham Biosciences). For detection, we used the Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific) according to manufacturers' recommendations. All primer sequences are described in Supplementary Table 1.

Luciferase assays. Neuro-2a cells were transfected using Lipofectamine 2000 (Invitrogen) with one of pCAG-Sox5, pCAG-Sox4, pCAG-Sox11, pCAG-Sox12 or empty pCAGEN, together with one of the pGL4 (Promega) luciferase vectors generated with enhancer sequences as described above. A *Renilla* luciferase plasmid (pRL, Promega) was co-transfected to control for transfection efficiency. The luciferase assays were performed 48 h after transfection using the dual-luciferase kit (Promega) according to the manufacturer's instructions. Primary cortical neurons were prepared from E14.5 heterozygous littermate control (*Sox4*$^{fl/+}$;*Sox11*$^{fl/+}$;*Emx1-cre* and *Sox4*$^{fl/+}$;*Sox11*$^{fl/+}$) and *Sox4;Sox11* cdKO (*Sox4*$^{fl/fl}$;*Sox11*$^{fl/fl}$;*Emx1-cre*) cortices and transfected with pGL4 or pGL4-E4 with pRL using the AMAXA Mouse Neuron Nucleofector Kit (VPG-1001; Lonza). At 48 h after transfection, the luciferase assays were performed as described above.

***In utero* electroporation.** Plasmid DNA (4 μg μl^{-1}) mixture (pCAGEN-Cre;pCALNL-Gfp or pCAGEN-Cre;pCALNL-Gfp;pCAGEN-fezf2) was injected into the lateral ventricles of embryonic mice at E12.5 and transferred into the cells of ventricular zone by electroporation (five 50-ms pulses of 40 V at 950-ms intervals) as described elsewhere[35,39]. Brains and tissue sections of electroporated animals were analysed for GFP expression after fixation with 4% paraformaldehyde at P0.

Quantitative RT–PCR. Neocortex was dissected from P0 brain and RNA was isolated using the RNeasy kit (Qiagen). Quantitative RT–PCR was performed using primers listed in Supplementary Table 1 and the SYBR FAST qPCR kit (KAPA Biosystems) according to the manufacturer's instructions. Gene expression levels were normalized to *Gapdh* expression.

51. Penzo-Méndez, A., Dy, P., Pallavi, B. & Lefebvre, V. Generation of mice harboring a Sox4 conditional null allele. *Genesis* **45,** 776–780 (2007).

52. Kawamoto, S. *et al.* A novel reporter mouse strain that expresses enhanced green fluorescent protein upon Cre-mediated recombination. *FEBS Lett.* **470,** 263–268 (2000).
53. Iwasato, T. *et al.* Dorsal telencephalon-specific expression of Cre recombinase in PAC transgenic mice. *Genesis* **38,** 130–138 (2004).
54. Liu, P. T., Jenkins, N. A. & Copeland, N. G. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res.* **13,** 476–484 (2003).
55. Visel, A., Thaller, C. & Eichele, G. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res.* **32,** D552–D556 (2004).

ARTICLE

doi:10.1038/nature11127

Chemical genetic discovery of targets and anti-targets for cancer polypharmacology

Arvin C. Dar[1]*, Tirtha K. Das[2]*, Kevan M. Shokat[1] & Ross L. Cagan[2]

The complexity of cancer has led to recent interest in polypharmacological approaches for developing kinase-inhibitor drugs; however, optimal kinase-inhibition profiles remain difficult to predict. Using a Ret-kinase-driven *Drosophila* model of multiple endocrine neoplasia type 2 and kinome-wide drug profiling, here we identify that AD57 rescues oncogenic Ret-induced lethality, whereas related Ret inhibitors imparted reduced efficacy and enhanced toxicity. *Drosophila* genetics and compound profiling defined three pathways accounting for the mechanistic basis of efficacy and dose-limiting toxicity. Inhibition of Ret plus Raf, Src and S6K was required for optimal animal survival, whereas inhibition of the 'anti-target' Tor led to toxicity owing to release of negative feedback. Rational synthetic tailoring to eliminate Tor binding afforded AD80 and AD81, compounds featuring balanced pathway inhibition, improved efficacy and low toxicity in *Drosophila* and mammalian multiple endocrine neoplasia type 2 models. Combining kinase-focused chemistry, kinome-wide profiling and *Drosophila* genetics provides a powerful systems pharmacology approach towards developing compounds with a maximal therapeutic index.

The cellular kinase-signalling network is a major regulator of cancer progression. Kinase-signalling pathways are often involved in pathogenesis, and kinase mutations are common and potent drivers of oncogenesis[1–4]. Targeting a single kinase has proven successful in some cases; examples include drugs that inhibit BCR–ABL, as well as members of the EGFR and RAF class of proteins[5–7]. However, results of this approach have been mixed[8–10]. Difficulties include rapidly emerging resistance as well as considerable toxicity that can limit dosing to levels that are insufficient for blocking tumour growth.

By contrast, most drugs approved for clinical use have multiple targets[11–13]. For many, or perhaps most, 'off-target' activities contribute to the overall efficacy of a drug. Sorafenib provides a recent example[14]: it was developed initially as an inhibitor of RAF kinase, but its efficacy in renal and hepatocellular cancer was later attributed to inhibition of VEGFR2 and PDGFR and potentially other targets[15]. Sorafenib highlights the therapeutic potential of targeting multiple kinases but also the uncertainty and serendipity of phenotype-based screening.

Most multiple endocrine neoplasia type 2 (MEN2) patients have an autosomal-dominant activating germline mutation in the RET (rearranged during transfection) receptor tyrosine kinase that is necessary and probably sufficient to direct a series of transformation events including medullary thyroid carcinoma (MTC)[16,17]. To identify candidate compounds with optimal polypharmacological profiles, we synthesized a panel of inhibitors with potency against RET (a traditional target-based approach) that additionally target distinct downstream kinases. We demonstrate how stepwise testing in *Drosophila* models of the disease subtype MEN2B[18] uncovered a spectrum of targets contributing to drug-induced efficacy and toxicity. Our results present a new approach to rational drug development that combines aspects of target- and phenotype-based drug discovery; it relies on whole-animal screening to both explore the mechanism of a drug and identify an optimal polypharmacological profile for suppressing tumours *in vivo*.

Identifying AD57 in a whole-animal *Drosophila* screen

We previously reported a *Drosophila* MEN2B model in which an activating intracellular mutated isoform of the *Drosophila* Ret orthologue (dRet) was targeted to the eye[18]. This $\mathrm{dRet}^{\mathrm{MEN2B}}$ model proved useful for validating whole-animal efficacy of the kinase inhibitor vandetanib (also known as ZD6474, Caprelsa)[19], a drug recently approved for sporadic MTC and for MTC arising in patients with MEN2 (ref. 20). To improve its utility for drug screening, we developed a quantitative viability assay that uses the GAL4/upstream activating system (UAS) to target oncogenic $\mathrm{dRet}^{\mathrm{MEN2B}}$ to multiple developing epithelial tissues (Fig. 1a; T.K.D. *et al.*, in preparation). Specifically, oncogene expression is driven by the *patched* (*ptc*) promoter, which directs expression in a dynamic pattern including developing epithelia (for example, wing, eye and leg) and other tissues[21]. We calibrated the $ptc > dRet^{MEN2B}$ assay to permit 50% survival to pupariation and 0% survival to adulthood. Oral administration of clinical kinase inhibitors[22,23] resulted in weak (vandetanib), mild (sunitinib) or stronger (sorafenib) rescue (Fig. 1b), validating our assay. Notably, sorafenib rescued some animals to adulthood but did not considerably increase the proportion that developed to pupariation, indicating some efficacy but also toxicity (reduced survival) at optimal doses.

We developed and screened a library of polypharmacological compounds that target Ret in addition to other classes of kinases[24] (Supplementary fig. 1). One compound, AD57, potently suppressed $ptc > dRet^{MEN2B}$ lethality in the larva, rescuing approximately 25% of animals to adulthood (Fig. 1b, c). Rescued adults also showed complete suppression of notum and scutellum defects that were observed in un-eclosed control pupae (Fig. 1c), and were fully active and fertile. AD57 demonstrated both an improved efficacy and toxicity profile in our assay compared with other kinase inhibitors (Fig. 1b).

AD57 exhibited improved activity compared to analogues

The overall structure of AD57-like compounds includes two fragments fused through a urea linker (Fig. 1d). Shared features include

[1] Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158, USA. [2] Department of Developmental and Regenerative Biology, Mount Sinai School of Medicine, New York, New York 10029, USA.
*These authors contributed equally to this work.

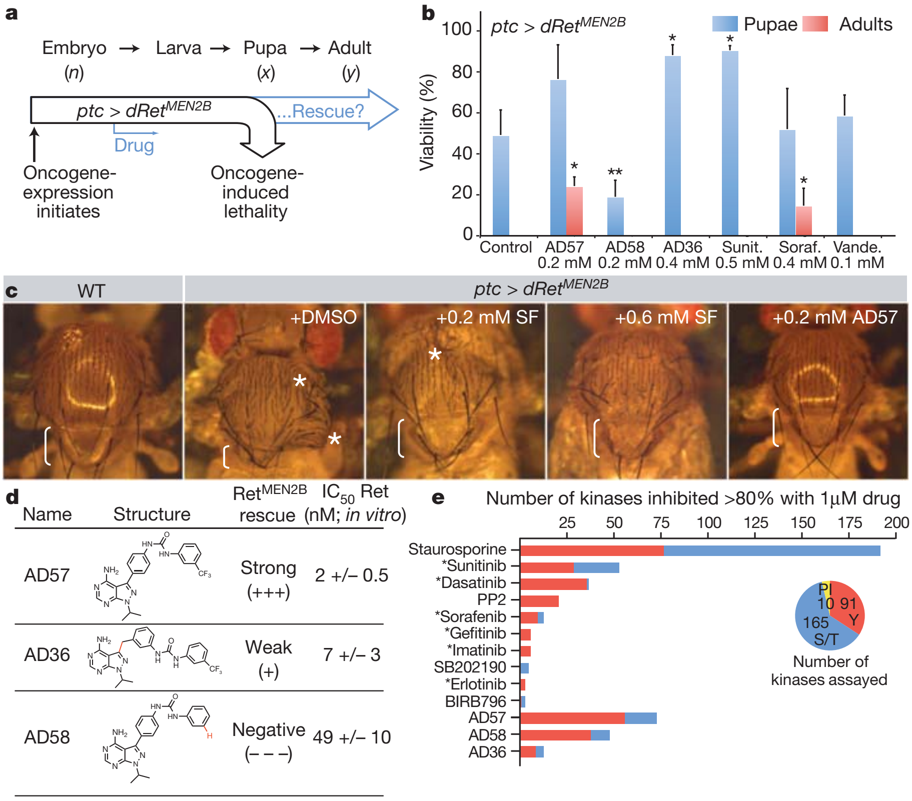


Figure 1 | Screening for an optimal therapeutic index in a *Drosophila* MEN2B model yields a polypharmacological kinase inhibitor. **a**, Suppression of dRet^{MEN2B}-induced developmental block and whole-animal toxicity were scored based on the number of embryos (n) that survived as pupae (x) and adults (y). **b**, Per cent viability of control- or drug-treated flies determined for pupae (x per n) and adults (y per n). AD57 emerged as the best single-agent hit from the screen. Asterisks indicate significance comparing to control using Student's t-test ($P < 0.05$ for adults in AD57 and sorafenib treatments, and $P < 0.05$ for pupae for the rest). Error bars denote s.e.m. Total n of 200, 75, 98, 54, 91, 280 and 209, from left to right. Soraf., sorafenib; Sunit., sunitinib; Vande., vandetanib. **c**, *ptc* > *dRet*^{MEN2B} adults have notum defects including excessive bristles (asterisks) and scutellum defects (brackets); controls (+ dimethylsulphoxide (DMSO)) died as un-eclosed adults. AD57 strongly suppressed whereas sorafenib (SF) weakly suppressed these defects, yielding fully eclosed adults. Width of each wild-type notum is ~0.75 mm. WT, wild type. **d**, Structure–activity relationships suggest that dRet inhibition alone is insufficient to rescue MEN2B flies. IC$_{50}$ values were determined against a purified form of human Ret. **e**, The AD series of compounds showed broad-spectrum kinase-inhibition profiles. Clinical (asterisks) and known kinase inhibitors are shown for comparison. The number of lipid (PI), tyrosine (Y) and serine/threonine (S/T) kinases tested are shown in the pie chart.

a pyrazolopyrimidine core that functions as a mimic of adenosine or hinge-binder and a hydrophobic element that binds within an allosteric pocket of the kinase domain (Supplementary Fig. 1b). AD36, a close analogue of AD57, contains a methylene group between the pyrazolopyrimidine ring and fused phenyl portion, whereas the analogue AD58 does not contain the trifluoromethyl group (Fig. 1d). These subtle structural changes led to substantial changes in biological activity; AD36 showed some efficacy (increased numbers of pupae but no adults), whereas AD58 induced considerable toxicity without detectable efficacy (fewer pupae and adults; Fig. 1b).

These results demonstrate the sensitivity of whole-body phenotyping in *Drosophila* to detect the effects of conservative structural differences between drug candidates. The difference between AD36 and AD57 was particularly notable because both demonstrate similar potency for Ret *in vitro* (Fig. 1d); indeed, our analysis of other kinase inhibitors indicated that efficacy did not correlate solely with inhibition of Ret (Fig. 2c and Supplementary Fig. 2; data not shown). This suggested that targeting of additional kinases is necessary for the biological efficacy of AD57.

AD57 suppressed *dRet*^{MEN2} transformation

We previously developed a wing-based assay for transformation and cell migration in which the *ptc-Gal4* driver directed oncogene expression in a stripe along the anterior–posterior axis[25]. Adapting it to *ptc* > *dRet*^{MEN2B} wings led to overproliferation, basal constriction and cell migration away from the *ptc* domain (Fig. 2a; T.K.D. *et al*, in preparation). Oral administration of AD57 blocked each of these phenotypes (Fig. 2a). Sorafenib, sunitinib and vandetanib showed substantially weaker rescue (Fig. 1b; data not shown). We conclude that oral administration of AD57 is particularly effective at suppressing dRet-mediated transformation at doses that are non-toxic to the fly.

In standard mammalian RET^{MEN2} models, AD57 potently inhibited the viability of MEN2B (MZ-CRC-1) and MEN2A (TT) patient-derived cell lines with a half-maximum inhibitory concentration (IC$_{50}$) that was several orders more potent than sorafenib, vandetanib, AD36 or AD58 (Supplementary Fig. 3a, b). In a conventional mouse xenograft model, AD57 significantly suppressed TT-based tumour growth at a dose (20 mg kg^{-1}) that demonstrated no detectable toxicity as assessed by animal weights (Supplementary Fig. 3c, d). Together, our data indicate that *Drosophila in vivo* assays provide a useful tool for identifying compounds with improved *in situ* efficacy and toxicity profiles.

Erk and Src inhibition suppressed dRet signalling

Our previous *Drosophila* genetic screens emphasized three major pathways for dRet^{MEN2B}-mediated transformation: Ras/Raf, Src and glucose metabolism/PI3K[18] (Fig. 2b; data not shown). Furthermore, assessing AD57, AD36 and AD58 in a broad *in vitro* mammalian-kinase panel indicated that small perturbations in the structure of AD57 led to considerable changes in kinase selectivity (Fig. 1e and Supplementary Tables 1–3). For example, AD57 is a potent inhibitor of the pathway-relevant human kinases BRAF, S6K (also known as RPS6KB1), mTOR and SRC (Fig. 2c). By comparison, AD58 is a much weaker inhibitor of S6K and BRAF but is more potent against mTOR; AD36 is nearly inactive against mTOR, S6K and SRC, potentially reflecting steric clash at the gatekeeper position (compare, for example, ABL(T315I), EGFR(T790M) and RET(V804L); Supplementary Fig. 4). We focused on effectors of RAS, PI3K and SRC, although other targeted pathways may also contribute to compound activity.

We demonstrated previously that activation of Src is sufficient to direct many of the aspects we observed within the *ptc* > *dRet*^{MEN2B} domain[25–27] and we explored its activity *in situ*. Expressing *ptc* > *dRet*^{MEN2B} led to high levels of activated phospho-Src at the basal invading front of transformed cells (Fig. 2a). In addition to suppressing invasion, oral administration of AD57 suppressed phospho-Src in basal regions of the wing epithelium (Fig. 2a). Distinctions with AD36 and AD58 were instructive. AD36 failed to suppress the invasion or basal migration of *ptc* > *dRet*^{MEN2B} cells and, as predicted by our *in vitro* assay, phospho-Src remained at high levels at the basal leading edge (Fig. 2a). Also as predicted, AD58

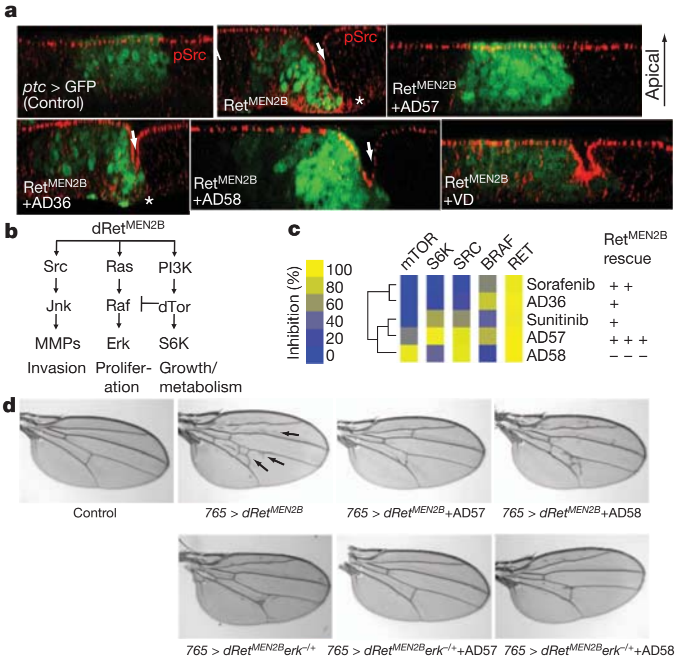


Figure 2 | Multiple-pathway inhibition by AD57 mitigates dRet-directed phenotypes. **a**, z-series confocal images of larval wing epithelia; virtual cross-section through tissue with apical up. Control tissue shows apical phospho-Src (pSrc) expression (red) in the junctions. *ptc* > *dRet*^{MEN2B} wing cells (green fluorescent protein; GFP$^+$) shifted basally (arrows) and invaded below adjacent wild-type tissue; phospho-Src emerged at the basal invading front (asterisks). These phenotypes were strongly suppressed by AD57 but not by AD36, AD58 or vandetanib (VD). Apical–basal distance is ~45 µm; imaged with ×63 oil. **b**, Partial list of signalling pathways activated by oncogenic dRet^{MEN2B}. MMPs, matrix metalloproteinases. **c**, Per cent *in vitro* kinase inhibition profiles (left) and relative rescue (right) are shown. Tree indicates similarity of compounds on the basis of hierarchical clustering of per cent kinase inhibition. **d**, Broad dRet^{MEN2B} expression led to ectopic wing veins (arrows), reflective of hyperactive Ras pathway signalling. The wing defects were suppressed by AD57 and enhanced by AD58. Removal of one functional copy of *erk*/*rolled* (*erk*$^{-/+}$) enhanced rescue by AD57 and AD58. Quantified in Supplementary Figure 5c.

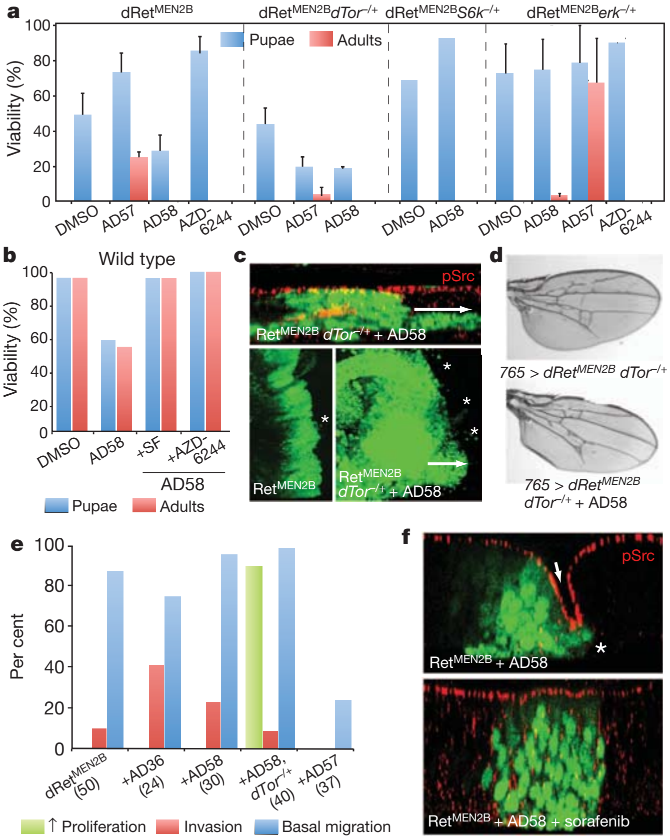


Figure 3 | Feedback downregulation of the Ras pathway through the anti-target Tor. **a**, Reducing *dTor* gene dosage decreased per cent viability of AD57- and AD58-treated dRet^{MEN2B} flies ($P < 0.05$ comparing pupae (AD57, AD58) or adults (AD57)). Conversely, reducing *erk* gene dosage enhanced survival of both ($P < 0.05$ comparing pupae (AD58) or adults (AD57)). Treatment with a specific MEK inhibitor alone, AZD6244, in control (*ptc* > *dRet*^{MEN2B}) or *erk*$^{-/+}$ flies did not rescue viability compared to AD57-treated flies, suggesting its level of Ras pathway suppression is close to optimal ($P > 0.05$ comparing AZD6244 pupae). Reducing S6K (*S6k*$^{-/+}$) partially mitigated toxicity from AD58 treatment. Column bars represent the mean of three separate experiments. Total *n* of 214, 58, 63, 130, 254, 66, 118, 52, 59, 190, 93, 114 and 96 from left to right. Error bars denote s.e.m. **b**, Decreased viability of wild-type flies by AD58 was mitigated by co-administration of sorafenib or AZD6244. Total *n* of 118, 47, 51 and 28 from left to right. **c**, Reducing *dTor* strongly enhanced AD58-mediated invasion (asterisks, arrow) and excess proliferation. Top panel, a lateral reconstruction; compare with Fig. 2a. Bottom panels represent an apical view, constructed as a z-series overlay of confocal images spanning the full depth of the wing disc epithelia. It shows how cells migrate from the *ptc* domain to distant sites (for example, arrow, asterisks), a phenotype strongly enhanced in the presence of *dTor*$^{-/+}$ plus AD58. Top panel (GFP+) width is ~150 µm; imaged with ×63. Bottom right panel width (GFP+) is ~150 µm; both bottom panels were imaged with ×40. **d**, Wing defects in *ptc* > *dRet*^{MEN2B} *dTor*$^{-/+}$ adults were further enhanced by AD58. **e**, Quantification of *ptc* > *dRet*^{MEN2B} phenotypes. Invasion was established by scoring for single or groups of GFP-labelled cells that relocated away from the *ptc* boundary (Fig. 3c, asterisks). Basal migration was scored as indentation of the apical surface (see Fig. 2a, arrows). Proliferation was scored as significant widening of the *ptc* boundary. The number of wings analysed under each condition is indicated in brackets. Reduced *dTor* increased proliferation in the presence of AD58 as well as reducing survival; all aspects were improved by feeding AD57 whereas increased invasion by feeding AD36 did not translate to reduced survival. **f**, Migration of *dRet*^{MEN2B}-transformed cells was blocked by co-treatment with AD58 plus sorafenib (bottom). Treatment with similar doses of AD58 (top) or sorafenib alone (not shown) did not suppress migration. Arrow indicates constriction of apical cell surface and asterisk indicates basal invading front. Apical–basal distance is ~45 µm; imaged with ×63.

suppressed basal phospho-Src accumulation, yet it failed to prevent invasion and basal migration (Fig. 2a). These data support the view that Src inhibition contributes to reducing invasion and basal migration, but suggest that other targets are also required.

Elevated Ras/Erk pathway activity leads to ectopic veins in the adult wing (for example, refs 28, 29). Expression of oncogenic dRet throughout the developing wing (*765* > *dRet*^{MEN2B}) led to disruption of the overall adult wing pattern, including ectopic wing veins. Reducing gene dosage of the *erk* orthologue *rolled* suppressed these phenotypes, confirming that wing vein formation is dependent on Ras/Erk activity (Fig. 2d). *dRet*^{MEN2B}-dependent wing phenotypes were suppressed by AD57 (Fig. 2d); by contrast, vandetanib had little effect (Supplementary Fig. 5a, b). Notably, the ectopic wing vein phenotype was slightly but consistently enhanced with AD58 treatment (Fig. 2d). This enhancement was suppressed by removing a functional copy of *erk* (Fig. 2d and Supplementary Fig. 5c), further indicating that AD58 treatment actually increased Ras pathway signalling. These data raise the possibility that AD58 toxicity was due to excess Ras pathway activity and that further suppressing Ras pathway activity would improve the overall efficacy of AD57.

Unbalanced dTor inhibition promotes toxicity

AD58 directed substantial whole-animal toxicity when fed to *ptc* > *dRet*^{MEN2B} or wild-type flies (Figs 1a, 3a, b and Supplementary Fig. 6), providing us with an opportunity to explore aspects of AD drug-class toxicity. On the basis of *in vitro* kinase data, AD58 is a stronger inhibitor of mTOR and a weaker inhibitor of RAF than AD57 (Fig. 2c and Supplementary Fig. 4c). Recently, mTOR has been demonstrated to provide feedback inhibition of the RAS pathway in mammals[30,31]. We therefore assessed whether the high toxicity observed for AD58 was due in part to high inhibition of Drosophila target of rapamycin (dTor) coupled with low inhibition of Raf, leading to hyperactivation of *Drosophila* Ras85D pathway signalling throughout the animal.

Reducing dTor (*ptc* > *dRet*^{MEN2B}, *dTor*$^{-/+}$) dominantly suppressed the efficacy of AD57 and enhanced the toxicity of AD58 (Fig. 3a). Quantitative phenotypic assessment indicated that AD58-induced toxicity was due primarily to an increase in proliferation (Fig. 3c, e). Also, reducing the gene dosage of *dTor* enhanced the AD58-induced ectopic wing vein formation (Fig. 3d) and suppressed efficacy of AD57 on wing vein patterning (Supplementary Fig. 5b), indicating that reducing dTor increased Erk activity. Notably, removing a genomic copy of the dTor target S6K suppressed AD58 toxicity (Fig. 3a), indicating that S6K is independent of the dTor feedback loop.

We also assessed whether reducing the activity of *Drosophila* Ras85D pathway components could abrogate the effects of dTor inhibition. AD58-mediated toxicity in wild-type flies was almost completely suppressed by co-feeding with the Raf inhibitor sorafenib or Mek inhibitor AZD6244 (Fig. 3b). Combining AD58 with sorafenib also resulted in considerable suppression of invasion and migration within *ptc* > *dRet*^{MEN2B} wing discs (Fig. 3f). Removing a genomic copy of *erk*/*rolled* considerably improved AD57 rescue (Fig. 3a). Together, these data indicate that both AD57 and AD58 act to inhibit dTor activity, but that failure of AD58 to suppress Raf kinase led to elevated Ras pathway activity. Elevated Erk, in turn, led to poor efficacy against the tumour and high whole-body toxicity (see Fig. 5 for pathway logic).

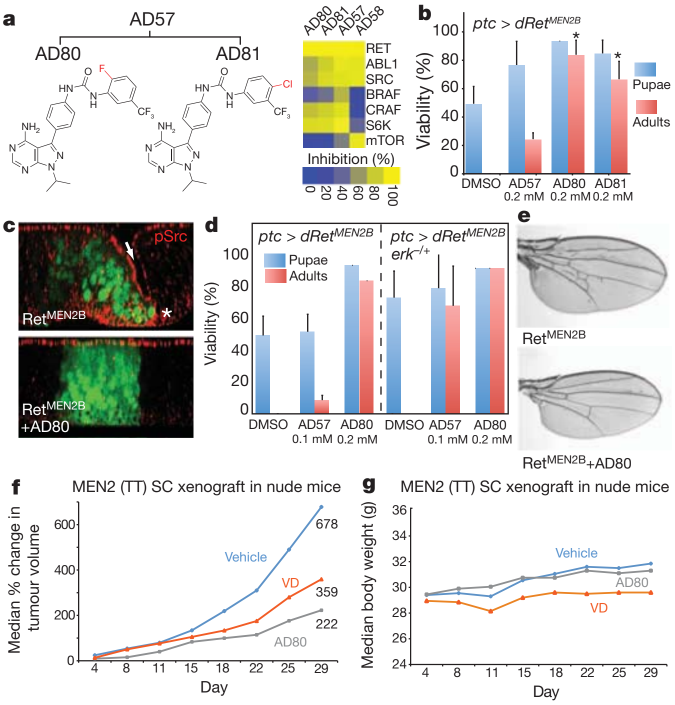


Figure 4 | Balanced kinase polypharmacology provides optimal efficacy and toxicity. **a**, Chemical structures of the AD57 derivatives AD80 and AD81 and percentage inhibition of relevant targets at 1 μM. Unlike AD57 and AD58, both lack significant inhibitory activity against mTOR. **b**, AD80 and AD81 showed improved rescue relative to AD57. **P* < 0.05, significance compared to AD57 in a two-tailed Student's *t*-test. Total *n* of 214, 58, 109 and 99 from left to right. Error bars denote s.e.m. **c**, Basal migration (arrow) of dRet^{MEN2} cells and basal phospho-Src (pSrc; asterisk) were blocked by AD80. Apical–basal distance is ~45 μm; imaged with ×63. **d**, Reducing *erk* gene dosage enhanced survival of AD57 (*P* < 0.05 for adult flies compared across genotypes) but not AD80 (*P* > 0.5 for adults compared across genotypes), suggesting that the Tor feedback loop was not altered by AD80 and that Erk was optimally suppressed in flies. Total *n* of 214, 43 and 109 from left to right. Error bars denote s.e.m. **e**, *765* > *dRet*^{MEN2B}-dependent extra wing vein phenotype was fully rescued by AD80. **f**, AD80 and vandetanib (VD) reduced tumour progression 3.1- and 1.9-fold, respectively, relative to vehicle-treated nude mice transplanted with TT cells. Change in tumour volume was calculated per mouse and shown are the median per group. Twenty vehicle- and ten drug-treated mice were analysed for each treatment group. **g**, Corresponding body-weight measurements.

AD80 and AD81 demonstrated an improved profile

Together, our genetic and chemical data indicate that an optimal drug for MEN2B would show activity against Ret, Src, S6K and Raf but limited activity against Tor. To test this logic and potentially improve AD57, we developed a series of new AD-based analogues. From our previously determined structure of AD57 in complex with c-Src we reasoned that modifying the terminal phenyl group of AD57 would selectively perturb dTor binding without altering inhibitory interactions with dRet, Raf, S6K or Src. We therefore generated two compounds, AD80 and AD81, into which ortho-fluorine and para-chlorine groups were respectively incorporated (Fig. 4a).

On the basis of their *in vitro* human kinase profiles, AD80 and AD81 inhibited RET, RAF, SRC and S6K, with greatly reduced mTOR activity relative to AD57 and AD58 (Fig. 4a and Supplementary Fig. 7). Oral administration of either AD80 or AD81 resulted in a notable 70–90% of animals developing to adulthood in our *Drosophila* *ptc* > *dRet*^{MEN2B} model, a considerable improvement over the efficacy observed with AD57 and all other compounds we have tested until now (Fig. 4b). Focusing on AD80, ectopic Src activation (Fig. 4c) and wing vein pattern phenotypes (Fig. 4e and Supplementary Fig. 6b) were strongly suppressed, indicating that Src and Ras activities were restored to normal levels. The result was phenotypically normal *ptc* > *dRet*^{MEN2B} adults, showing rescue that exceeded AD57 or sorafenib, which yielded adults with some cuticle defects. Notably, although reducing *erk* gene dosage (*ptc* > *dRet*^{MEN2B}, *erk*$^{-/+}$) in the fly considerably enhanced the efficacy of AD57 and AD58 in viability assays, it did not alter efficacy of AD80 treatment (Fig. 4d). This indicates that AD80 is optimal for Ras–Erk pathway inhibition (Supplementary Fig. 2).

The improved profile of AD80 also translated to mammalian MEN2 models. AD80 inhibited proliferation of MZ-CRC-1 and TT thyroid cancer cells in culture, probably through the induction of apoptosis (Supplementary Fig. 8). Immunoblot analysis demonstrated potent downregulation of phosphorylated Ret and several downstream biomarkers within these cells (Supplementary Fig. 9). AD80 also promoted enhanced tumour growth inhibition and reduced body-weight modulation relative to vandetanib in a mouse xenograft model (Fig. 4f, g and Supplementary Table 4).

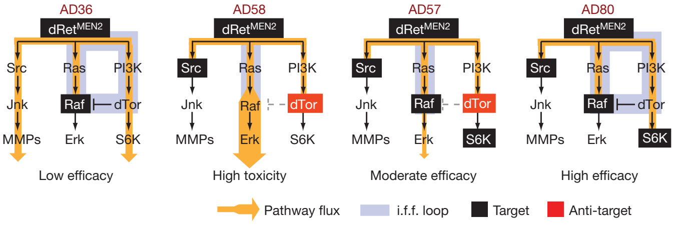


Figure 5 | Differential polypharmacology and outcomes from the AD compounds. Models to explain the AD series of compounds in dRet^{MEN2B} transgenic flies. Pathway components blocked by inhibitors have been boxed, with resulting flux indicated by orange lines and arrows, and the incoherent feed-forward (i.f.f.) loop highlighted in blue. Grey dashed lines indicate loss of dTor inhibition. Targets in black boxes contribute to efficacy whereas inhibition of the anti-target dTor (red) leads to hyperactivation of the Ras pathway, causing high toxicity in the MEN2 model. The polypharmacological profile of AD80 best addresses the three key pathways, providing high drug efficacy and optimal therapeutic index.

Discussion

Here we describe a systems pharmacology approach for cancer drug discovery that focuses on whole-animal testing, chemistry and genetics to identify a single agent with an optimized polypharmacological profile. Using a stepwise approach that combined genetics and chemistry, we identified AD80 and AD81 as polypharmacological agents with an optimal balance of activity against Ret, Raf, Src, Tor and S6K that show high efficacy with very low toxicity (Fig. 5). Our studies indicate that these drugs may be an improvement over existing compounds including vandetanib, a kinase inhibitor demonstrated by our group and others to act on Ret-based tumorigenesis[19,32] that has recently been approved for MTC patients. Details of human tumours can differ substantially from *Drosophila* cancer models and mouse xenografts, and the true predictive value of this approach must await further testing. A related approach is to assess drug combinations; however, in addition to the increased cost of clinically testing a mix of compounds, complex target-profile interactions and differing pharmacokinetics can make executing clinical trials challenging.

The connection between the Tor and Ras pathways within the MEN2B model is reminiscent of a general network motif termed an incoherent feed-forward loop[33]: here, dRet^{MEN2B} activates Ras but also represses Ras signalling by activating dTor. This network motif has been identified within diverse contexts, including transcriptional and neuronal networks, as a means to tune cellular responses to incoming signals[33]. Perhaps this motif will prove common within cancer signalling networks, providing a useful place to search for other anti-targets that limit the therapeutic benefits of kinase inhibitors.

METHODS SUMMARY

Inhibitor studies in *Drosophila*, fly stocks, genetics, histology, antibodies, mammalian cell and xenograft studies, imaging, western blotting, *in vitro* kinase assays, kinase-inhibitor profiling, chemical synthesis and other procedures were performed as described in Methods.

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

Received 18 July 2011; accepted 5 April 2012.

1. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
3. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
4. Cancer Genome Atlas Research Network.. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
5. Druker, B. J. Translation of the Philadelphia chromosome into therapy for CML. *Blood* **112,** 4808–4817 (2008).
6. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363,** 809–819 (2010).
7. Geyer, C. E. *et al.* Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* **355,** 2733–2743 (2006).
8. Boss, D. S., Beijnen, J. H. & Schellens, J. H. Clinical experience with aurora kinase inhibitors: a review. *Oncologist* **14,** 780–793 (2009).
9. Haura, E. B. *et al.* A Phase II study of PD-0325901, an oral MEK inhibitor, in previously treated patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* **16,** 2450–2457 (2010).
10. LoRusso, P. M. *et al.* Phase I pharmacokinetic and pharmacodynamic study of the oral MAPK/ERK kinase inhibitor PD-0325901 in patients with advanced cancers. *Clin. Cancer Res.* **16,** 1924–1937 (2010).
11. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nature Rev. Cancer* **10,** 130–137 (2010).
12. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnol.* **26,** 127–132 (2008).
13. Mestres, J. *et al.* The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.* **5,** 1051–1057 (2009).
14. Wilhelm, S. *et al.* Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nature Rev. Drug Discov.* **5,** 835–844 (2006).
15. Ahmad, T. & Eisen, T. Kinase inhibition with BAY 43-9006 in renal cell carcinoma. *Clin. Cancer Res.* **10,** 6388S–6392S (2004).
16. Lairmore, T. C. *et al.* A 1.5-megabase yeast artificial chromosome contig from human chromosome 10q11.2 connecting three genetic loci (*RET*, *D10S94*, and *D10S102*) closely linked to the *MEN2A* locus. *Proc. Natl Acad. Sci. USA* **90,** 492–496 (1993).
17. Almeida, M. Q. & Stratakis, C. A. Solid tumors associated with multiple endocrine neoplasias. *Cancer Genet. Cytogenet.* **203,** 30–36 (2010).
18. Read, R. D. *et al.* A *Drosophila* model of multiple endocrine neoplasia type 2. *Genetics* **171,** 1057–1081 (2005).
19. Vidal, M. *et al.* ZD6474 suppresses oncogenic RET isoforms in a *Drosophila* model for type 2 multiple endocrine neoplasia syndromes and papillary thyroid carcinoma. *Cancer Res.* **65,** 3538–3541 (2005).
20. Wells, S. A. Jr *et al.* Vandetanib in patients with locally advanced or metastatic medullary thyroid cancer: a randomized, double-blind Phase III trial. *J. Clin. Oncol.* **30,** 134–141 (2012).
21. Hinz, U., Giebel, B. & Campos-Ortega, J. A. The basic-helix-loop-helix domain of *Drosophila* lethal of scute protein is sufficient for proneural function and activates neurogenic genes. *Cell* **76,** 77–87 (1994).
22. Wilhelm, S. M. *et al.* BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Res.* **64,** 7099–7109 (2004).
23. Sun, L. *et al.* Discovery of 5-[5-fluoro-2-oxo-1,2- dihydroindol-(3Z)-ylidenemethyl]-2,4- dimethyl-1H-pyrrole-3-carboxylic acid (2-diethylaminoethyl)amide, a novel tyrosine kinase inhibitor targeting vascular endothelial and platelet-derived growth factor receptor tyrosine kinase. *J. Med. Chem.* **46,** 1116–1119 (2003).
24. Dar, A. C., Lopez, M. S. & Shokat, K. M. Small molecule recognition of c-Src via the imatinib-binding conformation. *Chem. Biol.* **15,** 1015–1022 (2008).
25. Vidal, M., Larson, D. E. & Cagan, R. L. Csk-deficient boundary cells are eliminated from normal *Drosophila* epithelia by exclusion, migration, and apoptosis. *Dev. Cell* **10,** 33–44 (2006).
26. Read, R. D., Bach, E. A. & Cagan, R. L. *Drosophila* C-terminal Src kinase negatively regulates organ growth and cell proliferation through inhibition of the Src, Jun N-terminal kinase, and STAT pathways. *Mol. Cell. Biol.* **24,** 6676–6689 (2004).
27. Vidal, M. *et al.* Differing Src signaling levels have distinct outcomes in *Drosophila*. *Cancer Res.* **67,** 10278–10285 (2007).
28. Sawamoto, K. *et al.* The *Drosophila* secreted protein Argos regulates signal transduction in the Ras/MAPK pathway. *Dev. Biol.* **178,** 13–22 (1996).
29. Guichard, A. *et al.* *rhomboid* and *Star* interact synergistically to promote EGFR/MAPK signaling during *Drosophila* wing vein development. *Development* **126,** 2663–2676 (1999).
30. Gedaly, R. *et al.* PI-103 and sorafenib inhibit hepatocellular carcinoma cell proliferation by blocking Ras/Raf/MAPK and PI3K/AKT/mTOR pathways. *Anticancer Res.* **30,** 4951–4958 (2010).
31. Carracedo, A. *et al.* Inhibition of mTORC1 leads to MAPK pathway activation through a PI3K-dependent feedback loop in human cancer. *J. Clin. Invest.* **118,** 3065–3074 (2008).
32. Carlomagno, F. *et al.* ZD6474, an orally available inhibitor of KDR tyrosine kinase activity, efficiently blocks oncogenic RET kinases. *Cancer Res.* **62,** 7284–7290 (2002).
33. Alon, U. Network motifs: theory and experimental approaches. *Nature Rev. Genet.* **8,** 450–461 (2007).

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

Acknowledgements We thank the Bloomington Stock Center, Vienna *Drosophila* RNAi Center and C. Pfleger for reagents. T.K.D. and R.C. were supported by National Institutes of Health grants R01CA109730 and R01CA084309 and American Cancer Society Grant 120616-RSGM-11-018-01-CDD. T.K.D. was also supported by American Cancer Society Grant 120886-PFM-11-137-01-DDC. We thank members of the Shokat and Cagan laboratories for discussions. We thank members of the SelectScreen team at Invitrogen, in particular K. Vogel, for performing kinase-profiling services. K.M.S. thanks NIH R01EB001987, P01 CA081403-11 and The Waxman Foundation.

Author Contributions A.C.D. and T.K.D. contributed equally and are listed alphabetically. A.C.D. and T.K.D. conceived and designed experiments with K.M.S. and R.C. A.C.D. performed chemical synthesis, modelling, IC_{50} measurements, informatics and western blots on cancer cell lines. T.K.D. performed, imaged and analysed all *Drosophila* assays and cancer cell line viability assays. All authors discussed experimental data and wrote the manuscript.

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 K.M.S. (shokat@cmp.ucsf.edu).

METHODS

Inhibitor studies in flies. AZD6244 (Cal Biochem), sorafenib, sunitinib (LC Labs) and new AD-series drugs were dissolved in DMSO buffer and then diluted in molten (~50 °C)-enriched fly food and left to solidify at room temperature (25 °C) to yield the indicated final drug concentrations. 30–60 embryos of each genotype were raised on drug-containing food (500–1,000 µl) in 5-ml vials until they matured as third-instar larvae (wing disc migration and invasion assay) or allowed to proceed to adulthood (viability assay and wing vein quantification assay). Experiments were done in duplicate and repeated at least three times.

Fly stocks, genetics and subcloning. Fly stocks were obtained from Bloomington stock centres and C. Pfleger (*rolled*[1]). *UAS-dRet*^{MEN2B} flies were generated by ligating a partial EcoR1-digested *glass* multimerized repeat–*dRet*^{MEN2B} DNA fragment[18] into EcoR1 site of *pUAST* vector. Transgenic flies were generated by standard protocol.

Histology and antibodies. For wing disc invasion and migration analysis third-instar discs were staged and fixed in 4% paraformaldehyde. Immunofluorescence was performed as described[34]. Antibodies used were anti-phospho-Src(Y^{418}) (Invitrogen). AlexaFluor secondary antibodies were used for all immunofluorescence experiments. Confocal imaging used a Leica DM5500 Q microscope and image analysis was performed using Adobe Photoshop.

MTT assays using cancer lines. MZ-CRC-1 (MEN2B) and TT (MEN2A) cell lines were cultured in DMEM buffer and Ham's F12K media, respectively, supplemented with 10% bovine serum albumin (BSA) and a penicillin and streptomycin antibiotics mix. Cells were grown in 75 cm^2 sterile polystyrene culture flasks to 80% confluency, trypsinized and re-seeded in equal aliquots into 96-well plates. After 2 days and ~50% confluency, media was removed and replaced with DMSO or drug-containing media. Cells were allowed to grow for another 6 days, after which the thiazolyl blue tetrazolium bromide (MTT) assay was performed. Cell media was removed and replaced with MTT-containing media (1 mg ml^{-1} final concentration) and cells were allowed to grow at 37 °C for another 3.5 h. MTT media was removed and MTT precipitate dissolved in 4 mM HCl, 0.1% NP40 in isopropanol, solvent by shaking for 1 h. Spectrophotometric readings at 590 nm and 630 nm using a 96-well-plate reader were used to establish growth and viability of cells. Each drug dose was tested in quadruplicates and experiments repeated twice.

Western blotting of fly wing disc tissues. A total of ten third-instar discs of each treatment were dissolved in lysis buffer (50 mM Tris, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA) supplemented with protease-inhibitor cocktail and phosphatase-inhibitor cocktail (Sigma). Total protein in each sample was quantified using BIORAD protein assay. Samples were boiled, resolved on SDS–PAGE and transferred by standard protocols. Membranes were stripped with SIGMA Restore stripping buffer and re-probed with other antibodies to assess the signal under exactly the same loading conditions. Antibodies used were from Cell Signaling Technology. Only phospho-Src antibody was from Invitrogen.

Whole-mount imaging of fly notum and wings. For fly notum images, after completion of the viability assay, un-eclosed pupae were dissected out of their pupal cases, placed on double-sided tape and imaged under the Leica MZ16F stereomicroscope. Eclosed adults were imaged similarly. For adult wing vein analysis, wings from male flies were dissected and kept in 100% ethanol overnight, mounted on slides in 80% glycerol in phosphate-buffered saline solution and imaged by regular light microscopy using a Leica DM5500 Q microscope.

Xenograft analysis. 5×10^6 TT cells were injected subcutaneously into one flank of male *nu nu* mice. Mice showing established growing tumours were separated into vehicle or drug-treatment groups. A similar range of tumour sizes was selected for each experiment (vehicle versus AD57; vehicle versus AD80 versus vandetanib). Vehicle, AD57 (20 mg kg^{-1}), AD80 (30 mg kg^{-1}) or vandetanib (50 mg kg^{-1}) were administered by oral gavage (per os) once daily, five times a week. Tumour and body-weight measurements were performed three times per week. Mouse experiments were carried out by Washington Biotechnology (accreditation no. A192-01) according to the Public Health Services guideline, set forth by the Office for Laboratory and Animal Welfare division of the National Institutes of Health.

Western blotting of cancer cell lines. MZ-CRC-1 (MEN2B) and TT (MEN2A) cell lines were grown in 24-well plates in DMEM buffer (+ 4.5 g l^{-1} glucose; without L-Glu and pyruvate) and Ham's F12K (ATCC) media respectively; each supplemented with 10% heat-inactivated FBS and penicillin and streptomycin antibiotics. Cells were treated for 1 h with inhibitors or vehicle (0.1% DMSO). After treatment, media was removed and cells were washed twice with cold PBS and then lysed in radioimmunoprecipitation assay buffer (25 mM Tris, pH 7.6, 150 mM NaCl, 1% NP-40, 0.1% SDS) containing protease and phosphatase inhibitors (Roche). Cell lysates were separated by SDS–PAGE, transferred to nitrocellulose and blotted for the indicated proteins using commercial antibodies (all were from Cell Signaling Technology). For measuring cleaved poly ADP ribose polymerase (PARP) and cleaved caspase 3, cells were treated identically with the following exceptions: cells were incubated with inhibitors for 72 h, causing a number of cells to become non-adherent. Both adherent and non-adherent cells were combined before cell lysis and immunoblot analysis.

IC_{50} value measurements. A recombinant glutathione *S*-transferase (GST)–RET kinase domain fusion (Invitrogen) was diluted in a mix containing phosphoacceptor peptide (132 µM final concentration; sequence: EAIYAAPFKKK), buffer (10 mM $MgCl_2$, 10 mM HEPES 7.2, 40 ng BSA) and varying concentrations of inhibitor. Reactions were initiated by addition of 100 µM cold ATP supplemented with 5 µCi γ^{32}P-ATP. After 15 min at room temperature, 2 µl of the reactions (out of 25 µl total volume) were spotted onto P81 phosphocellulose paper (Whatman). Blots were washed at least 5 times over 1 h in 1% (v/v) phosphoric acid and blots were then dried and transferred radioactive counts were measured by phosphorimaging using a Typhoon Scanner (Molecular Dynamics). Quantification was conducted with ImageQuant software and titration data were fit to a sigmoidal dose response to derive IC_{50} values in the Prism software package. Inhibitors were diluted threefold over a final concentration range of 0.0005–100 µM to derive dose–response curves. Experiments were completed three times to derive mean and standard error measurements. IC_{50} values for KDR, SRC, ABL, C-RAF, mTOR, EGFR and AKT1 were determined similarly with the following exceptions. Dephosphorylated casein was used as a substrate for mTOR (Invitrogen). Inactive MEK1 (Millipore) was used as a substrate for C-RAF (Millipore). Poly Glu-Tyr was used as the substrate for KDR (Invitrogen), EGFR (Invitrogen), SRC (purified as previously described[24]) and ABL (purified as previously described[24]). Myelin basic protein was used as a substrate for AKT1 (Invitrogen).

Kinase-inhibitor profiling. AD36, AD57, AD58, AD80 and AD81 were assayed by Invitrogen to derive percentage inhibition of kinase activity. All compounds were screened at 1 µM and values are shown in Supplementary Tables 1–3. Detailed procedures for kinase reactions, ATP concentrations used and Z'-LYTE or Adapta assay formats are described in the SelectScreen Customer Protocol (http://www.invitrogen.com/kinaseprofiling). Kinase-inhibition data for 1 µM inhibitor of each of the clinical and tool compounds staurosporine, sunitinib, dasatinib, pyrazolopyrimidine 2, gefitinib, imatinib, SB202190, erlotinib and BIRB796 were obtained from Invitrogen (http://tools.invitrogen.com/content/sfs/brochures/Activity_Assay_Kinase_Selectivity_Data.pdf). Clustering and visualization were performed with Cluster 3.0 and Java TreeView (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm).

Chemical synthesis. The AD compounds were synthesized using a seven-step chemical synthesis that is described in the Supplementary Methods. Final products were characterized by nuclear magnetic resonance spectroscopy and liquid-chromatography–mass spectrometry.

34. Brachmann, C. B. *et al.* The *Drosophila* Bcl-2 family member dBorg-1 functions in the apoptotic response to UV-irradiation. *Curr. Biol.* **10,** 547–550 (2000).

doi:10.1038/nature11152

Fluoride ion encapsulation by Mg^{2+} ions and phosphates in a fluoride riboswitch

Aiming Ren[1], Kanagalaghatta R. Rajashankar[2,3] & Dinshaw J. Patel[1]

Significant advances in our understanding of RNA architecture, folding and recognition have emerged from structure–function studies on riboswitches, non-coding RNAs whose sensing domains bind small ligands and whose adjacent expression platforms contain RNA elements involved in the control of gene regulation. We now report on the ligand-bound structure of the *Thermotoga petrophila* fluoride riboswitch, which adopts a higher-order RNA architecture stabilized by pseudoknot and long-range reversed Watson–Crick and Hoogsteen A•U pair formation. The bound fluoride ion is encapsulated within the junctional architecture, anchored in place through direct coordination to three Mg^{2+} ions, which in turn are octahedrally coordinated to water molecules and five inwardly pointing backbone phosphates. Our structure of the fluoride riboswitch in the bound state shows how RNA can form a binding pocket selective for fluoride, while discriminating against larger halide ions. The *T. petrophila* fluoride riboswitch probably functions in gene regulation through a transcription termination mechanism.

The field of RNA structure, folding and recognition has been propelled forward by the discovery of metabolite-sensing bacterial non-coding RNA elements, termed riboswitches, which have been shown to affect RNA-mediated gene regulation[1,2]. Most riboswitches are positioned within 5′-untranslated regions of genes associated with transport and metabolism of their cognate metabolites. Riboswitches interconvert between metabolite-free and metabolite-bound conformations, depending on metabolite concentration, with the sensing domain involved in metabolite recognition, whereas the adjacent expression platform contains RNA elements that control translational initiation, transcription termination or ribozyme-mediated cleavage. So far, structure–function studies have been undertaken on riboswitches that target purines and their analogues, amino acids, coenzymes and phosphoamino sugars, as well as Mg^{2+} ions[3,4]. Structural studies on the compact ligand-bound sensing domains of the above riboswitches have elucidated the structural principles underlying RNA folding topology and recognition, whereby endogenous RNA molecules can fold and form pockets that recognize small metabolites and discriminate against closely related analogs[5,6].

Of particular interest is how RNA as a negatively charged polyphosphate can bind ligands that are also negatively charged. The earliest studies to address this question focused on the cofactor thiamine pyrophosphate (TPP) riboswitch[2], where structural insights established that a pair of hydrated Mg^{2+} ions mediated interactions between the diphosphates of TPP and guanine base edges (rather than backbone phosphates) of the RNA[7,8]. This structural information was critical for guiding subsequent studies on ligand-induced folding of the TPP riboswitch[9,10]. Similar structural principles involving a bridging hydrated Mg^{2+} ion were also used for recognition of the monophosphate of flavin mononucleotide (FMN) by its riboswitch[11,12].

Recently, a riboswitch associated with *crcB* motif non-coding RNAs from *Pseudomonas syringae* has been identified that targets fluoride ion with a K_d of approximately 60 μM and discriminates against other halogen ions[13]. This riboswitch is common to bacterial and archaeal species and was found to activate the expression of genes that encode putative fluoride transporters. Given the small size and negative charge of the fluoride ion, it seems remarkable that RNA can form a small enough pocket to target it and discriminate against larger halide ions.

Structure of fluoride-bound riboswitch

The conserved secondary fold of *crcB* RNA motif fluoride-sensing riboswitches was deduced following sequence conservation and covariational analysis among bacterial and archaeal species, as well as in-line probing and mutational studies[13]. The resulting analysis identified two helical stems connected by a large asymmetric internal loop, with the overhang at the 5′-end capable of adopting pseudoknot-like higher-order interactions.

We used isothermal titration calorimetry (ITC) to establish that the 52-mer RNA sequence corresponding to the sensing domain of the *T. petrophila* fluoride riboswitch (Fig. 1a) bound fluoride ion (on addition of KF) with a K_d of 135 ± 9 μM under 5 mM Mg^{2+} conditions (Fig. 1b). The stoichiometry of binding approaches 1:1 ($N = 0.87$), with estimated thermodynamic parameters of $\Delta H = -2.5 \pm 0.1\ \mathrm{kcal\,mol^{-1}}$ and $\Delta S = 9.3\ \mathrm{cal\,mol^{-1}\,K^{-1}}$. Similar binding parameters were observed under 1 mM Mg^{2+} conditions, but no binding was observed in the absence of Mg^{2+} ion (Supplementary Fig. 1a, b, respectively).

The sensing domain of the *T. petrophila* fluoride riboswitch (Fig. 1a) yielded crystals in the presence of fluoride ion that diffracted to 2.3 Å resolution. We solved the structure of this fluoride riboswitch by co-crystallizing it with $Ir(NH_3)_6^{3+}$, and capitalizing on the anomalous properties of iridium to solve the phase problem. (Supplementary Fig. 2). The three-dimensional structure in the bound state is shown in Fig. 1c with different elements colour-coded as shown in Fig. 1a. The most striking feature is that the bound fluoride ion (red ball, Fig. 1c), positioned within the centre of the riboswitch fold, is surrounded and directly coordinated by three metal ions (cyan balls, Fig. 1c; metal–fluoride distances of 1.8-2.0 Å). A close-up stereo view of the ligand-binding pocket in the same perspective as in Fig. 1c, with the emphasis on the fluoride ion (in red), three coordinating metal ions (in cyan) and five inwardly pointing backbone phosphates (non-bridging oxygens in pink and phosphorus atoms in yellow) is shown in stereo in Fig. 1d.

[1]Structural Biology Program, Memorial Sloan-Kettering Center, New York, New York 10065, USA. [2]NE-CAT, Advanced Photon Source, Argonne National Laboratory, Chicago, Illinois 60439, USA. [3]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA.

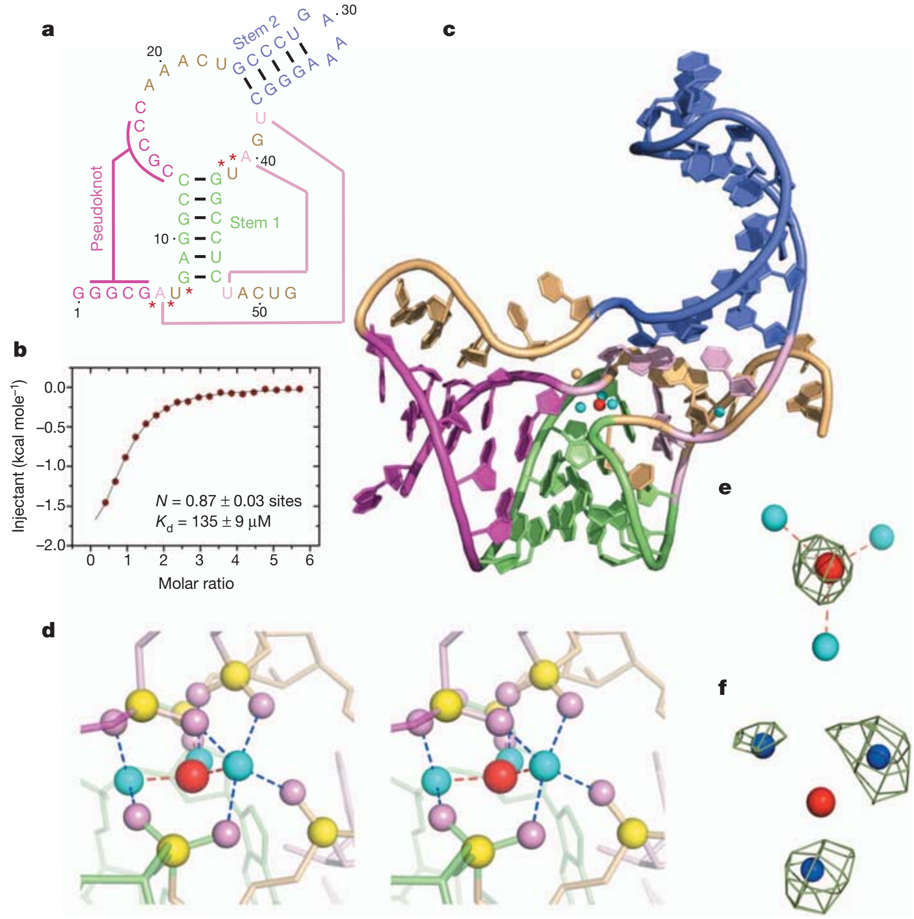


Figure 1 | Sequence, binding affinity and structure of the sensing domain of the *T. petrophila* fluoride riboswitch in the ligand-bound state. a, Secondary structure schematic of the 52-nucleotide sensing domain of the *T. petrophila* fluoride riboswitch used in this study. The colour coding highlights regular stem segments (blue and green) and long-range pseudoknot pairing interactions (magenta and pink; connected by lines), and bases not involved in pairing interactions (gold). Five phosphates, whose non-bridging oxygens form a shell around the three hydrated metal ions, with the metal ions in turn coordinated to the fluoride ion, are labelled by red stars. **b**, ITC binding curve for binding of KF to the *T. petrophila* fluoride riboswitch in 5 mM Mg^{2+}-containing buffer (for experimental conditions, see Methods). **c**, Crystal structure (at 2.3 Å) of the fluoride riboswitch in the ligand-bound state. The colour coding of RNA segments follows that shown in panel **a**, with the fluoride ion shown by a red ball and directly coordinated metal ions shown by cyan balls. **d**, A close-up stereo view of the ligand-binding pocket in the same perspective as in panel **c**, with the emphasis on the fluoride ion, three coordinating metal ions and five inwardly pointing backbone phosphates. **e**, $F_o - F_c$ omit electron density map contoured a 4σ level calculated following deletion of the fluoride ion. **f**, Anomalous density at 9σ level at positions of metal ions 1, 2 and 3 for crystals of the complex soaked in 50 mM Mn^{2+}-containing solution.

Fluoride coordinated to three Mg^{2+} ions

We have identified the fluoride ion binding site in the $F_o - F_c$ omit map for fluoride ion (Fig. 1e) and assign all three metal ions coordinated to the fluoride ion, to Mg^{2+} ions, based on observation of anomalous signals following soaking of the crystals in 50 mM Mn^{2+} ion-containing solution (Fig. 1f and Supplementary Fig. 3). The additional metal ion (in orange, Fig. 1c) positioned close by but not coordinated to the fluoride ion (metal–fluoride distance of 4.0 Å), is assigned to a K^+ cation (buffer contained K-acetate), on the basis of the observed anomalous signal at its position following soaking of crystals in Cs^+ ion-containing (Supplementary Fig. 4a) and Tl^+ ion-containing (Supplementary Fig. 4b) solutions.

Riboswitch adopts pseudoknot scaffold

The higher-order fold of the fluoride-bound riboswitch is stabilized by pseudoknot formation involving residues G2-G3-C4-G5 of the 5′-overhang segment and residues C14-G15-C16-C17 of the internal loop, thereby forming a regular duplex composed of four stacked Watson–Crick G-C base pairs (in magenta, Fig. 2a). This validates the prediction of pseudoknot formation amongst the same residues[13], though our crystal structure shows stabilization by four rather than the postulated five base pairs anticipated in solution. In this regard, 5′-terminal G1 and 3′-terminal U51 and G52 are involved in crystal packing contacts in our structure of the fluoride riboswitch (Supplementary Fig. 5). None of the residues in the C18 to U23 segment of the internal loop are involved in pairing interactions, but rather C18 stacks on terminal Watson–Crick G2-C17 pair, with a sharp turn at the C18-A19 step, followed by continuous stacking within the A19-A20-A21 and C22-U23 steps (Fig. 2a). The junctional architecture in the vicinity of the fluoride-binding site is additionally stabilized by formation of long-range single-base pseudoknot-like pairing between A6 and U38 (predicted previously[13]) and between A40 and U48 (Fig. 2b). Both form non-canonical pairs, with A6•U38 pairing through a reversed Watson–Crick alignment (Fig. 2c) and A40•U48 pairing through a reversed Hoogsteen alignment (Fig. 2d). Note that unpaired U7 and G39 are mutually interdigitated (Fig. 2b) and contribute to the formation of the junctional architecture. The tracing of these RNA segments in the $2F_o - F_c$ electron density maps of the fluoride riboswitch are shown in Supplementary Fig. 6a, b.

Coordination of Mg^{2+} ions by phosphates

The negatively charged fluoride ion (in red) is directly coordinated to three Mg^{2+} ions (in cyan) labelled M1, M2 and M3 (stereo view in Fig. 3a), with the fluoride ion positioned somewhat out of the plane

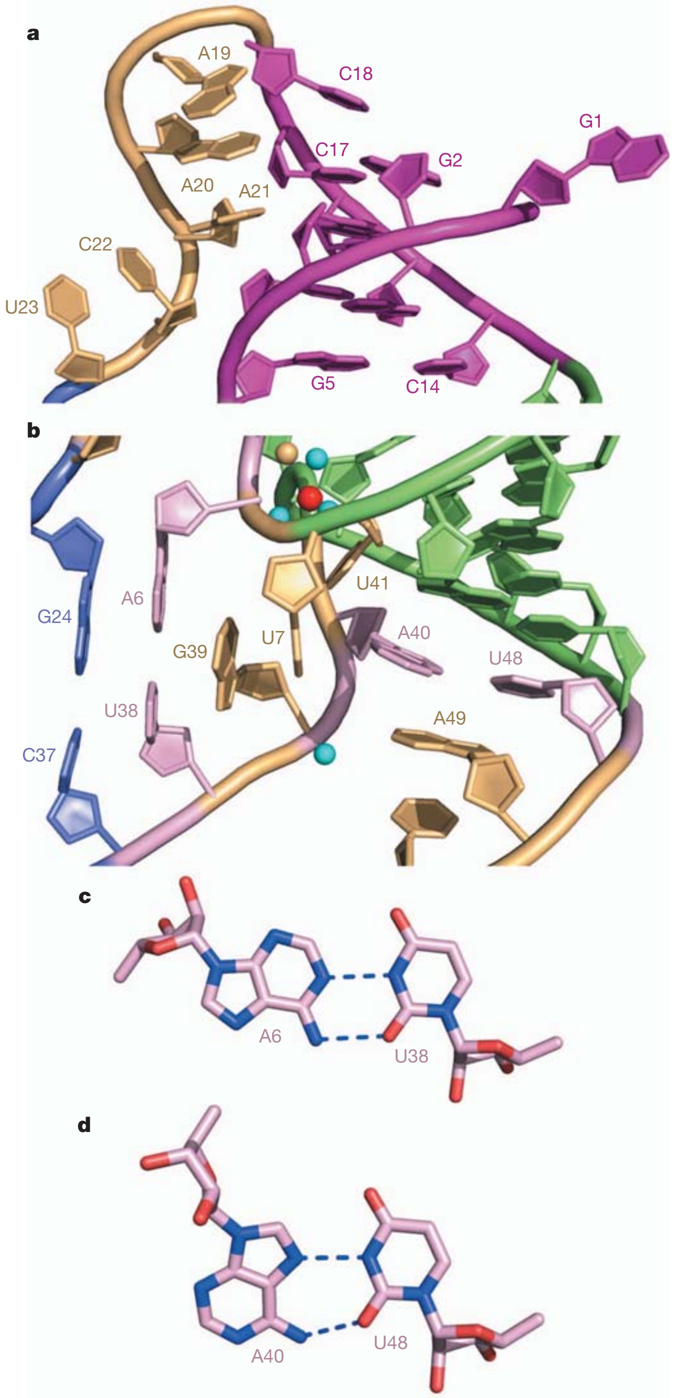


Figure 2 | Details of long-range interactions within the structure of the *T. petrophila* fluoride riboswitch in the ligand-bound state. **a**, An expanded region of the structure highlighting pseudoknot formation involving residues G2-G3-C4-G5 of the 5′-overhang segment and residues C14-G15-C16-C17 of the large internal loop (in magenta), as well as continuous stacking within the A19-A20-A21 and C22-U23 steps. **b**, An expanded region of the structure highlighting long-range single-base pseudoknot-like pairing between A6 and U38, and between A40 and U48. Note the mutual interdigitation between unpaired U7 and G39 that contribute to formation of the junctional architecture. **c**, Long-range reversed Watson–Crick A6•U38 pair formation. **d**, Long-range reversed Hoogsteen A40•U48 pair formation.

(0.49 Å) formed by the three Mg^{2+} ions (Fig. 3b). Stereo views of the corresponding $F_o - F_c$ omit electron density maps for fluoride ion, three Mg^{2+} ions and bound water molecules are shown in Supplementary Fig. 7a (contoured at 3σ) and b (contoured at 7σ). The three Mg^{2+} ions in turn are coordinated by water molecules and five inwardly pointing non-bridging phosphate oxygens (in pink; phosphorus in yellow), with three of these phosphates involved in bidentate coordination (pA6, pU7 and pG42), and two others involved in monodentate coordination (pG8 and pU41) (Fig. 3a). Stereo views of the corresponding $F_o - F_c$ omit electron density maps for the five nucleotides with inwardly pointing phosphates are shown in Supplementary Fig. 7c (contoured at 3σ) and d (contoured at 5σ).

In essence, our structure identifies a unique solution for how a negatively charged RNA scaffold can target a negatively charged fluoride ligand. The fluoride ion is surrounded by and coordinated to an inner shell of three Mg^{2+} ions, which in turn are surrounded and coordinated to an outer shell of five backbone phosphates and water molecules. Notably, the five participating phosphates are located within two distinct segments of the sequence, with three of them residing within the 5′-overhang G5pA6pU7pG8 segment, while two other reside within the A40pU41pG42 internal loop segment, as labelled in Fig. 1a by red asterisks. Because of the unique orientation of these five inwardly directing backbone phosphates, 8 of the 14 internal loop residues are stacked but not involved in hydrogen-bond pairing, and differ from large internal loops of other structurally characterized riboswitches, where ligand binding results in a compaction mediated by maximal hydrogen-bond pairing and stacking of loop residues[14–19].

Metal ions and their coordinated water molecules were identified on the basis of $2F_o - F_c$ and $F_o - F_c$ maps guided by the coordination geometries. The positioning of the fluoride ion, the non-bridging phosphate oxygens and the water molecules that constitute the three Mg^{2+} ion cluster that is coordinated to the fluoride ion is shown in stereo in Fig. 3c, with individual Mg^{2+} ions adopting the anticipated octahedral-like alignments as shown in Fig. 3d. Mg^{2+} coordination within the square-planar arrangements are shown by dashed blue lines, whereas coordination in apical positions are shown by dashed red lines, together with distances listed in Å in Fig. 3d. Notably, the fluoride riboswitch does not bind chloride ion (added KCl) in the presence of 20 mM Mg^{2+} as monitored by ITC (Supplementary Fig. 1c). In addition, it does not bind fluoride ion (added KF) in the presence of 20 mM Li^{+} ions (Supplementary Fig. 1d), in contrast to the observed binding under Mg^{2+} ion conditions (Fig. 1b).

Fluoride–metal ion coordination in proteins

Of additional note, the fluoride ion adopts an apical position in the octahedral coordination geometries for all three Mg^{2+} ions (Fig. 3d). Mg^{2+} ion M1 is coordinated by four non-bridging phosphate oxygens aligned in a square planar arrangement, whereas Mg^{2+} ions M2 and M3 are coordinated by two non-bridging phosphate oxygens (Fig. 3d). Additional support for our model of the coordination geometry of three Mg^{2+} ions around fluoride in the fluoride riboswitch comes from the highly similar coordination geometry observed in the 1.9 Å structure of fluoride-inhibited pyrophosphatase (crystals grown from 1 mM $MnCl_2$ and 5 mM NaF-containing buffer)[20]. Here, a fluoride ion is coordinated by a similar three-metal ion arrangement in the pyrophosphatase system (Supplementary Fig. 8). In the fluoride-inhibited pyrophosphatase structure, two Mn^{2+} ions and one Na^{+} ion are coordinated by oxygens atoms from one pyrophosphate (POP) molecule and from four aspartate carboxylate groups, and water molecules (Supplementary Fig. 8a).

We are aware that fluoride anion is a potent hydrogen bond acceptor[13,21], but currently have no evidence in support of direct F^{-}–H hydrogen bonding in the structure of the fluoride riboswitch, nor was such hydrogen bonding reported for the fluoride-inhibited pyrophosphatase system[20].

Previous studies of metals in RNA folding, stability and catalysis have focused on monovalent and divalent cations[22], including Mg^{2+} clusters bridging closely positioned phosphates in 5S RNA[23], P4-P5-P6 fragment of group I introns[24] and Mg^{2+}-sensing riboswitches[25,26]. Our current contribution, based on the discovery of a fluoride riboswitch[13],

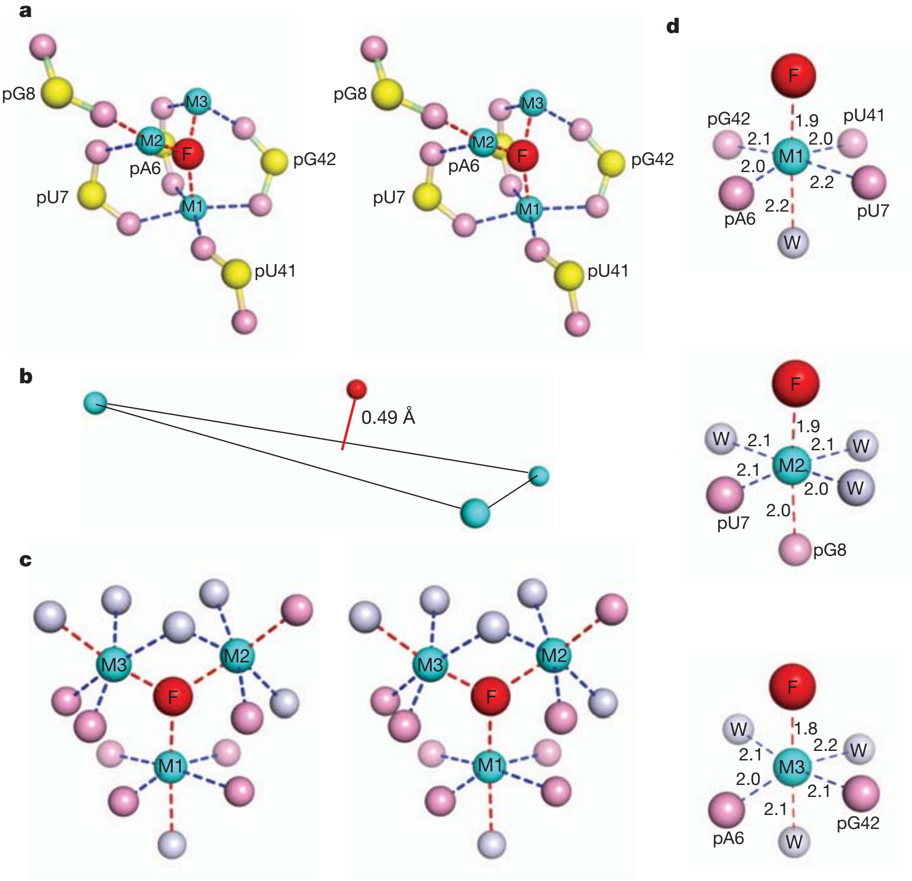


Figure 3 | Details of the fluoride ion binding site in the *T. petrophila* fluoride riboswitch in the ligand-bound state. a, An expanded stereo view highlighting the non-bridging phosphate oxygens directly coordinated to the three Mg^{2+} ions labelled M1, M2 and M3, that are in turn directly coordinated to the fluoride ion. In terms of nomenclature, the phosphate-labelled pA6 corresponds to the G5pA6 step. Three of the phosphates are involved in bidentate coordination with the metals, whereas two are involved in monodentate coordination. **b**, The fluoride ion is positioned 0.49 Å above the plane formed by the three metal ions. **c**, A stereo view of the coordination geometries of fluoride ion with surrounding Mg^{2+} ions, and in turn of Mg^{2+} with surrounding non-bridging phosphate oxygens and water molecules. Metal coordination within square-planar positions are shown by dashed blue lines, whereas coordination in apical positions are shown by dashed red lines in panels **a**, **c** and **d**. **d**, These panels show the octahedral coordination geometries around Mg^{2+} ions M1, M2 and M3, together with coordination distances in Å.

has defined how a RNA scaffold combines an inner shell of metal ions and an outer shell of phosphates to completely encapsulate a fluoride ion.

Folding and mechanism of action

We have recorded 900 MHz imino proton NMR spectra of the fluoride riboswitch in the free and fluoride bound states, which establish a conformational transition (chemical shift changes) and compaction (additional peaks) through higher-order structure generation on complex formation, with slow exchange between the free and bound forms (Supplementary Fig. 9). We have not been able to crystallize the ligand-free form of the fluoride riboswitch in the presence of Mg^{2+} ions. In the absence of this structure, we can only speculate on the folding energy landscape of the fluoride riboswitch as to how it dynamically assembles to recognize its ligand. This could occur through either hierarchical or simultaneous positioning of Mg^{2+} ions, RNA functional groups and ligand for recognition. This issue of folding energy landscape, as well as energetic costs associated with transfer of the fluoride ion from free to the riboswitch-bound state, could perhaps be addressed in the future by comprehensive single-molecule studies.

Previous functional studies implied that the *Bacillus cereus* and *Pseudomonas syringae* fluoride-responsive riboswitches control gene expression by regulating transcription termination and translational initiation, respectively[13]. To deduce mechanistic insights into gene regulation by the *T. petrophila* fluoride riboswitch from a structural perspective would require crystallographic characterization of the complete riboswitch containing both the sensing domain and expression platform in the ligand-free and bound states, which is beyond our current capabilities. Nevertheless, we note that that the sensing domain and adjacent expression platform of the *T. petrophila* fluoride riboswitch could interconvert between two conformations, one in which stem 1 forms in the presence of bound fluoride (ON state) and an alternate conformation at low fluoride concentrations where stem 1 is disrupted as a result of forming a rho-independent transcription terminator (OFF state) involving a stable stem loop ending in uracils at the 3′-end (Supplementary Fig. 10). Thus, it appears that both the *B. cereus* (supported by functional data)[13] and *T. petrophila* fluoride riboswitches are likely to adopt a transcription termination mechanism for gene regulation.

METHODS SUMMARY

Details of RNA preparation, purification and complex formation, as well as crystallization, X-ray data collection and refinement are described in detail in Methods.

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

Received 8 March; accepted 23 April 2012.
Published online 13 May 2012.

1. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111,** 747–756 (2002).

2. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419,** 952–956 (2002).
3. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29,** 11–17 (2004).
4. Winkler, W. C. & Breaker, R. R. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59,** 487–517 (2005).
5. Serganov, A. & Patel, D. J. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature Rev. Genet.* **8,** 776–790 (2007).
6. Montange, R. K. & Batey, R. T. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37,** 117–133 (2008).
7. Serganov, A. *et al.* Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441,** 1167–1171 (2006).
8. Thore, S., Leinungdut, M. & Ban, N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312,** 1208–1211 (2006).
9. Lang, K., Rieder, R. & Micura, R. Ligand-induced folding of the *thiM* TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. *Nucleic Acids Res.* **35,** 5370–5378 (2007).
10. Anthony, P. C., Perez, C. F., Garcia-Garcia, C. & Block, S. M. Folding energy landscape of the thiamine pyrophosphate riboswitch aptamer. *Proc. Natl Acad. Sci. USA* **109,** 1485–1489 (2012).
11. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111,** 747–756 (2002).
12. Serganov, A., Huang, L. & Patel, D. J. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *Nature* **458,** 233–237 (2009).
13. Baker, J. L. *et al.* Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335,** 233–235 (2012).
14. Mandal, M. *et al.* Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113,** 577–586 (2003).
15. Batey, R. T., Gilbert, S. D. & Montange, R. K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432,** 411–415 (2004).
16. Serganov, A. *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11,** 1729–1741 (2004).
17. Mandal, M. *et al.* A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* **306,** 275–279 (2004).
18. Huang, L., Serganov, A. & Patel, D. J. Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Mol. Cell* **40,** 774–786 (2010).
19. Butler, E. B., Xiong, Y., Wang, J. & Strobel, S. A. Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem. Biol.* **18,** 293–298 (2011).
20. Heikinheimo, P. *et al.* Toward a quantum-mechanical description of metal-assisted phosphoryl transfer in pyrophosphatase. *Proc. Natl Acad. Sci. USA* **98,** 3121–3126 (2001).
21. Auffinger, P., Hays, F. H., Westhof, E. & Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl Acad. Sci. USA* **101,** 16789–16794 (2004).
22. Hanna, R. & Doudna, J. A. Metal ions in ribozyme folding and catalysis. *Curr. Opin. Chem. Biol.* **4,** 166–170 (2000).
23. Correll, C. C., Freeborn, B., Moore, P. B. & Steitz, T. A. Metals, motifs and recognition in the crystal structure of a 5S RNA domain. *Cell* **91,** 705–712 (1997).
24. Cate, J. H., Hanna, R. L. & Doudna, J. A. A magnesium ion core at the heart of a ribozyme domain. *Nature Struct. Biol.* **4,** 553–558 (1997).
25. Cromie, M. J., Shi, Y., Latifi, T. & Groisman, E. A. An RNA sensor for intracellular Mg^{2+}. *Cell* **125,** 71–84 (2006).
26. Dann, C. E. III *et al.* Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130,** 878–892 (2007).

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

Acknowledgements We thank J. Goldberg and C. Lima for discussion and insights, Y. Liu for recording imino proton NMR spectra, and the personnel of the synchrotron beamlines 24-ID-C/E at the Advanced Photon Source, Argonne National Laboratory for their assistance. This research was funded by NIH grant GM34504 to D.J.P.

Author Contributions A.R. generated and purified RNA constructs, grew diffraction quality crystals, collected synchrotron data sets and solved the structure of the fluoride riboswitch in the bound state under the supervision of D.J.P.; K.R.R. assisted in crystallographic aspects of the structure determination, including finding unique solutions to the positioning of the Mg and fluoride ions. D.J.P. wrote the manuscript with the assistance of the other authors, all of whom discussed the results and commented on the manuscript.

Author Information Atomic coordinates of the structure of the fluoride riboswitch in the bound state have been deposited in the RCSB Protein Data Bank under the accession code 4ENC for the native structure and 4ENB for the $Ir(NH_3)_6{}^{3+}$-containing structure of the fluoride-bound riboswitch, as well as 3VRS, 4ENA and 4EN5 for crystals of the complex soaked in Mn^{2+}-, Cs^{+}- and Tl^{+}-containing solutions. 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 D.J.P. (pateld@mskcc.org).

METHODS

RNA preparation, purification and complex formation. The *crcB* motif of the *T. petrophila* fluoride riboswitch followed by the HDV ribozyme was transcribed *in vitro* using T7 RNA polymerase[27]. The transcribed RNA was purified by denaturing polyacrylamide gel electrophoresis (PAGE), followed by anion-exchange chromatography and ethanol precipitation. To generate the complex of the sensing domain of the *T. petrophila* fluoride riboswitch with fluoride ion, 10 mM KF was added to the buffer consisting of 100 mM potassium-acetate, pH 6.8 and 5 mM $MgCl_2$. After annealing at 60 °C for 10 min, the complex was purified by gel-filtration chromatography, before setting up crystallization trials.

Crystallization. Crystals of the fluoride anion-bound *T. petrophila* fluoride riboswitch grew at 20 °C over a period of 1 week using the sitting-drop vapour diffusion approach after mixing the complex at an equimolar ratio with the reservoir solution containing 0.1 M MES-sodium, pH 6.5, 35–40% (w/v) 2-methyl-2,4-pentanediol (MPD). To generate heavy-atom-bound crystals, 5 mM $Ir(NH_3)_6^{3+}$ was mixed with 0.5 mM complex before setting up crystallization trials. The $Ir(NH_3)_6^{3+}$ bound crystals grew from a solution containing 0.1 M sodium-cacodylate, pH 7.0, 20 mM spermine, 0.2 M strontium chloride and 20% MPD over a period of 4 days.

For heavy atom and cation soaking, crystals grown from 0.1 M sodium-cacodylate, pH 7.0, 20 mM spermine, 160 mM KCl, 50 mM $MgCl_2$ and 20% MPD were transferred into the mother solution of the crystals with $MgCl_2$ replaced with 50 mM $MnCl_2$, or with KCl replaced with 50 mM Tl-acetate or 100 mM CsCl for 24 h.

For cryoprotection, crystals were passed through several 5 μl drops of the stabilizing solution, which was the reservoir solution with MPD replaced by 30% MPD and 5% glycerol.

X-ray data collection and refinement. X-ray diffraction data were collected on flash-frozen crystals of the fluoride anion-bound *T. petrophila* fluoride riboswitch at NE-CAT beamlines at the Advanced Photon Source, Argonne National Laboratory and processed using the HKL2000 program (HKL Research). The structure (space group: $P2_12_12$) was determined using single wavelength anomalous dispersion (SAD) technique using anomalous signal from four iridium atoms with the HKL2MAP program[28] and PHENIX[29] suite. The RNA model was built in COOT[30] and refined in PHENIX[29] and REFMAC[31] using 2.3 Å native data set of the ligand-bound fluoride riboswitch. Metal ions and their coordinated water molecules were identified on the basis of $2F_o - F_c$ and $F_o - F_c$ maps guided by the coordination geometries. Fluoride ion was added to the model at the last stage based on the experimental and refined maps, coupled with electrostatic analysis. The X-ray statistics of the native and iridium-containing crystals are listed in Supplementary Table 1. Anomalous data sets of Mn^{2+}, Cs^+ and Tl^+ were collected at the wavelength of 1.7712 Å. The X-ray statistics of the complex crystals soaked in Mn^{2+}-, Cs^+- and Tl^+-containing solutions are listed in Supplementary Table 1. Mn^{2+}, Cs^+ and Tl^+ cations were positioned on the basis of the anomalous electron density maps (Supplementary Figs 3 and 4a, b). Cations were interpreted as Mg^{2+} or K^+ on the basis of the anomalous maps of their mimics, coordination geometry and distances[32].

NMR spectra. Imino proton NMR spectra (10 to 15 ppm) of the fluoride riboswitch in the free and bound state were recorded on a 900 MHz Bruker NMR spectrometer with cryoprobe using a jump-and-return pulse for H_2O solvent suppression. NMR spectra were recorded in buffer containing 50 mM K-acetate-d_3, 5 mM Mg-sulphate, 90%H_2O/10%D_2O, pH 6.8 at 25 °C.

Isothermal titration calorimetry. All experiments were performed on a MicroCal ITC200 calorimeter at 20 °C. Prior to titration, 0.3–0.5 mM RNA samples of the fluoride-bound riboswitch were dialysed overnight at 4 °C against experimental buffer containing 50 mM potassium acetate, pH 6.8, and 0 to 20 mM $MgSO_4$ or other cations to remove bound fluoride. RNAs were refolded by heating at 60 °C for 10 min and followed by cooling on ice. For measurements, KF dissolved in the dialysis buffer at 10 mM concentration was typically titrated into the RNA in the sample cell ($v = 207$ μl) by 20 serial injections of 2 μl each, with a 0.5 μl s^{-1} rate, 180 s intervals between injections, and a reference power of 6 μcal s^{-1}. The thermograms were integrated and analysed by using Origin 7.0 software (MicroCal).

27. Pikovskaya, O. *et al.* Preparation and crystallization of riboswitch-ligand complexes. *Methods Mol. Biol.* **540,** 115–128 (2009).

28. Pape, T. & Schneider, T. R. *HKL2MAP*: a graphical user interface for phasing with *SHELX* programs. *J. Appl. Cryst.* **37,** 843–844 (2004).

29. Adams, P. D. *et al. PHENIX*: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).

30. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).

31. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).

32. Feig, A. L. & Uhlenbeck, O. C. in *The RNA World* 2nd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 287–319 (Cold Spring Harbor Laboratory Press, 1999).

LETTER

doi:10.1038/nature11062

The age of the Milky Way inner halo

Jason S. Kalirai[1,2]

The Milky Way galaxy has several components, such as the bulge, disk and halo. Unravelling the assembly history of these stellar populations is often restricted because of difficulties in measuring accurate ages for low-mass, hydrogen-burning stars[1,2]. Unlike these progenitors, white dwarf stars[3], the 'cinders' of stellar evolution, are remarkably simple objects and their fundamental properties can be measured with little ambiguity[4,5]. Here I report observations of newly formed white dwarf stars in the halo of the Milky Way, and a separate analysis of archival data in the well studied 12.5-billion-year-old globular cluster Messier 4. I measure the mass distribution of the remnant stars and invert the stellar evolution process to develop a mathematical relation that links this final stellar mass to the mass of their immediate progenitors, and therefore to the age of the parent population. By applying this technique to a small sample of four nearby and kinematically confirmed halo white dwarf stars, I calculate the age of local field halo stars to be 11.4 ± 0.7 billion years. The oldest globular clusters formed 13.5 billion years ago. Future observations of newly formed white dwarf stars in the halo could be used to reduce the uncertainty, and to probe relative differences between the formation times of the youngest globular clusters and the inner halo.

Figure 1a illustrates the discovery of almost 2,000 white dwarfs in the nearest globular star cluster, Messier 4. The deep Hubble Space Telescope Advanced Camera for Surveys observations were recently obtained (Guest Observer Program 10146, Principal Investigator L. Bedin)[6], and analysed using new methods[7]. The stellar evolution process that produces these remnants runs in a predictable, continuous way like clockwork, because all of the Messier 4 stars formed at the same time, 12.5 ± 0.5 billion years ago[8]. The brightest objects at visual magnitudes of 22.5–23.5 represent the newly formed remnants of progenitor hydrogen-burning stars that have just exhausted their nuclear fuel. The mass of the progenitor stars can be calculated accurately from stellar evolution models because the age of the population is well measured[9]. For Messier 4, this mass is $M_{\rm initial} = 0.802^{+0.007}_{-0.011} M_{\rm Sun}$, where $M_{\rm Sun}$ is the mass of the Sun. The fainter stars on the Messier 4 white dwarf cooling sequence, at visual magnitudes of 28–29, are 'older white dwarfs' that evolved from more massive progenitors earlier in the star cluster's history.

The rich white-dwarf cooling sequence of the globular cluster Messier 4 offers a rare opportunity to anchor a new relation that links the final mass of stellar evolution to the age of the population. In Fig. 1b–g, I present Keck Telescope multi-object spectroscopy of a half-dozen newly formed white dwarfs in Messier 4. The composition of white dwarfs is simple: a carbon/oxygen core at high pressure surrounded by a helium mantle and a thin atmosphere of hydrogen[10]. Unlike A-dwarfs, the hydrogen-atom Balmer lines are strongly pressure broadened. I reproduced these observed profiles with the latest white-dwarf atmosphere models, which include updated Stark broadening calculations of the hydrogen atom[11]. The fundamental parameters of each star, including the temperature, gravity and mass, were measured through a simultaneous fit to both low- and high-order Balmer lines. These results indicate that the mass of white dwarfs forming today in Messier 4 is $M_{\rm final} = 0.529 \pm 0.012 M_{\rm Sun}$. This is in excellent agreement with both theoretical predictions of the masses of white dwarfs forming today in globular clusters[12], and with an independent (but indirect) measurement in the 12.5-billion-year-old cluster NGC 6752 (ref. 13).

The absolute calibration of both the initial and final stellar mass at a well-measured (old) age provided the necessary input to calculate the formation time of the Milky Way field halo. The only missing ingredient was knowledge of the mass distribution of white dwarfs that are forming today in the halo. Previous searches[14–16] for white dwarfs in the halo have successfully uncovered cool remnants with temperatures of less than 5,000 K. Such white dwarfs are difficult to date. The cooling rates of the stars depend on their masses, and the masses cannot be measured owing to a lack of spectral features at these temperatures. The total age of each star is the combined age of the progenitor lifetime and the (appreciable) white-dwarf cooling age.

The spectra of four nearby field white dwarfs is presented in Fig. 1h–k, along with a theoretical fit based on the same updated models used to analyse the Messier 4 remnants[11]. These four stars were carefully selected from 398 white dwarfs in the SPY survey (SN Ia Progenitor Survey), and are considered kinematic members of the Galactic halo on the basis of three-dimensional velocity measurements[17,18]. The temperatures of these stars confirm their nature as newly formed white dwarfs from progenitors that have just exhausted their hydrogen supply. Using the same method as described above for the Messier 4 white dwarfs, I measured the average mass of the four Milky Way halo white dwarfs to be $M_{\rm final} = 0.551 \pm 0.005 M_{\rm Sun}$.

The mass distribution of the six white dwarfs at the bright tip of the Messier 4 cooling sequence, and the four newly formed white dwarfs in the Galactic halo is shown in Fig. 2a. Through the uniform treatment of both populations, I found that the halo white dwarfs have a larger mass by approximately 4% ($0.02 M_{\rm Sun}$). I interpret this difference to reflect a small difference in the mass of the stellar core of the progenitor star that is currently leaving the hydrogen-burning stage in each population. For such low-mass hydrogen-burning stars, the post-hydrogen-burning evolutionary timescales are almost identical over small changes in mass, and recent studies have now convincingly demonstrated that progressively lower-mass hydrogen-burning stars evolve to form lower-mass white dwarfs[19–21].

To calibrate the measured difference in the mass of white dwarfs forming today in these two populations, I first constructed a new relation between the initial and final mass of stars. The relation is anchored on the Messier 4 measurement of $M_{\rm initial} = 0.802^{+0.007}_{-0.011} M_{\rm Sun}$ and $M_{\rm final} = 0.529 \pm 0.012 M_{\rm Sun}$. The $\Delta M = 0.02 M_{\rm Sun}$ difference in the core mass of the star implies an initial mass of the progenitor Milky Way halo stars of $0.825^{+0.009}_{-0.013} M_{\rm Sun}$. Next, using the same stellar evolution models that were applied[9] to the globular cluster Messier 4, I calculated the age of the stellar halo near the position of the Sun to be 11.4 ± 0.7 billion years (Fig. 2b). The uncertainty in this measurement derives from the spread in masses of the small sample of four halo white dwarfs, and can be improved considerably by increasing the number of spectroscopically measured remnants in future studies. The relation that links the mass of remnants forming today to the parent population's age is:

$$\log[{\rm Age\ (Gyr)}] = (\log[M_{\rm final}/M_{\rm Sun} + 0.270] - 0.201)/-0.272$$

This new relation is directly calibrated on the globular cluster age scale, as defined through a homogeneous imaging survey of 60 clusters with

[1]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [2]Center for Astrophysical Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA.

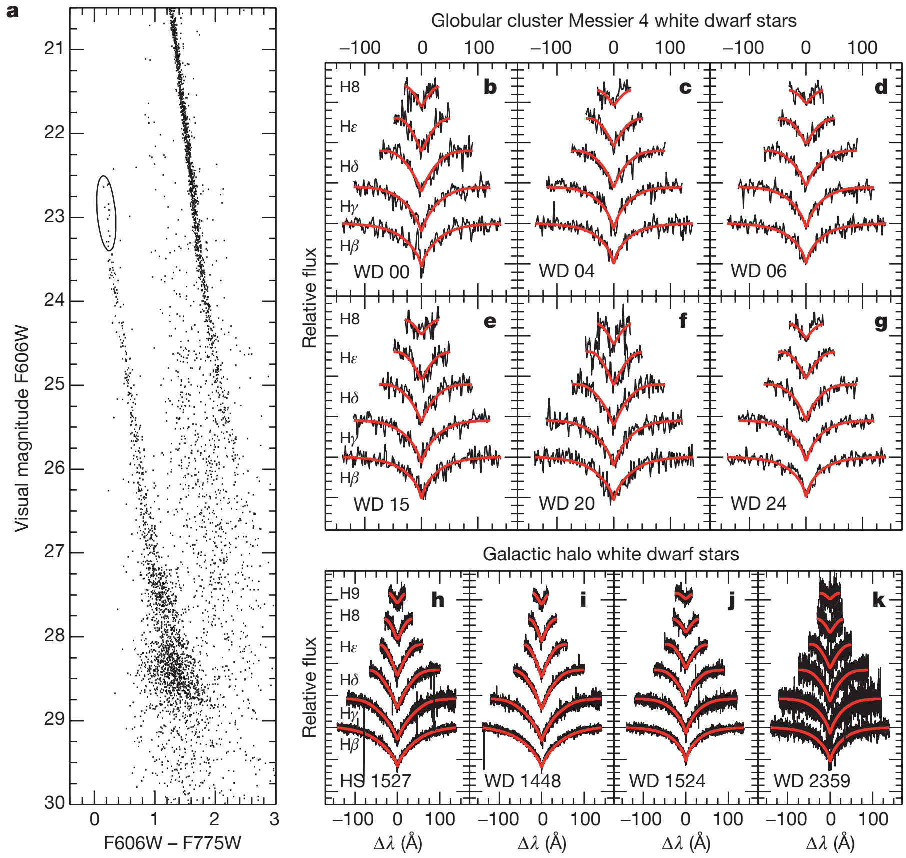


Figure 1 | Spectroscopic examination of white dwarfs in Messier 4 and the Milky Way halo. **a**, The stellar cinders of previous-generation hydrogen-burning main-sequence stars pile up along the white-dwarf cooling sequence of the nearby, 12.5-billion-year-old globular cluster Messier 4 (ref. 8). The brightest of these white dwarfs (within the ellipse) have been cooling for less than 100 million years, and are therefore the end products of stars that are just today evolving from the hydrogen-burning phase (that is, stars with $M_{\mathrm{initial}} = 0.802^{+0.007}_{-0.011}M_{\mathrm{Sun}}$ for Messier 4). **b**–**j**, Spectroscopic observations of the brightest Messier 4 white dwarfs (**b**–**g**) and the newly formed white dwarfs in the Milky Way halo from the SPY survey (**h**–**j**). Specifically, of the seven halo candidates in SPY[18], I accepted the three stars that are more than 3.5σ outliers from the thin and thick disk distributions (WD 1524, HS 1527 and WD 1448) as well as a fourth star that has a clear halo-like orbit (WD 2359). The white dwarf WD 0252 was rejected using these criteria, and I also note that its mass is $<0.40M_{\mathrm{Sun}}$ and so it could not have formed through a normal channel of stellar evolution (for example, it could be a helium-core white dwarf that was in a binary). The temperatures of the four stars were measured to be 14,000–20,000 K, and so their cooling ages are only 25–300 million years. Each panel illustrates all of the broad Balmer absorption lines in a single star that belongs to the respective populations. These pressure-broadened spectral features are reproduced using the latest white-dwarf atmosphere models (with updated Stark broadening physics) to reveal the temperatures, gravities and masses of the stars[11]. To avoid systematic errors from the two different resolutions, the model spectra were convolved to the resolution of the respective data sets (using a Gaussian function) before the fitting.

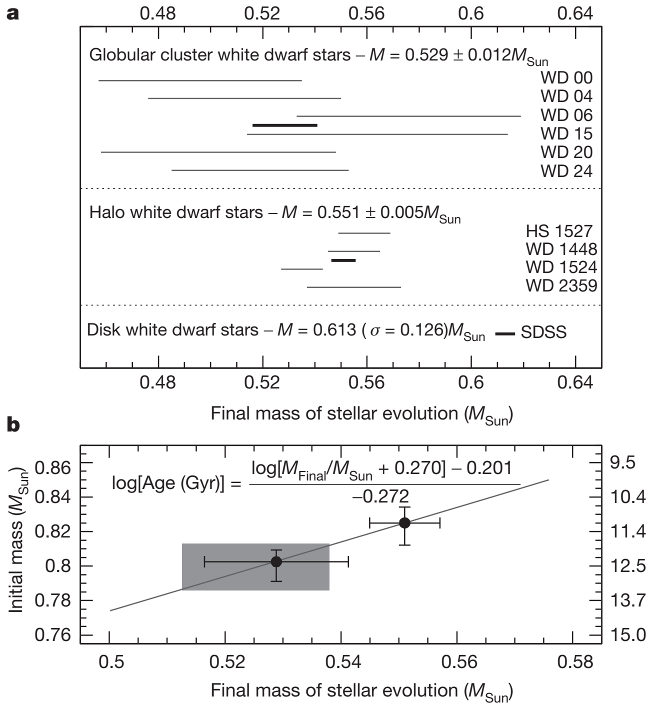


Figure 2 | The remnant mass and population age of the Milky Way halo. **a**, The masses of newly formed white dwarfs in Messier 4 (top) and the Milky Way halo (middle). The grey bars illustrate the individual mass measurements and the black bar represents the mean of the sample. The length of the bar indicates the uncertainty in the measurement. The average mass of the bright Messier 4 white dwarfs is lower than the halo white dwarfs by approximately $0.02M_{\mathrm{Sun}}$. **b**, The heavier white dwarfs in the Milky Way halo formed from heavier progenitors[10,20,21]. The well-measured age of Messier 4 provides an anchor with which to convert the mass difference to an age difference. For a measured mass of $0.551 \pm 0.005M_{\mathrm{Sun}}$ for the field halo white dwarfs, the age of the population is 11.4 ± 0.7 billion years, similar to that of the Messier 4 cluster (12.5 ± 0.5 billion years old), given the uncertainty. This age is 2 billion years younger than the oldest globular clusters, which formed 13.5 billion years ago on this scale (the thick grey bar illustrates the age dispersion of the Galactic globular clusters)[8,23]. The mass distribution of white dwarfs in the Galactic disk from high signal-to-noise Sloan Digital Sky Survey (SDSS) spectra, measured using the same techniques and models as described above, is $M_{\mathrm{final}} = 0.613M_{\mathrm{Sun}}$ ($1\sigma = 0.126M_{\mathrm{Sun}}$)[28]. The distribution also has a shallow tail to higher masses (not shown). The progenitor lifetimes of these stars are much shorter than the halo and globular clusters, confirming that the Milky Way disk formed the bulk of its stars well after the halo formed. The newly defined relation for calculating the age of the Galactic halo cannot be directly used to date the Galactic disk for two reasons. First, the disk has had an extended star-formation history and these white dwarfs are not all newly formed. Second, the progenitors of the disk white dwarfs had a metallicity very different from that of the halo and therefore the mass–age relation is different.

the Hubble Space Telescope[22]. The entire data set was subjected to a uniform set of stellar evolution models with updated physics[8,9,23]. The results indicate a clear age–metallicity gradient in the population, defined strongly by more than 50 of the clusters. Figure 2b shows that the Milky Way's most metal-poor clusters formed 13.5 billion years ago and the most metal-rich systems formed 12 billion years ago. Only nine of the clusters are younger than Messier 4. These results therefore suggest that the local Milky Way halo formed 2 billion years after the first globular clusters, approximately at the same time as the last clusters.

Observations of galaxies in the nearby Universe demonstrate that the process of galaxy assembly proceeds in a hierarchical framework. The stellar halo of the Milky Way represents the premier hunting ground in which to unravel the archaeology of when and how these processes occurred. Recent observations in the Milky Way suggest that the galaxy's halo may have two populations that are distinct in their abundances and kinematics[24,25]. The most recent smoothed particle hydrodynamics + N-body simulations of the formation of stellar halos are also finding evidence for a dual origin to the halo. In addition to an outer halo that is dominated by accreted stars from satellite disruption, half of the stars in the inner few tens of kiloparsecs of galaxy halos may be stars that formed *in situ*[26,27]. The local white dwarf population in this study is akin to this latter population, and the derived age measurement of 11.4 ± 0.7 billion years agrees very well with the prediction that 70% of the *in situ* population in galaxy-formation simulations formed by a redshift of 3 (that is, 11.5 billion years ago)[26].

Stars that are now in the outer, accreted halo of the Milky Way are predicted to have formed a few billion years before the *in situ* star formation of the inner halo[27]. This component is also probably more metal-poor than the $[\mathrm{Fe/H}] = -1.6$ inner halo[24]. Future surveys of newly formed white dwarf stars with the kinematic characteristics of the outer halo population can test this. For example, if the outer halo was accreted 13.5 billion years ago (consistent with the age of the oldest globular clusters), then I predict that the masses of white dwarf stars forming today should be $0.51M_{\mathrm{Sun}}$ or lower, depending on how metal-poor the population is.

Received 12 January; accepted 14 March 2012.
Published online 30 May 2012.

1. Soderblom, D. R. The ages of stars. *Annu. Rev. Astron. Astrophys.* **48,** 581–629 (2010).
2. Jofré, P. & Weiss, A. The age of the Milky Way halo stars from the Sloan Digital Sky Survey. *Astron. Astrophys.* **533,** A59 (2011).
3. Paczyński, B. Evolution of single stars. I. Stellar evolution from main sequence to white dwarf or carbon ignition. *Acta Astron.* **20,** 47–58 (1970).
4. Wegner, G. & Schulz, H. Spectroscopy of suspected peculiar DA white dwarfs. I—Equivalent widths and line profiles. *Astron. Astrophys.* **43** (Suppl.), 473–478 (1981).
5. Bergeron, P., Saffer, R. A. & Liebert, J. A spectroscopic determination of the mass distribution of DA white dwarfs. *Astrophys. J.* **394,** 228–247 (1992).
6. Bedin, L. R. *et al.* The end of the white dwarf cooling sequence in M4: an efficient approach. *Astrophys. J.* **697,** 965–979 (2009).
7. Kalirai, J. S. *et al.* A deep, wide-field, and panchromatic view of 47 Tuc and the SMC with HST: observations and data analysis methods. *Astron. J.* **143,** 11 (2012).
8. Dotter, A. *et al.* The ACS survey of galactic globular clusters. IX. Horizontal branch morphology and the second parameter phenomenon. *Astrophys. J.* **708,** 698–716 (2010).
9. Dotter, A. *et al.* The Dartmouth Stellar Evolution Database. *Astrophys. J.* **178** (Suppl.), 89–101 (2008).
10. Shapiro, S. L. & Teukolsky, S. A. *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects* 663 (Wiley-Interscience, 1983).
11. Tremblay, P.-E. & Bergeron, P. Spectroscopic analysis of DA white dwarfs: Stark broadening of hydrogen lines including nonideal effects. *Astrophys. J.* **696,** 1755–1770 (2009).
12. Renzini, A. & Fusi Pecci, F. Tests of evolutionary sequences using color-magnitude diagrams of globular clusters. *Annu. Rev. Astron. Astrophys.* **26,** 199–244 (1988).
13. Moehler, S. *et al.* Spectral types and masses of white dwarfs in globular clusters. *Astron. Astrophys.* **420,** 515–525 (2004).
14. Oppenheimer, B. R. *et al.* Direct detection of galactic halo dark matter. *Science* **292,** 698–702 (2001).
15. Gates, E. *et al.* Discovery of new ultracool white dwarfs in the Sloan Digital Sky Survey. *Astrophys. J.* **612,** L129–L132 (2004).
16. Kilic, M. *et al.* Visitors from the halo: 11 Gyr old white dwarfs in the solar neighborhood. *Astrophys. J.* **715,** L21–L25 (2010).
17. Napiwotzki, R. *et al.* Search for progenitors of supernovae type Ia with SPY. *Astron. Nachr.* **322,** 411–418 (2001).
18. Pauli, E.-M. *et al.* 3D kinematics of white dwarfs from the SPY project. II. *Astron. Astrophys.* **447,** 173–184 (2006).
19. Kalirai, J. S. *et al.* Stellar evolution in NGC 6791: mass loss on the red giant branch and the formation of low-mass white dwarfs. *Astrophys. J.* **671,** 748–760 (2007).
20. Kalirai, J. S. *et al.* The initial-final mass relation: direct constraints at the low-mass end. *Astrophys. J.* **676,** 594–609 (2008).
21. Kalirai, J. S. *et al.* The masses of population II white dwarfs. *Astrophys. J.* **705,** 408–425 (2009).
22. Sarajedini, A. *et al.* The ACS survey of galactic globular clusters. I. Overview and clusters without previous Hubble Space Telescope photometry. *Astron. J.* **133,** 1658–1672 (2007).
23. Dotter, A., Sarajedini, A. & Anderson, J. Globular clusters in the outer galactic halo: new Hubble Space Telescope/Advanced Camera for Surveys imaging of six globular clusters and the galactic globular cluster age-metallicity relation. *Astrophys. J.* **738,** 74 (2011).
24. Carollo, D. *et al.* Two stellar components in the halo of the Milky Way. *Nature* **450,** 1020–1025 (2007).
25. Beers, T. C. *et al.* The case for the dual halo of the Milky Way. *Astrophys. J.* **746,** 34 (2012).
26. Zolotov, A. *et al.* The dual origin of stellar halos. *Astrophys. J.* **702,** 1058–1067 (2009).
27. Font, A. S. *et al.* Cosmological simulations of the formation of the stellar haloes around disc galaxies. *Mon. Not. R. Astron. Soc.* **416,** 2802–2820 (2011).
28. Tremblay, P.-E., Bergeron, P. & Gianninas, A. An improved spectroscopic analysis of DA white dwarfs from the Sloan Digital Sky Survey Data Release 4. *Astrophys. J.* **730,** 128 (2011).

Acknowledgements The data presented in this paper were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W. M. Keck Foundation. I acknowledge the significant cultural role and reverence that the summit of Mauna Kea has within the indigenous Hawaiian community. I am fortunate to have had the opportunity to conduct observations from this mountain. This work was also based on observations obtained at the Paranal Observatory of the European Southern Observatory. I thank A. Dotter, P. Bergeron, and P.-E. Tremblay for discussions related to stellar evolution and ages. I also thank U. Heber for providing us reduced spectra from the SPY Survey.

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 the author (jkalirai@stsci.edu).

doi:10.1038/nature11203

Ultraviolet-radiation-induced methane emissions from meteorites and the Martian atmosphere

Frank Keppler[1], Ivan Vigano[1,2], Andy McLeod[3], Ulrich Ott[1,4], Marion Früchtl[2] & Thomas Röckmann[2]

Almost a decade after methane was first reported in the atmosphere of Mars[1,2] there is an intensive discussion about both the reliability of the observations[3,4]—particularly the suggested seasonal and latitudinal variations[5,6]—and the sources of methane on Mars. Given that the lifetime of methane in the Martian atmosphere is limited[1,6], a process on or below the planet's surface would need to be continuously producing methane. A biological source would provide support for the potential existence of life on Mars, whereas a chemical origin would imply that there are unexpected geological processes[7]. Methane release from carbonaceous meteorites associated with ablation during atmospheric entry is considered negligible[8]. Here we show that methane is produced in much larger quantities from the Murchison meteorite (a type CM2 carbonaceous chondrite) when exposed to ultraviolet radiation under conditions similar to those expected at the Martian surface. Meteorites containing several per cent of intact organic matter reach the Martian surface at high rates[9], and our experiments suggest that a significant fraction of the organic matter accessible to ultraviolet radiation is converted to methane. Ultraviolet-radiation-induced methane formation from meteorites could explain a substantial fraction of the most recently estimated atmospheric methane mixing ratios[3,4]. Stable hydrogen isotope analysis unambiguously confirms that the methane released from Murchison is of extraterrestrial origin. The stable carbon isotope composition, in contrast, is similar to that of terrestrial microbial origin; hence, measurements of this signature in future Mars missions may not enable an unambiguous identification of biogenic methane.

Global mean methane (CH_4) mixing ratios between 8 and 15 parts per billion by volume (p.p.b.v.) have been reported for Mars[1,2,4], and some studies have suggested[5,10] spatial and seasonal variations ranging from 0 to 70 p.p.b.v. The validity of some of the spectroscopic detections has recently been questioned[3] by suggesting an upper limit to CH_4 of the order of 3 p.p.b.v., whereas another recent analysis[4] of observations made in 2006 reveals a CH_4 mixing ratio of 10 p.p.b.v. around the deepest canyon (Valles Marineris) and 3 p.p.b.v. outside this region. Nevertheless, since the discovery of CH_4 in the Martian atmosphere there has been a continuing debate about its origin, and several possible sources[7,11–13], including the existence of microbes on Mars, have been proposed[1]. Eight possible mechanisms have been listed[12] that may be involved in its production, including subsurface clathrates and serpentinization of olivine, geothermal outgassing or biological methanogenic processes. However, all hypotheses have substantial shortcomings and/or are suggested to produce only a fraction of the estimated total flux of 200–300 t yr^{-1} (refs 1, 4, 13). For example, CH_4 release as a product of ablation and pyrolysis on atmospheric entry of meteoritic carbonaceous chondrites was estimated[8] to account for only a negligible fraction of Martian CH_4. However, much of the meteoritic material (mostly micrometeorites from interplanetary dust resembling carbonaceous chondrites in composition and containing a few per cent of organic matter) that is estimated to reach the Martian surface remains unmelted[9] and may become a potential source of CH_4.

On Earth, CH_4 can be formed from plant litter and from various organic molecules when irradiated in the laboratory with ultraviolet wavelengths (UVC, <280 nm; UVB, 280–320 nm; UVA, 320–400 nm)[14,15], but in the natural environment stratospheric ozone absorbs shorter solar ultraviolet wavelengths—and only a fraction of UVA and UVB wavelengths reaches the surface. On Mars, in contrast, due to the thin atmosphere and the low abundance of ozone, the situation is distinctly different[16]. Organic compounds such as amino acids and carboxylated molecules can be degraded by ultraviolet irradiation under simulated Martian atmospheric conditions[11,17–19], and hence it was suggested that photodegradation of organic matter on Mars may exceed the rate at which organic compounds are supplied by meteoritic infall.

We irradiated samples of Murchison—a large and well investigated carbonaceous chondrite of type CM2 that fell in Australia on 28 September 1969—with ultraviolet radiation comparable to that occurring under Martian surface conditions, and examined the production of CH_4 (see Methods and Supplementary Information). The sample was ground, resulting in a particle size distribution peaking around 100 μm, similar to the dominant size of unmelted micrometeorite infall[20]. Under 'control' conditions (22 °C, 1 bar N_2, no ultraviolet) we observed no CH_4 release. Release of CH_4 became detectable on commencing ultraviolet irradiation at an intensity close to that reported for the Martian surface at 20° latitude and assuming a dust optical density of 0.1 (ref. 16; see Supplementary Table 1). Emission started almost instantaneously and CH_4 levels increased to ~100 p.p.b.v. after 600–800 s and thereafter slightly decreased (Fig. 1 and Supplementary Fig. 4). The average emission rate during the

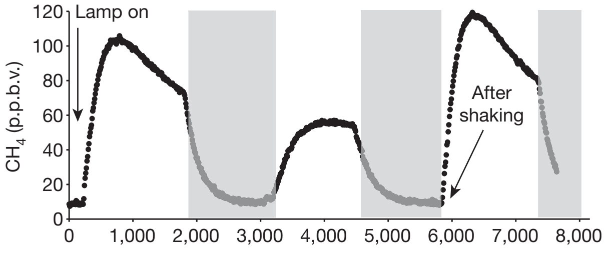


Figure 1 | Methane mixing ratios during ultraviolet irradiation of Murchison meteorite samples in a photochemical reactor. The reactor (Supplementary Fig. 1) was continuously flushed with N_2 gas (7–10 p.p.b.v. CH_4 background). Grey areas indicate when the reactor was bypassed in order to measure only the background flushing gas. Applied irradiances from the xenon-arc lamp were kept constant at ~36, 4.8 and 0.7 W m^{-2} for UVA, UVB and UVC wavelengths, respectively (Supplementary Table 1). The left arrow indicates when the lamp was switched on, the right arrow shows the time when the sample was redistributed.

[1]Department of Atmospheric Chemistry, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany. [2]Institute for Marine and Atmospheric Research Utrecht, Utrecht University, 3584CC Utrecht, The Netherlands. [3]School of GeoSciences, University of Edinburgh, Crew Building, The King's Buildings, West Mains Road, Edinburgh EH9 3JN, UK. [4]University of West Hungary, Savaria Campus, H-9700 Szombathely, Hungary.

first period of irradiation was 1,680 ± 158 ng CH_4 g^{-1} h^{-1}. These emissions were much larger than those found for plant pectin and also larger than those of a terrestrial Oxisol from Hawaii (see Supplementary Information). The extrapolated cumulative CH_4 released after one day of irradiation is calculated to be ~1.64 μg CH_4, and if we assume an ultraviolet penetration depth into minerals between 20 nm (ref. 21) and a more realistic 130 nm (see Supplementary Information), we obtain an 8.5–55% conversion of the accessible carbon to CH_4 (2.3–14.6 mg CH_4 g^{-1}). This is about five orders of magnitude higher than total CH_4 release observed for carbonaceous chondrites during freeze–thaw disaggregation[22]. After redistributing the irradiated ground sample by shaking the reactor (Fig. 1), the CH_4 mixing ratio increased again to values of ~120 p.p.b.v., slightly higher than at the start of the experiment, indicating that particle alignment to the ultraviolet source is important. Furthermore, we found a linear relationship between CH_4 formation and the applied ultraviolet irradiance (Supplementary Fig. 3), whereas the composition of the gas phase (N_2, O_2 or CO_2) did not affect emission rates. We also detected emissions of other volatile organic compounds (ethane, ethene and propane), but in much smaller quantities (see Supplementary Information).

For comparison, we also studied a solid fragment of Murchison of similar weight which presented a much smaller irradiated surface, and found CH_4 release on an area basis (μg CH_4 m^{-2} h^{-1}) that was similar (210 ± 57 μg CH_4 m^{-2} h^{-1}) to that for the ground sample (171 ± 16 μg CH_4 m^{-2} h^{-1}) (Supplementary Fig. 5).

The mineral composition of the meteoritic material may play an important role in CH_4 formation, because it may catalyse surface reactions of organic matter that produce CH_4. Formation of CH_4 via the photo-Kolbe reaction might occur in Martian regolith[11]. In addition, metal oxides and silicates have been described as catalysts for CH_4 formation when irradiated with light and in the presence of organic matter, CO and CO_2 (ref. 23). Murchison samples contain substantial amounts of haematite and serpentine-like minerals that could potentially serve as powerful catalysts. Interestingly, the Hawaiian soil that also produced considerable amounts of CH_4 also contained substantial amounts of iron oxides and silicates, and might be considered to have properties similar to those recently used as Martian soil analogues[11,18,24].

Surface temperatures on Mars vary tremendously, from −143 °C at the poles to +17 °C for the warmest soils of equatorial regions[25]. Therefore, we investigated CH_4 emissions between −190 and +20 °C (at 1,000 mbar N_2). Between temperatures of −190 °C and −40 °C, release of CH_4 was below our detection limit. Above −20 °C, CH_4 release was observed with emission rates increasing with temperature by about 18 ng CH_4 g^{-1} h^{-1} $°C^{-1}$, reaching up to ~800 ng CH_4 g^{-1} h^{-1} at 20 °C (red data points in Fig. 2). During subsequent cooling CH_4 emission rates decreased, but CH_4 formation was observable down to −60 °C (~200 ng CH_4 g^{-1} h^{-1}), indicating that once CH_4 generation is activated the reaction can proceed at lower temperatures. In a subsequent second heating cycle (starting at −60 °C) the temperature-dependent emission profile was repeated, with strongly enhanced emissions starting at −30 °C. We also examined the influence of pressure and found that for a lower pressure of 5–10 mbar, which closely simulates Martian atmospheric pressure conditions, CH_4 emission was already observed at −80 °C and was enhanced by a factor of ~2–3 compared to the experiments at 1,000 mbar.

The observed dependence of CH_4 emission on wavelength (see Supplementary Fig. 9) suggests that UV radiation at wavelengths between 295 and 305 nm is particularly important for generating CH_4 from meteoritic material, as previously reported for dry plant matter[14].

Additional information comes from the stable isotopic composition of CH_4 released from Murchison by ultraviolet treatment. The hydrogen isotope values clearly confirm that the source of the observed CH_4 is extraterrestrial (that is, from the meteorite), with δD values of H in CH_4 (δD-CH_4) ranging from +1,070‰ to −90‰, depending on the experiment (Fig. 3a). These values are unlike those of CH_4 sources on Earth (Fig. 3b). In contrast, the terrestrial Hawaiian soil sample (results labelled 'UV-soil' in Fig. 3b) produced δD-CH_4 values of −330‰ to −350‰, entirely within the range of known terrestrial CH_4 sources. The carbon signature, $\delta^{13}C$-CH_4 for CH_4 from Murchison, was in the range −57‰ to −32‰, values that are commonly found for terrestrial sources including biogenic ones. This can be compared with carbon isotope signatures found for various types of organic matter in Murchison, which are in the range −13.4‰ to +36‰ for soluble organic matter and −13‰ for the (dominant) insoluble organic matter (IOM), respectively[22,26]. For bulk organic matter (BOM) and single organic compounds such as amino acids, $\delta^{13}C$ values ranging from −2‰ to −21‰ and from −1‰ to +41‰ have been reported[27], while inorganic carbon (carbonate) shows average $\delta^{13}C$ values of +43‰ (refs 27, 28). Thus, it appears that ultraviolet irradiation causes substantial fractionation between the precursor carbon and emitted CH_4. This accords with previous observations of non-microbial CH_4 formation from terrestrial organic substances by ultraviolet treatment, where isotope fractionation between CH_4 and the bulk organic matter was between −25‰ and −40‰ (ref. 29). The terrestrial soil sample showed an evolution of $\delta^{13}C$-CH_4 values similar to the meteorite, but shifted by approximately 10–15‰ to more negative values (Supplementary Fig. 8), which nicely reflects the difference between $\delta^{13}C$ values for the (original) BOM of the two samples (terrestrial soil BOM: −26‰). Note that the isotopic composition of CH_4 released from Murchison under ultraviolet irradiation is distinct from extraterrestrial CH_4 captured in carbonaceous chondrites and liberated by freeze–thaw disaggregation[22] (Fig. 3). Thus, different mechanisms must have been involved in its generation.

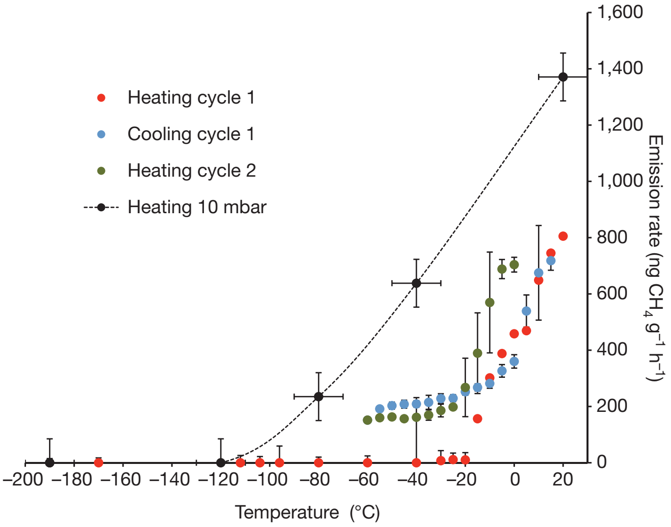


Figure 2 | Methane emission rates from ultraviolet irradiation of Murchison meteorite samples as a function of temperature and at different pressures. Data were obtained during heating cycle 1 (red), heating cycle 2 (green) and a cooling cycle (blue) at 1,000 mbar N_2, whereas black points show data obtained by simulating heating at the Martian atmospheric pressure (~10 mbar N_2). Applied irradiances from the xenon-arc lamp were kept constant at ~22, 2.9 and 0.44 W m^{-2} for UVA, UVB and UVC wavelengths, respectively (Supplementary Table 1). Data show mean ± s.d., $n = 3$.

Photodegradation of meteoritic organic matter and formation of CH_4 may be important on planets or planetary bodies that are exposed to high ultraviolet levels—such as Mars, or comets when their temperatures increase as they approach the Sun. Ultraviolet-induced organic matter degradation on Mars may prevent the accumulation of complex organic molecules on the surface of Martian soil[11,18,30]. However, this assumes that ultraviolet radiation can actually decompose a large fraction of the total carbon contained in the meteorites, which may be challenged because of the short penetration depth of ultraviolet light into minerals. Incoming particles with intact carbon

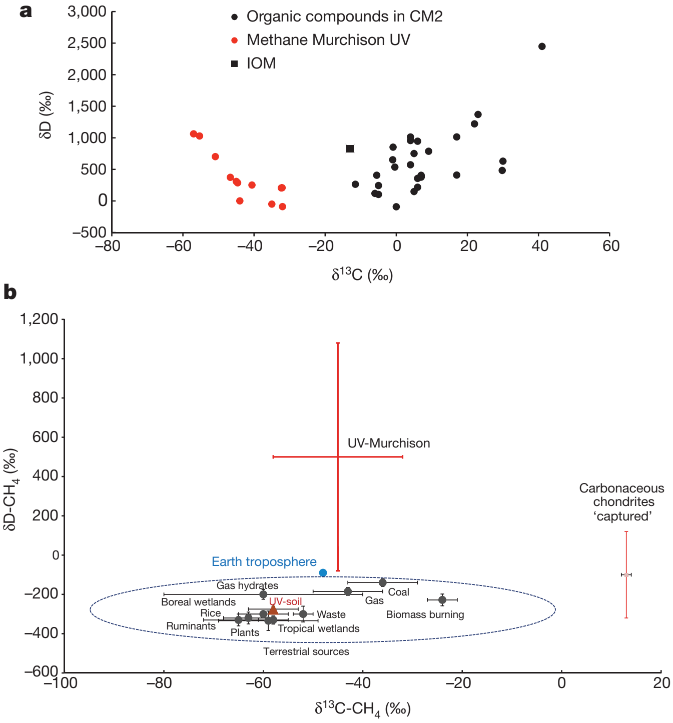


Figure 3 | Stable carbon and hydrogen isotope composition of CH_4 from terrestrial sources and from carbonaceous chondrites. **a**, Comparison of stable carbon and hydrogen isotope values of CH_4 released from Murchison samples by ultraviolet treatment ('Methane Murchison UV'; this study) with that of organic matter ('Organic compounds in CM2'; data from ref. 28) and insoluble organic matter ('IOM'; data from ref. 26) extracted from Murchison. **b**, Range of stable carbon and hydrogen isotope values of CH_4 emitted from ultraviolet irradiation of Murchison samples ('UV-Murchison'; this study) compared with terrestrial CH_4 sources (within dotted elliptical line; individual source values (filled circles) from ref. 29) and captured CH_4 from carbonaceous chondrites (released by freeze-thaw disaggregation; data (at right) from ref. 22). The brown triangle represents isotopic values from Hawaiian soil exposed to ultraviolet irradiation (this study). The filled blue circle reflects the current average isotope composition of Earth's troposphere, which plots outside the range of sources because of the kinetic isotope fractionation in the removal reaction with hydroxyl ($OH^•$).

are centred in size around 100 μm (ref. 20), and a simple estimate, assuming an ultraviolet penetration depth of 20 nm (ref. 24) to a more realistic 130 nm (see Supplementary Information) would thus suggest that only a small fraction of their carbon is actually exposed to ultraviolet radiation. However, there are many forms of physical and chemical weathering on Mars (including dust storms) that may remove protecting silicates and eventually bring most of the carbon close enough to the surface to be affected by ultraviolet radiation (see Supplementary Information). This assumption is supported by the observations of the Viking and Phoenix missions to Mars, which could not clearly identify whether organic matter is present in the surface layers[31].

Non-photochemical degradation processes could also destroy chemically labile organics, including oxidizing sources in Martian soils such as the recently detected perchlorates[32]. On the basis of our observations and of previous studies[11,17–19] showing very fast degradation kinetics of both simple and complex molecular organic matter during ultraviolet irradiation, we postulate that a significant fraction of organic matter (in the range of 10–50%) from extra-Martian sources will be rapidly converted to CH_4 (see Supplementary Information for details). We now compare this to previous assessments of CH_4 production on Mars.

An incoming meteorite flux of 2,700–59,000 t yr^{-1} (mostly micrometeorites with a composition resembling that of the carbonaceous chondrites[9], with an estimated organic matter content of 2%) contains 54–1,180 t C, which encompasses the estimated infall range[20] (240 t yr^{-1}) of unaltered meteoritic carbon on Mars. From this, 8–787 t CH_4 could be produced at an organic carbon conversion of 10–50%. Such estimates compare favourably with the suggested source strength of 126–270 t CH_4 (refs 1, 8) that is required to explain an equilibrium level of 8–10 p.p.b.v. of CH_4 on Mars[1,4,13] (if a CH_4 lifetime of 300–600 yr is considered). Note however that a recent review[3] argued against the detection of CH_4 on Mars above ~3 p.p.b.v., based on a re-evaluation of the spectroscopic data. Our present findings provide strong support for the view that CH_4 could actually be produced on Mars in quantities that would be consistent with such a global mixing ratio (at the lower p.p.b.v. level), not taking into account possible additional contributions from other suggested sources.

On a local scale, we would expect the highest CH_4 release rates in regions of Mars with high solar irradiance, particularly at the equator and extending to higher latitudes during northern and southern midsummer, but also following periods of surface disturbance (for example, dust storms). Assuming that Martian soil contains 2–40% meteoritic debris[9,33], that all of it is carbonaceous chondritic in composition, and that the fluxes from Murchison determined in the laboratory are representative for Mars under summer conditions, we calculate fluxes in the range 0.46–9.2 ng CH_4 m^{-2} s^{-1}. This range covers the flux values (≥ 1 ng CH_4 m^{-2} s^{-1}) required to explain the large seasonal variations in northern midsummer reported by Mumma *et al.*[5]. However, our estimated fluxes would only last for a few hours a day when temperature and ultraviolet irradiance are high, and only until the accessible organic matter of the meteorite is exhausted, whereas continuous release rates were calculated by Mumma *et al.*[5] for a 'filling time' of the plume of 0.5 Martian years (~344 Earth days). Even the upper limit to our annual estimate of 787 t CH_4 cannot explain the large amounts (19,000 t CH_4) estimated by Mumma *et al.*[5], unless much more meteoritic matter falls on Mars—for example, randomly in larger impact events[13]—or other previously suggested mechanisms[1,7,11–13,30] are active. Without such events, it is unlikely that the spatial and temporal variations of photodegradation of micrometeorites would create observable variations in the mixing ratio of CH_4.

Finally, we note that possible biological activity on Mars is an intriguing and controversial subject, for which the detection of CH_4 has been considered supporting evidence. Several missions will travel to Mars this decade to search for the potential presence of life. Application of stable isotope data has been suggested as a powerful tool in the identification of extraterrestrial life, because terrestrial biotic isotope signatures both from microbial and thermogenic (maturation of sedimentary kerogen) sources usually have relatively depleted $\delta^{13}C$-CH_4 values between −60‰ and −30‰ (Fig. 3b). Caution has to be exercised, however, in interpreting such a signature as proof of biological activity, as our results show that Murchison meteoritic matter under ultraviolet exposure produces $\delta^{13}C$-CH_4 values that are quite similar.

METHODS SUMMARY

Meteoritic and terrestrial samples. Samples of the Murchison meteorite, both solid fragments (~1–2 g) and ground matter with a peak in the particle size distribution at ~100 μm, were used for irradiation experiments. A fraction of the powder was sterilized (autoclaved) in order to test for possible microbial activity. In addition two terrestrial samples, plant pectin and soil (Oxisol) from Hawaii (Maui Island), were tested.

Reactor, radiation sources and experiments. Experiments used 200 mg of ground sample or solid fragments inside a leak-tight ultraviolet-transparent quartz reactor. Samples were placed in the reactor (see Supplementary Fig. 1) and continuously flushed with either nitrogen, synthetic air or common compressed air, depending on the experiment. A 1,000 W xenon-arc lamp was used to simulate the ultraviolet spectrum on the surface of Mars (Supplementary Fig. 2 and Supplementary Table 1). Methane mixing ratio was determined using cavity ring-down spectroscopy (CRDS) with a precision of 1 p.p.b.v.

Stable isotope measurements. Gas samples from the photochemical reactor were collected using evacuated glass vials or electropolished stainless steel canisters.

These samples were injected into a continuous flow-isotope ratio mass spectrometer (CF-IRMS) system for high precision analysis of δ^{13}C-CH_4 and δD values. δ^{13}C (‰) values are reported relative to Vienna PDB (VPDB) and defined by the equation $\delta^{13}C = (R_{sample}/R_{VPDB} - 1)$ with $R = {}^{13}C/{}^{12}C$. δ^{13}D values (‰) are reported relative to Vienna Standard Mean Oceanic Water (VSMOW) and defined by the equation $\delta D = (R_{sample}/R_{VSMOW} - 1)$ with $R = {}^{2}H/{}^{1}H$. For detailed information regarding experimental set up and full methods, see Supplementary Information.

Received 18 December 2011; accepted 27 April 2012.
Published online 30 May 2012.

1. Krasnopolsky, V. A., Maillard, J. P. & Owen, T. C. Detection of methane in the Martian atmosphere: evidence for life? *Icarus* **172,** 537–547 (2004).
2. Formisano, V., Atreya, S., Encrenaz, T., Ignatiev, N. & Giuranna, M. Detection of methane in the atmosphere of Mars. *Science* **306,** 1758–1761 (2004).
3. Zahnle, K., Freedman, R. S. & Catling, D. C. Is there methane on Mars? *Icarus* **212,** 493–503 (2011).
4. Krasnopolsky, V. A. Search for methane and upper limits to ethane and SO_2 on Mars. *Icarus* **217,** 144–152 (2012).
5. Mumma, M. J. *et al.* Strong release of methane on Mars in northern summer 2003. *Science* **323,** 1041–1045 (2009).
6. Lefèvre, F. & Forget, F. Observed variations of methane on Mars unexplained by known atmospheric chemistry and physics. *Nature* **460,** 720–723 (2009).
7. Etiope, G., Oehler, D. Z. & Allen, C. C. Methane emissions from Earth's degassing: implications for Mars. *Planet. Space Sci.* **59,** 182–195 (2011).
8. Court, R. W. & Sephton, M. A. Investigating the contribution of methane produced by ablating micrometeorites to the atmosphere of Mars. *Earth Planet. Sci. Lett.* **288,** 382–385 (2009).
9. Flynn, G. J. & McKay, D. S. An assessment of the meteoritic contribution to the Martian soil. *J. Geophys. Res. B* **95,** 14497–14509 (1990).
10. Geminale, A., Formisano, V. & Sindoni, G. Mapping methane in Martian atmosphere with PFS-MEX data. *Planet. Space Sci.* **59,** 137–148 (2011).
11. Shkrob, I. A., Chemerisov, S. D. & Marin, T. W. Photocatalytic decomposition of carboxylated molecules on light-exposed Martian regolith and its relation to methane production on Mars. *Astrobiology* **10,** 425–436 (2010).
12. Schuerger, A. C., Clausen, C. & Britt, D. Methane evolution from UV-irradiated spacecraft materials under simulated Martian conditions: implications for the Mars Science Laboratory (MSL) mission. *Icarus* **213,** 393–403 (2011).
13. Atreya, S. K., Mahaffy, P. R. & Wong, A. S. Methane and related trace species on Mars: origin, loss, implications for life, and habitability. *Planet. Space Sci.* **55,** 358–369 (2007).
14. Vigano, I. *et al.* Effect of UV radiation and temperature on the emission of methane from plant biomass and structural components. *Biogeosciences* **5,** 937–947 (2008).
15. McLeod, A. R. *et al.* Ultraviolet radiation drives methane emissions from terrestrial plant pectins. *New Phytol.* **180,** 124–132 (2008).
16. Cockell, C. S. *et al.* The ultraviolet environment of Mars: biological implications past, present, and future. *Icarus* **146,** 343–359 (2000).
17. ten Kate, I. L. *et al.* Amino acid photostability on the Martian surface. *Meteorit. Planet. Sci.* **40,** 1185–1193 (2005).
18. Stoker, C. R. & Bullock, M. A. Organic degradation under simulated Martian conditions. *J. Geophys. Res. E* **102,** 10881–10888 (1997).
19. Chun, S. F. S., Pang, K. D., Cutts, J. A. & Ajello, J. M. Photocatalytic oxidation of organic compounds on Mars. *Nature* **274,** 875–876 (1978).
20. Flynn, G. J. The delivery of organic matter from asteroids and comets to the early surface of Mars. *Earth Moon Planets* **72,** 469–474 (1996).
21. Wong, N. *et al.* in *Proceedings of the Twelfth ASCE Aerospace Division International Conference on Engineering, Science, Construction, and Operations in Challenging Environments and the Fourth NASA/ARO/ASCE Workshop on Granular Materials in Lunar and Martian Exploration* (eds Song, G. & Malla, R.B.) 29–35 (ASCE, 2010).
22. Butterworth, A. L., Aballain, O., Chappellaz, J. & Sephton, M. A. Combined element (H and C) stable isotope ratios of methane in carbonaceous chondrites. *Mon. Not. R. Astron. Soc.* **347,** 807–812 (2004).
23. Bartoszek, M., Wecks, M., Jakobs, G. & Mohlmann, D. Photochemically induced formation of Mars relevant oxygenates and methane from carbon dioxide and water. *Planet. Space Sci.* **59,** 259–263 (2011).
24. Allen, C. C., Gooding, J. L., Jercinovic, M. & Keil, K. Altered basaltic glass — a terrestrial analog to the soil of Mars. *Icarus* **45,** 347–369 (1981).
25. Kieffer, H. H. *et al.* Thermal and albedo mapping of Mars during the Viking Primary Mission. *J. Geophys. Res.* **82,** 4249–4291 (1977).
26. Robert, F. & Epstein, S. The concentration and isotopic composition of hydrogen, carbon and nitrogen in carbonaceous meteorites. *Geochim. Cosmochim. Acta* **46,** 81–95 (1982).
27. Sephton, M. A. *et al.* Investigating the variations in carbon and nitrogen isotopes in carbonaceous chondrites. *Geochim. Cosmochim. Acta* **67,** 2093–2108 (2003).
28. Sephton, M. A. Organic compounds in carbonaceous meteorites. *Nat. Prod. Rep.* **19,** 292–311 (2002).
29. Vigano, I. *et al.* The stable isotope signature of methane emitted from plant material under UV irradiation. *Atmos. Environ.* **43,** 5637–5646 (2009).
30. ten Kate, I. L. Organics on Mars? *Astrobiology* **10,** 589–603 (2010).
31. Navarro-González, R., Vargas, E., de la Rosa, J., Raga, A. C. & McKay, C. P. Reanalysis of the Viking results suggests perchlorate and organics at midlatitudes on Mars. *J. Geophys. Res. E* **115,** 12010, http://dx.doi.org/10.1029/2010JE003599 (2010).
32. Hecht, M. H. *et al.* Detection of perchlorate and the soluble chemistry of Martian soil at the Phoenix Lander site. *Science* **325,** 64–67 (2009).
33. Clark, B. C. & Baird, A. K. Is the Martian lithosphere sulfur rich? *J. Geophys. Res.* **84,** 8395–8403 (1979).

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

Acknowledgements F.K. was supported by the ESF (EURYI award to F.K.) and DFG (KE 884/2-1). Measurements at Utrecht University were funded by the Dutch NWO (grant number 865.07.001). A.M. was supported by the NERC (NE/F020422/1) and a Royal Society Leverhulme Trust senior research fellowship. We thank M. Greule, F. Althoff, A. Jugold, C. Tubbesing and H. F. Schöler for technical support; H. van Weelden from the Utrecht Medical Center for measuring the spectral distribution of the Xe lamp and calibrating the Waldmann ultraviolet meter; C. Brenninkmeijer, J. Lelieveld and C. Cockell for advice; and C. Jäger for providing data on ultraviolet absorption by minerals.

Author Contributions F.K., I.V., A.M. and T.R. planned the study. I.V., F.K. and M.F. carried out the experiments. F.K., I.V., A.M., U.O. and T.R. worked on the scientific interpretation 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 F.K. (frank.keppler@mpic.de).

doi:10.1038/nature11200

Late Miocene decoupling of oceanic warmth and atmospheric carbon dioxide forcing

Jonathan P. LaRiviere[1], A. Christina Ravelo[1], Allison Crimmins[1]†, Petra S. Dekens[1]†, Heather L. Ford[1], Mitch Lyle[2] & Michael W. Wara[1]†

Deep-time palaeoclimate studies are vitally important for developing a complete understanding of climate responses to changes in the atmospheric carbon dioxide concentration (that is, the atmospheric partial pressure of CO_2, p_{CO_2})[1]. Although past studies have explored these responses during portions of the Cenozoic era (the most recent 65.5 million years (Myr) of Earth history), comparatively little is known about the climate of the late Miocene (~12–5 Myr ago), an interval with p_{CO_2} values of only 200–350 parts per million by volume but nearly ice-free conditions in the Northern Hemisphere[2,3] and warmer-than-modern temperatures on the continents[4]. Here we present quantitative geochemical sea surface temperature estimates from the Miocene mid-latitude North Pacific Ocean, and show that oceanic warmth persisted throughout the interval of low p_{CO_2} ~12–5 Myr ago. We also present new stable isotope measurements from the western equatorial Pacific that, in conjunction with previously published data[5–10], reveal a long-term trend of thermocline shoaling in the equatorial Pacific since ~13 Myr ago. We propose that a relatively deep global thermocline, reductions in low-latitude gradients in sea surface temperature, and cloud and water vapour feedbacks may help to explain the warmth of the late Miocene. Additional shoaling of the thermocline after 5 Myr ago probably explains the stronger coupling between p_{CO_2}, sea surface temperatures and climate that is characteristic of the more recent Pliocene and Pleistocene epochs[11,12].

High-latitude climate reconstructions from the oxygen isotopic composition ($\delta^{18}O$) of benthic foraminifera[3] reveal a long-term cooling trend over the past ~50 Myr that occurred in conjunction with decreasing p_{CO_2} (ref. 2). However, although CO_2 levels were near pre-industrial values (280 p.p.m.v.) during the late Miocene (even the highest end of both the alkenone and leaf stomata estimates of CO_2 indicate that late Miocene CO_2 levels were less than the modern values of ~390 p.p.m.v.), high-latitude climate was too warm to support the growth of large Northern Hemisphere ice sheets[2,3] (Fig. 1a, f and Supplementary Information).

Climate modellers have tested whether external boundary conditions, such as the reduced topography of mountainous regions during the late Miocene[13], could have lowered the p_{CO_2} threshold for glaciation[14–16]; however, these tests focused on regional ice sheet growth rather than global temperatures and have not accounted for climate conditions outside the high latitudes. Palaeoclimate estimates from vegetation reconstructions suggest that warmer-than-modern conditions existed not just in the high latitudes but were globally widespread ~12–7 Myr ago[4] (Supplementary Information). Vegetation probably acted as a strong warming feedback in the late Miocene[17]; however, the boundary conditions that underlie such a vegetation distribution are not well constrained. Ocean circulation would have been integral to the global climate system of the late Miocene, but very little quantitative data exist to constrain surface circulation. For this reason, we reconstructed changes in sea surface temperature (SST) in the mid-latitude North Pacific Ocean and monitored the depth of the western tropical Pacific thermocline (the boundary between the warm surface ocean layer and the subsurface cold deep ocean) for the past ~13 Myr.

The SST estimates are derived from sediments collected at three Ocean Drilling Program (ODP) sites: Site 1010 (30° N, 118° W) in the subtropical east Pacific; Site 1021(39° N, 128° W) in the northeast Pacific at the seaward side of the northern edge of the California Current; and Site 1208 (36° N, 158° E) in the northwest Pacific at the transition zone between the subtropical and subarctic gyres (Fig. 2). SST estimates are based on the alkenone unsaturation proxy ($U^{k'}{}_{37}$) using the calibration of ref. 18. The 1010 and 1021 SST reconstructions are continuous since ~13 Myr before present. The SST reconstruction from Site 1208 is continuous since ~10 Myr before present.

The warmest SSTs occurred at the beginning of the late Miocene with subsequent cooling over the length of all three records. At subtropical east Pacific Site 1010, SSTs cooled by 5 °C between 9 and 5.8 Myr ago, and cooled an additional ~8 °C from the early Pliocene warm period, about 3.7 Myr ago, into the recent Pleistocene ice ages (Fig. 1e). A similar pattern of SST cooling was observed at northeast Pacific Site 1021; SSTs cooled by ~5 °C by 5.8 Myr ago and, after a ~3 °C increase from ~5.8 to 4.5 Myr ago, subsequently cooled an additional ~8.5 °C into the ice ages (Fig. 1e). At the northwest Pacific Site 1208, SSTs decreased by ~3 °C by 5.8 Myr ago, and continued to cool by an additional ~4 °C from 2.7 Myr ago and into the ice ages (Fig. 1d). Overall, when compared to SST estimates from the western Pacific warm pool[5], our new records of warm subtropical SSTs reveal that late Miocene meridional SST gradients were reduced relative to those of the Pliocene. Site movement by plate tectonics from one ocean temperature regime to another can explain no more than ~2 °C of the SST trend since 13 Myr ago (Supplementary Information).

Our SST records provide, despite some regional variability, the first documentation that late Miocene SSTs across a broad swathe of the North Pacific were significantly warmer than present (by 5–8 °C), and that there was nearly unidirectional cooling over the past 13 Myr. Furthermore, our records are consistent with other palaeodata[4,13,19], including bottom-water temperature estimates[19] (Fig. 1b), which indicate that the climate was warmer during the late Miocene than during the early Pliocene warm period. Thus, the preponderance of data, including our new records, indicates that global temperatures of the late Miocene, with relatively low p_{CO_2} of <350 p.p.m.v., exceeded that of the early Pliocene warm period, with relatively high p_{CO_2} of >350 p.p.m.v. (Fig. 1f). This decoupling between temperature and atmospheric p_{CO_2} trends requires an explanation. One possibility is that changes in boundary conditions (for example, continental topography, ocean basin shape) played a major role in determining the sensitivity of Earth's climate to CO_2 forcing.

During the late Miocene, the Central American Seaway (CAS) was open, the Indonesian Seaway was wider than at present, and the Bering

[1] Ocean Sciences Department, University of California, Santa Cruz, California 95064, USA. [2] Department of Oceanography, Texas A&M University, College Station, Texas 77843, USA. †Present addresses: US Environmental Protection Agency, Climate Change Division, Washington DC 20460, USA (A.C.); Department of Geosciences, San Francisco State University, San Francisco, California 94132, USA (P.S.D.); Freeman Spogli Institute for International Studies, Stanford Law School, Stanford, California 94305, USA (M.W.W.).

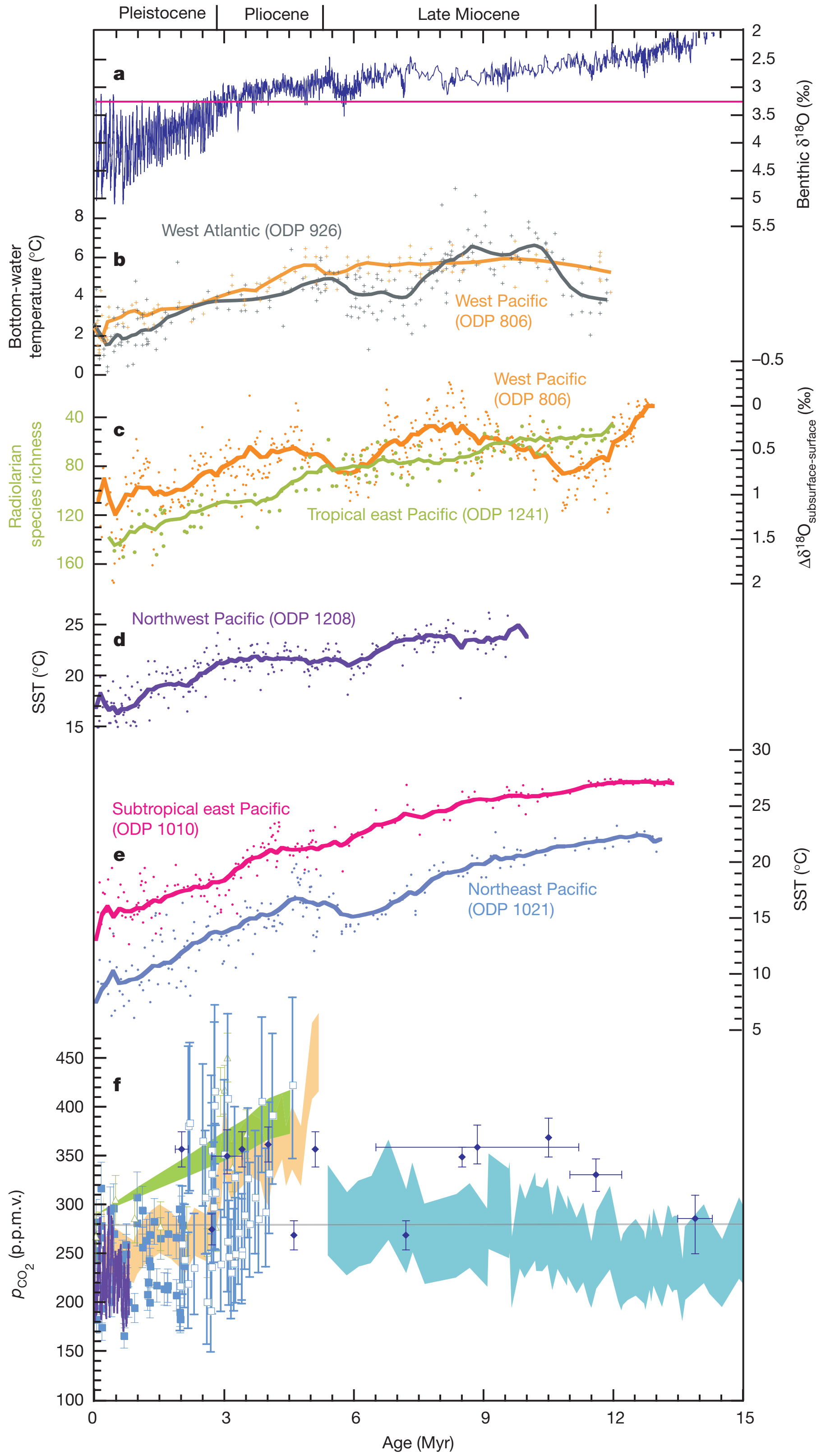


Figure 1 | Late Neogene oceanic conditions and atmospheric p_{CO_2}. **a**, Benthic foraminifera δ^{18}O record of high-latitude climate change[3,30]. Pink line denotes modern δ^{18}O. **b**, Mg/Ca-derived bottom-water temperatures from ODP Sites 806 and 926 (ref. 19). **c**, Oxygen isotopic difference between thermocline and surface foraminifera from Site 806 (orange curve, right-hand vertical axis) is inversely related to thermocline depth[5–7]. Radiolarian species richness from Site 1241[10] (green curve, left-hand vertical axis) is a reflection of thermocline depth. **d**, Alkenone SST estimates from Site 1208. **e**, Alkenone SST estimates from Sites 1021 and 1010. **f**, Estimates of atmospheric p_{CO_2} from ice cores (purple line), boron isotopes (open squares, filled squares and triangles), alkenones (green[11], gold and blue shading), and leaf stomata (diamonds). Grey line marks pre-industrial p_{CO_2} concentrations (280 p.p.m.v.). Vertical error bars represent reported uncertainty in boron isotope and leaf stomata p_{CO_2} estimates. Reported age uncertainties for leaf stomata estimates are denoted with horizontal bars. Green, gold and blue shading indicates the range between published maximum and minimum alkenone p_{CO_2} estimates. See Supplementary Information for p_{CO_2} data sources. Heavy lines in **b**–**e** represent Stineman smoothing curves applied in KaleidaGraph software (v4.1.3; http://www.synergy.com/).

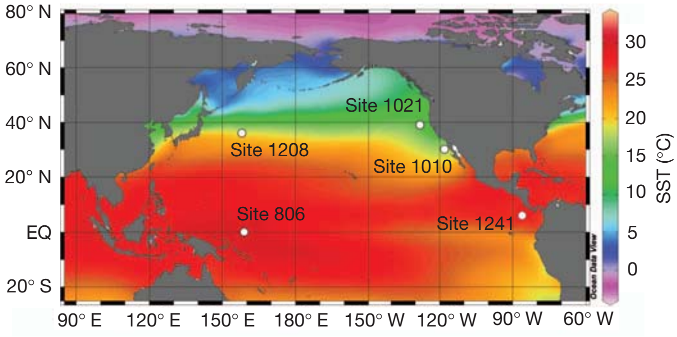


Figure 2 | Sites used in this study. ODP Sites 1208 (36° N, 158° E), 1021 (39° N, 128° W), 1010 (30° N, 118° W), 806 (0° N, 159° E) and 1241 (6° N, 86° W) overlaid on a map of mean annual SSTs[31]. EQ, Equator.

Strait was closed[13]; however, by the end of the early Pliocene the CAS was closed, the geography of the Indonesian Seaway was more similar to its modern configuration, and the Bering Strait was open[13]. Although none of the existing modelling sensitivity studies indicate that these tectonic changes could directly explain the warm, mid-latitude North Pacific temperatures observed in our reconstructions[20,21], modelling of the early Pliocene climate (5–3 Myr ago) does suggest that the CAS may have played a role in determining the depth of the thermocline. Results from general circulation models indicate that when the CAS was open, the modelled tropical thermocline was deeper[22,23], consistent with observations of the tropical Pacific thermocline[5–10,22]. Such a change in thermocline depth is important, because the ventilated thermocline maintains the balance between high-latitude oceanic heat loss and low-latitude heat gain[24]; a change in thermocline depth implies changes in surface ocean conditions that determine cloud, atmospheric water vapour and SST distributions[25]. Thus, changes in thermocline depth may be driven by internal dynamics or by changes in external boundary conditions, such as oceanic gateways, and could help to explain the climate of the late Miocene.

To assess changes in the shallow, wind-driven circulation of the ventilated thermocline, we monitored relative changes in thermocline depth at west Pacific ODP Site 806 (0° N, 159° E) (Fig. 2) using the δ^{18}O values of shells from surface-dwelling and subsurface-dwelling planktonic foraminifera. We made new stable isotopic measurements of *Globorotalia tumida* for the interval of ~4.8–0 Myr ago, which, in conjunction with previously published data[5–7] on this and other species, provides records of the δ^{18}O of *Globigerinoides sacculifer*, a surface dweller, and of *G. tumida*/*Globorotalia menardii*/*Globorotalia fohsi*, subsurface dwellers, for the past ~13 Myr. The difference between these species, $\Delta\delta^{18}\mathrm{O}_{\text{subsurface-surface}}$, reflects thermocline depth[5–7], with low values indicating a thick mixed layer and deep thermocline, and high values indicating a thinner mixed layer and shallower thermocline. Whereas the previously published data showed two pronounced intervals of a relatively thin mixed layer ~13–11 and ~7–5.8 Myr ago, the additional isotopic data in our study reveal a long-term trend in $\Delta\delta^{18}\mathrm{O}_{\text{subsurface-surface}}$ and indicate that the western equatorial Pacific thermocline has gradually shoaled since ~13 Myr ago (Supplementary Information). Evidence for thermocline changes in the eastern tropical Pacific comes from radiolarian species richness (ODP 1241; 6° N, 86° W), which shows a monotonic increase since ~12 Myr ago[10], and is consistent with a long-term increase in the number of ecological niches available as the thermocline became shallower (Fig. 1c). In addition, the authors of ref. 10 interpreted the radiolarian assemblage changes since ~4.2 Myr ago to indicate the shoaling of the eastern tropical Pacific thermocline from a relatively deep configuration that existed for the majority, if not all, of the late Miocene. Furthermore, foraminifera faunal reconstructions throughout the tropical Pacific show that the thermocline has generally shoaled from a relatively deep position in the middle Miocene to a shallower depth by the end of the late Miocene[8,9].

Overall, our data indicate that the oceanic state of the late Miocene was similar to that of the early Pliocene warm period, though more extreme (with warmer SSTs, smaller SST gradients and a deeper thermocline). Modelling of the early Pliocene conditions demonstrates that a tight coupling between the deep thermocline, expanded tropical warmth, and the reduction in meridional and zonal SST gradients resulted in mean global temperatures 3–4 °C warmer than today and the suppression of Northern Hemisphere glaciation[26]. The models, which are constrained by Pliocene proxy data, show that the expanded tropical warmth results in enhanced subtropical evaporation, greenhouse warming from water vapour, and warming from an increase in subtropical high 'greenhouse clouds'. These processes form a feedback loop that further facilitates maintenance of a deep thermocline and a warm climate with expanded tropical warmth[27]. Applying this idea to the Miocene provides a framework to understand our new SST and thermocline observations, all of which are consistent with warmer-than-modern global temperatures, as is the case in the early Pliocene. Furthermore, when applied to the Miocene, the Pliocene body of work suggests that, in the absence of large changes in p_{CO_2}, a tectonically driven change in upper ocean structure and tropical warm pool expanse could have, by itself, affected global temperatures through changes in atmospheric water vapour (a greenhouse gas) concentrations and in planetary albedo (through cloud type and distribution).

Changes in thermocline depth could explain the differences between late Miocene and Pliocene climate responses to atmospheric p_{CO_2} forcing. The shoaling of the thermocline has been directly linked to intensification of tropical SST gradients[22,28] and the strength of Walker and Hadley atmospheric circulation. However, when the thermocline is sufficiently deep (as it was in the tropical Pacific during the late Miocene), its movement is not coupled with SSTs. An increase in coupling appears to occur when the thermocline is shallow enough to pass some threshold depth. This threshold can explain why the thermocline depth variation in the western equatorial Pacific—a region where the thermocline is currently deeper than the surface Ekman layer of wind-driven mixing—did not appear to be coupled to SSTs until the thermocline in the eastern tropical Pacific shoaled adequately; such shoaling happened in the early Pliocene or, at the earliest, during the conclusion of the late Miocene. Once the thermocline became sufficiently shallow to affect SSTs, the climate system seems to have become sensitive to climate perturbations that had previously been inconsequential, including those driven by changes in p_{CO_2}. For example, with a shallow thermocline any small change (for example, in winds or in upwelling strength) that affected the low-latitude SSTs would be accompanied by strong positive feedbacks; changes in surface pressure gradients could reinforce initial changes in winds and in atmospheric water vapour and cloud formation[26] that amplify global temperature change. A shallower thermocline and accompanying enhanced climate sensitivity could explain why the Northern Hemisphere ice ages began in the Pliocene, rather than in the Miocene at comparable p_{CO_2} levels, and why a close coupling between glacial–interglacial climate cycles and p_{CO_2} developed after ~2.7 Myr ago[29].

Differences in the oceanic gateway boundary conditions of the late Miocene and Pliocene may have been the ultimate cause of the increase in climate sensitivity to p_{CO_2} forcing; the ocean basin configuration of the Pliocene, rather than the configuration that existed for the majority of the late Miocene, enabled the thermocline to shoal past a depth that was critical for coupling the thermocline and SSTs. However, although the closing of the CAS is the best candidate for forcing major upper-ocean structure changes in the earliest Pliocene, much work is needed to verify this idea and to test the effects of other ocean gateways on thermocline depth. Future work should aim to increase the geographic coverage of the surface and subsurface oceanographic reconstructions with an emphasis on ocean gateway regions (for example, the CAS, Indonesian Seaway and Bering Strait). Such evidence would help to constrain the timing and the nature of ocean circulation change, and therefore climate change, associated with tectonic events in the late Miocene and early Pliocene.

METHODS SUMMARY

Lipids were extracted from 0.5–5 g of crushed sediment with either a 3:1 dichloromethane:methanol mix or pure dichloromethane using a Dionex ASE200 accelerated solvent extractor. The total lipid extract was evaporated to dryness under N_2 and redissolved in 100–200 µl of toluene with hexatriacontane and heptatriacontane internal standards. Separation of organic compounds was carried out on an HP6890 gas chromatograph equipped with a flame ionization detector. Long-term reproducibility of liquid standard replicates included in each gas chromatography run was within ±0.007 $U^{k'}_{37}$ units, which is equivalent to ±0.2 °C (s.d., $n = 139$). We monitored the long-term precision of the entire method by processing a sediment standard with each batch of samples. Reproducibility for the sediment standards used for the 1208, 1021 and 1010 sites was ±0.016 (s.d., $n = 25$), ±0.013 (s.d., $n = 27$) and ±0.015 (s.d., $n = 9$) $U^{k'}_{37}$ units, respectively. The standard error of the estimate for the global SST calibration of ref. 18 is ±1.5 °C.

Fossil shells of *G. tumida* were analysed for oxygen isotopic composition ($\delta^{18}O$) using a Fisons Prism III dual inlet isotope ratio mass spectrometer. The precision of NBS-19 (NIST-8544) and of an in-house Carrera Marble standard was better than 0.08‰ for $\delta^{18}O$. Measurements of $\delta^{18}O$ are reported relative to Vienna-Pee Dee Belemnite (V-PDB).

Received 15 November 2011; accepted 2 May 2012.

1. Hansen, J. *et al.* Target atmospheric CO_2: where should humanity aim? *Open Atmos. Sci. J.* **2,** 217–231 (2008).
2. Ruddiman, W. F. A paleoclimatic enigma? *Science* **328,** 838–839 (2010).
3. Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292,** 686–693 (2001).
4. Pound, M. J. *et al.* A Tortonian (Late Miocene, 11.61–7.25 Ma) global vegetation reconstruction. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **300,** 29–45 (2011).
5. Nathan, S. A. & Leckie, R. M. Early history of the Western Pacific Warm Pool during the middle to late Miocene (~13.2–5.8 Ma): role of sea-level change and implications for equatorial circulation. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **274,** 140–159 (2009).
6. Wara, M. W., Ravelo, A. C. & Delaney, M. L. Permanent El Niño-like conditions during the Pliocene warm period. *Science* **309,** 758–761 (2005).
7. Chaisson, W. P. & Ravelo, A. C. Pliocene development of the east-west hydrographic gradient in the equatorial Pacific. *Paleoceanography* **15,** 497–505 (2000).
8. Keller, G. Depth stratification of planktonic foraminifers in the Miocene ocean. *Geol. Soc. Am.* **163,** 177–195 (1985).
9. Kennett, J. P., Keller, G. & Srinivasan, M. S. Miocene planktonic foraminiferal biogeography and paleoceanographic development of the Indo-Pacific region. *Geol. Soc. Am.* **163,** 197–236 (1985).
10. Kamikuri, S., Motoyama, I., Nishi, H. & Iwai, M. Evolution of Eastern Pacific Warm Pool and upwelling processes since the middle Miocene based on analysis of radiolarian assemblages: Response to Indonesian and Central American Seaways. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **280,** 469–479 (2009).
11. Pagani, M., Liu, Z. H., LaRiviere, J. & Ravelo, A. C. High Earth-system climate sensitivity determined from Pliocene carbon dioxide concentrations. *Nature Geosci.* **3,** 27–30 (2010).
12. Siegenthaler, U. *et al.* Stable carbon cycle-climate relationship during the late Pleistocene. *Science* **310,** 1313–1317 (2005).
13. Lyle, M. *et al.* Pacific Ocean and Cenozoic evolution of climate. *Rev. Geophys.* **46,** RG2002 (2008).
14. DeConto, R. M. *et al.* Thresholds for Cenozoic bipolar glaciation. *Nature* **455,** 652–656 (2008).
15. Foster, G., Lunt, D. & Parrish, R. Mountain uplift and the glaciation of North America — a sensitivity study. *Clim. Past* **6,** 707–717 (2010).
16. Lunt, D., Foster, G., Haywood, A. & Stone, E. Late Pliocene Greenland glaciation controlled by a decline in atmospheric CO_2 levels. *Nature* **454,** 1102–1105 (2008).
17. Knorr, G., Butzin, M., Micheels, A. & Lohmann, G. A warm Miocene climate at low atmospheric CO_2 levels. *Geophys. Res. Lett.* **38,** L20701, http://dx.doi.org/10.1029/2011GL048873 (2011).
18. Müller, P. J., Kirst, G., Ruhland, G., von Storch, I. & Rosell-Melé, A. Calibration of the alkenone paleotemperature index $U^{K'}_{37}$ based on core-tops from the eastern South Atlantic and the global ocean (60°N-60°S). *Geochim. Cosmochim. Acta* **62,** 1757–1772 (1998).
19. Lear, C. H., Rosenthal, Y. & Wright, J. D. The closing of a seaway: ocean water masses and global climate change. *Earth Planet. Sci. Lett.* **210,** 425–436 (2003).
20. Lunt, D. J., Valdes, P. J., Haywood, A. & Rutt, I. C. Closure of the Panama Seaway during the Pliocene: implications for climate and Northern Hemisphere glaciation. *Clim. Dyn.* **30,** 1–18 (2007).
21. Schneider, B. & Schmittner, A. Simulating the impact of the Panamanian seaway closure on ocean circulation, marine productivity and nutrient cycling. *Earth Planet. Sci. Lett.* **246,** 367–380 (2006).
22. Steph, S. *et al.* Early Pliocene increase in thermohaline overturning: a precondition for the development of the modern equatorial Pacific cold tongue. *Paleoceanography* **25,** PA2202, http://dx.doi.org/10.1029/2008PA001645 (2010).
23. Zhang, X. *et al.* Changes in equatorial Pacific thermocline depth in response to Panamanian seaway closure: insights from a multi-model study. *Earth Planet. Sci. Lett.* **317–318,** 76–84 (2012).
24. Boccaletti, G., Pacanowski, R. C., Philander, S. G. H. & Fedorov, A. V. The thermal structure of the upper ocean. *J. Phys. Oceanogr.* **34,** 888–902 (2004).
25. Philander, S. G. & Fedorov, A. V. Role of tropics in changing the response to Milankovich forcing some three million years ago. *Paleoceanography* **18,** 1045, http://dx.doi.org/10.1029/2002PA000837 (2003).
26. Brierley, C. M. & Fedorov, A. V. Relative importance of meridional and zonal sea surface temperature gradients for the onset of the ice ages and Pliocene-Pleistocene climate evolution. *Paleoceanography* **25,** PA2214, http://dx.doi.org/10.1029/2009PA001809 (2010).
27. Fedorov, A., Brierley, C. & Emanuel, K. Tropical cyclones and permanent El Niño in the early Pliocene epoch. *Nature* **463,** 1066–1070 (2010).
28. Fedorov, A. V. *et al.* The Pliocene paradox (mechanisms for a permanent El Niño). *Science* **312,** 1485–1489 (2006).
29. Herbert, T., Peterson, L., Lawrence, K. & Liu, Z. Tropical ocean temperatures over the past 3.5 million years. *Science* **328,** 1530–1534 (2010).
30. Lisiecki, L. E. & Raymo, M. E. A. Pliocene-Pleistocene stack of 57 globally distributed benthic delta $\delta^{18}O$ records. *Paleoceanography* **20,** PA1003, http://dx.doi.org/10.1029/2004PA001071 (2005).
31. Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P. & Garcia, H. E. in *World Ocean Atlas 2005, NOAA Atlas NESDIS 61* Vol. 1 (ed. Levitus, S.) 182 (US Government Printing Office, 2006).

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

Acknowledgements We thank the Ravelo laboratory group for discussions. We also thank J. Zachos and P. Koch for comments on the manuscript. L. Lajoie, P. Talmage and T. M. Aung assisted in sample preparation and analysis. D. Andreasen and R. Franks provided analytical support. This research used samples and/or data provided by the Integrated Ocean Drilling Program (IODP). Funding for this research was provided by NSF grant OCE0902047.

Author Contributions J.P.L. and A.C.R. did the primary data analysis and wrote the paper with intellectual feedback from all authors. J.P.L. generated alkenone temperature reconstructions; J.P.L., P.S.D., H.L.F., A.C. and M.W.W. analysed foraminifera $\delta^{18}O$.

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

LETTER

doi:10.1038/nature11141

Early differentiation and volatile accretion recorded in deep-mantle neon and xenon

Sujoy Mukhopadhyay[1]

The isotopes ^{129}Xe, produced from the radioactive decay of extinct ^{129}I, and ^{136}Xe, produced from extinct ^{244}Pu and extant ^{238}U, have provided important constraints on early mantle outgassing and volatile loss from Earth[1,2]. The low ratios of radiogenic to non-radiogenic xenon (^{129}Xe/^{130}Xe) in ocean island basalts (OIBs) compared with mid-ocean-ridge basalts (MORBs) have been used as evidence for the existence of a relatively undegassed primitive deep-mantle reservoir[1]. However, the low ^{129}Xe/^{130}Xe ratios in OIBs have also been attributed to mixing between subducted atmospheric Xe and MORB Xe, which obviates the need for a less degassed deep-mantle reservoir[3,4]. Here I present new noble gas (He, Ne, Ar, Xe) measurements from an Icelandic OIB that reveal differences in elemental abundances and ^{20}Ne/^{22}Ne ratios between the Iceland mantle plume and the MORB source. These observations show that the lower ^{129}Xe/^{130}Xe ratios in OIBs are due to a lower I/Xe ratio in the OIB mantle source and cannot be explained solely by mixing atmospheric Xe with MORB-type Xe. Because ^{129}I became extinct about 100 million years after the formation of the Solar System, OIB and MORB mantle sources must have differentiated by 4.45 billion years ago and subsequent mixing must have been limited. The Iceland plume source also has a higher proportion of Pu- to U-derived fission Xe, requiring the plume source to be less degassed than MORBs, a conclusion that is independent of noble gas concentrations and the partitioning behaviour of the noble gases with respect to their radiogenic parents. Overall, these results show that Earth's mantle accreted volatiles from at least two separate sources and that neither the Moon-forming impact nor 4.45 billion years of mantle convection has erased the signature of Earth's heterogeneous accretion and early differentiation.

The noble gases play an important role in understanding mantle structure and in quantifying mass and volatile fluxes into and out of the mantle[1,5–10]. Interpretation of the differences in noble gas composition between MORBs and OIBs has primarily relied upon steady-state mantle models that require ^{129}Xe and primordial noble gases such as ^{3}He, ^{22}Ne and ^{36}Ar in the volatile-depleted upper mantle to be derived from a primitive volatile-rich lower mantle via plume mass flow[8,9]. Mixtures of the plume-derived noble gases, radiogenic noble gases produced in the upper mantle, and subducted atmospheric Ar and Xe into the MORB source produce the noble gas isotopic compositions observed in MORBs. Such interpretations have been challenged on the grounds that a primitive lower mantle is at odds with geodynamic models of whole mantle convection[11], seismological observations of plate subduction into the lower mantle[12], and the depleted geochemical characteristics of OIBs[13]. However, the steady-state models could still be correct, if instead of the whole lower mantle, the primitive layer were much smaller, such as the D″ layer at the base of the mantle[9]. Numerous alternative interpretations of the noble gas observations have also been proposed, particularly to explain high ^{3}He/^{4}He ratios in OIBs within the framework of whole mantle convection, as some geodynamic models suggest that preserving primitive layers over the age of the Earth is problematic[11]. These interpretations assign high ^{3}He/^{4}He ratios to non-primordial processed mantle material[14–16] and include the preservation of high ^{3}He/^{4}He ratios in U-Th depleted residues of mantle melting.

Relative elemental abundances of the noble gases and precise measurements of isotopic ratios of Ne, Ar and Xe can distinguish between the different hypotheses put forward to explain the noble gas observations. For example, the steady-state models predict that MORBs and OIBs should have the same elemental abundances and that MORB-Xe is related to OIB-Xe through addition of atmospheric and fission-produced Xe. Whereas the elemental abundance pattern and the isotopic ratios of the upper mantle are relatively well-defined[3,17], the composition of the OIB mantle source remains poorly known; this is because low noble gas concentrations and post-eruptive atmospheric contamination often result in the mantle Ne, Ar and Xe compositions being overprinted. To overcome this problem, single large pieces of the DICE 10 basaltic glass from Iceland[18,19] were analysed by step-crushing under vacuum.

The new observations from Iceland show three steps where ^{20}Ne/^{22}Ne ratios reach 12.88 ± 0.06, 12.76 ± 0.01 and 12.74 ± 0.02 (Fig. 1). These values are unequivocally higher than the upper-mantle composition of ≤ 12.5, constrained from continental well gases[3,20]. Previously, the only exceptions to a mantle ^{20}Ne/^{22}Ne ratio of ≤ 12.5 were carbonatites from the 380-Myr-old plume in the Kola peninsula of Russia, with a maximum measured ^{20}Ne/^{22}Ne of 13.04 ± 0.2 (ref. 21). Thus, the Kola and Iceland plumes require the ^{20}Ne/^{22}Ne ratio of the MORB and plume sources to be different. Because ^{20}Ne and ^{22}Ne are primordial, and atmospheric Ne is not subducted back into the mantle in significant quantities[3], the ^{20}Ne/^{22}Ne ratio in the mantle does not evolve over time. Consequently, the new observations from Iceland and those from the Kola plume suggest that at least two separate reservoirs contributed neon during Earth's accretion[20].

Elemental abundance ratios provide additional insights into the origin of the observed difference in ^{20}Ne/^{22}Ne composition between MORBs and OIBs. The noble gases have different solubilities in a basalt melt, and so magmatic degassing will fractionate the elemental ratios in the melt compared to the mantle source. The relative abundances of ^{4}He, ^{21}Ne and ^{40}Ar in the DICE 10 sample are, however, in the same proportion as the mantle production rates for these isotopes (Supplementary Fig. 3), indicating that the sample preserves unfractionated mantle elemental ratios.

Excellent correlations are observed between isotope ratios and elemental ratios, allowing precise determination of the mantle source elemental abundance ratios. Figure 2 demonstrates that the Iceland plume has a lower ^{3}He/^{22}Ne ratio than the MORB source, consistent with previous observations that plume and MORB sources may have different ^{3}He/^{22}Ne ratios[18,22]. In addition, the new Icelandic observations indicate that the plume source has higher ^{3}He/^{36}Ar and ^{22}Ne/^{36}Ar ratios than the MORB source (Fig. 2; Supplementary Table 1). Although the high-pressure behaviours of the noble gases are not well constrained, the differences in the elemental abundance ratios between MORBs and the Iceland sample do not appear to be linked in any systematic way to diffusive fractionation or ancient magmatic degassing, such as degassing from a magma ocean (Fig. 2). Thus, the differences in ^{20}Ne/^{22}Ne and ^{3}He/^{22}Ne ratios observed between the Iceland source and the MORB source probably reflect an accretional

[1] Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

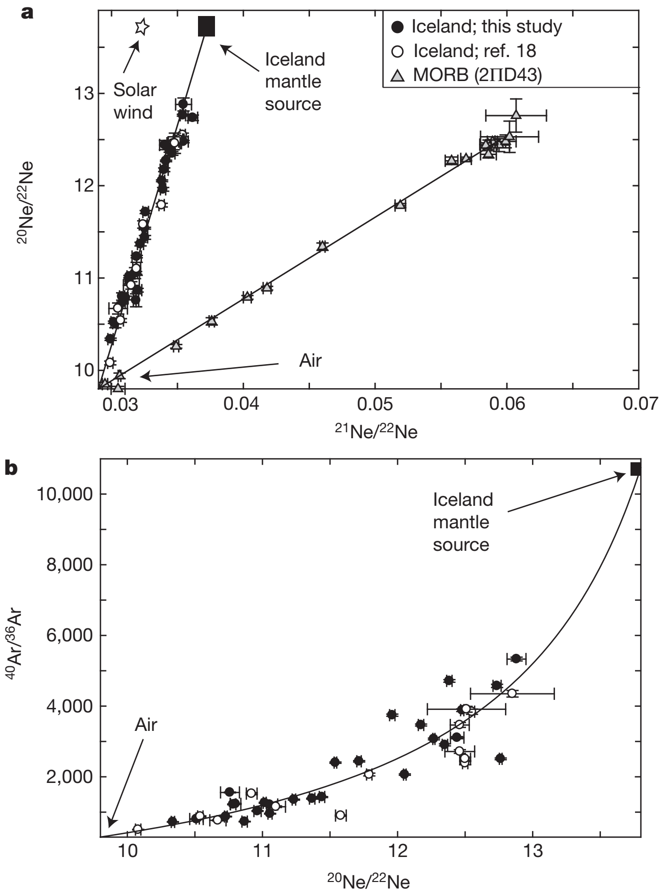


Figure 1 | Differences in neon and argon isotopic composition between MORB and the Iceland plume. **a**, Neon three-isotope plot showing the new analyses of the DICE 10 sample (filled circles) from Iceland in comparison to previously published data for this sample (open circles; ref. 18) and the gas-rich 'popping rock' (2ΠD43) from the north Mid-Atlantic Ridge (open triangles; ref. 17). Error bars are 1σ, and for clarity, two previous analyses[18] with large error bars have not been shown. Step-crushing of a mantle-derived basalt produces a linear trend that reflects variable amounts of post-eruptive air contamination in vesicles containing mantle Ne. The slope of the line is a function of the ratio of nucleogenic ^{21}Ne to primordial ^{22}Ne, with steeper slopes indicating a higher proportion of primordial ^{22}Ne and, thus, a less degassed mantle source. The slope of the Iceland line based on the new analyses is consistent with that obtained previously[18]. Importantly, $^{20}Ne/^{22}Ne$ ratios of 12.88 ± 0.06 are distinctly higher than the MORB source $^{20}Ne/^{22}Ne$ of $\leq$12.5 as constrained from continental well gases[20]. **b**, Ne–Ar compositions of individual step crushes of the DICE 10 sample. ^{40}Ar is generated by radioactive decay of ^{40}K, and low $^{40}Ar/^{36}Ar$ ratios are indicative of a less degassed mantle. The data reflect mixing between a mantle component and post-eruptive atmospheric contamination. A least-squares hyperbolic fit through the data yields a $^{40}Ar/^{36}Ar$ ratio of 10,745 ± 3,080, corresponding to a mantle solar $^{20}Ne/^{22}Ne$ ratio of 13.8. This Ar isotopic ratio is used as the mantle source value for Iceland in Figs 2 and 3. Symbols as in **a**; error bars are 1σ.

signal, with the MORB mantle more similar to the meteoritic Ne-B[20] composition and the deep mantle similar to a solar Ne composition[20,21].

Taken together, the differences in $^{20}Ne/^{22}Ne$, He/Ne and Ne/Ar ratios between MORBs and the Iceland plume require that heterogeneities from the early Earth still exist in the present day mantle, and the new Xe measurements provide conclusive evidence for this interpretation. Previous studies have observed lower $^{129}Xe/^{130}Xe$ ratios in OIBs compared to MORBs[3,17,18]. The lower measured $^{129}Xe/^{130}Xe$ ratios in OIBs could reflect syn- to post-eruptive atmospheric contamination of the lavas, mixing between subducted atmospheric Xe and MORB-type Xe[3,4,18], or different I/Xe ratios for the plume and MORB sources[1]. Ar–Xe mixing systematics (Fig. 3) constrain the Iceland mantle source $^{129}Xe/^{130}Xe$ ratio to be 6.98 ± 0.07, significantly lower than the MORB source ratio of 7.9 ± 0.14 (ref. 3). Therefore, syn- to post-eruptive

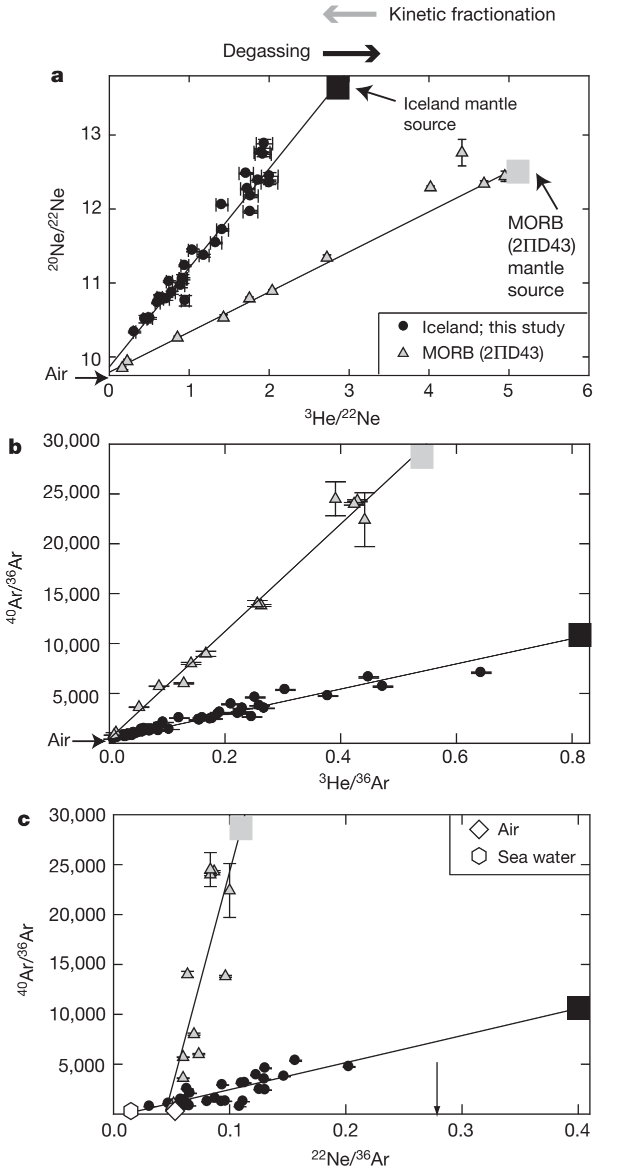


Figure 2 | Differences in elemental abundances and isotope ratios between MORB and the Iceland plume. Error bars are 1σ. **a**, $^{3}He/^{22}Ne$ versus $^{20}Ne/^{22}Ne$; **b**, $^{3}He/^{36}Ar$ versus $^{40}Ar/^{36}Ar$; and **c**, $^{22}Ne/^{36}Ar$ versus $^{40}Ar/^{36}Ar$. The mantle source composition for 2ΠD43 (filled grey square in all panels) is based on the $^{40}Ar/^{36}Ar$ and $^{20}Ne/^{22}Ne$ ratios as defined in ref. 30, and the mantle source composition for Iceland (filled black square in all panels) is based on Fig. 1. The grey and black arrows at the top of the figure indicate how elemental ratios evolve as a result of kinetic fractionation and solubility controlled degassing, respectively. Good linear relationships are observed between isotope ratios and elemental ratios, which reflect mixing between mantle-derived noble gases and post-eruptive atmospheric contamination. Lines are robust linear regressions through the data with the atmospheric contaminant near the origin and the mantle source at the other end. Arrow in **c** indicates the minimum $^{22}Ne/^{36}Ar$ ratio of the Iceland mantle source given the measured $^{40}Ar/^{36}Ar$ ratio of 7,047 (Supplementary Table 6). Because of systematic differences in noble gas solubilities and diffusivities, the differences in elemental abundances are not likely to be generated through ancient fractionation associated with diffusion or magmatic outgassing. For example, kinetic fractionation should lead to higher $^{3}He/^{22}Ne$ and higher $^{3}He/^{36}Ar$–$^{22}Ne/^{36}Ar$ ratios. However, the Iceland source has a lower $^{3}He/^{22}Ne$ and higher $^{3}He/^{36}Ar$–$^{22}Ne/^{36}Ar$. Likewise, adding recycled atmospheric gases to the MORB source cannot produce the plume noble gas compositions. Finally, **c** shows that preferential recirculation of atmospheric Ar into the plume source does not explain the higher $^{22}Ne/^{36}Ar$ of the plume source and because of the difference in MORB and OIB $^{22}Ne/^{36}Ar$ ratios, adding radiogenic ^{40}Ar to the plume composition is not likely to generate the $^{40}Ar/^{36}Ar$ ratio in MORBs.

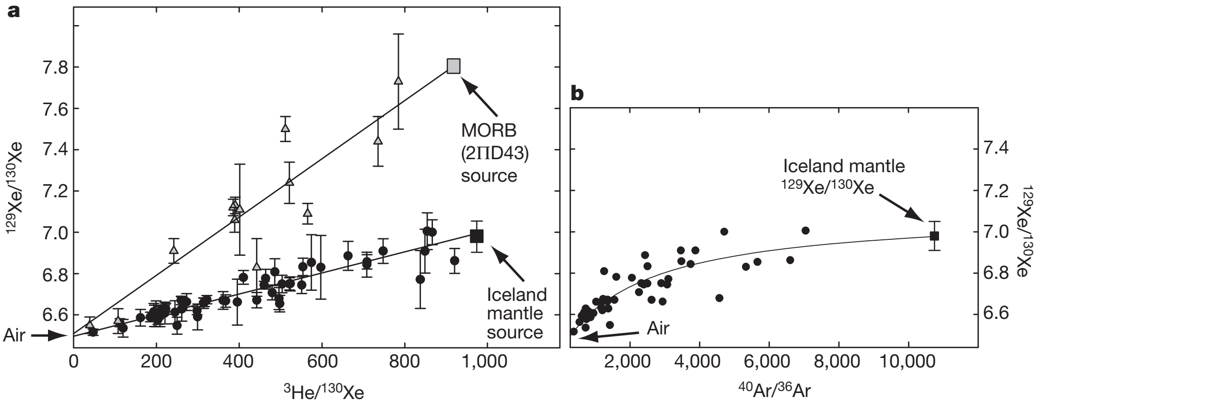


Figure 3 | Differences in Xe isotopic composition between MORB and the Iceland plume. **a**, Correlation between ^{129}Xe and ^{3}He in the 'popping rock' MORB (2ΠD43)[17] and Iceland (DICE 10). Error bars are 1σ. Data points are individual step crushes that reflect different degrees of post-eruptive atmospheric contamination in the vesicles. Air lies near the origin and the mantle compositions at the other end of the linear arrays. The straight lines are robust regressions through the data. Because mixing in this space is linear, the lines also represent the trajectories along which the mantle sources will evolve when mixed with subducted air. The new observations from Iceland demonstrate that the Iceland plume ^{129}Xe/^{130}Xe ratio cannot be generated solely through adding recycled atmospheric Xe to the MORB source, and vice versa. Thus, two mantle reservoirs with distinct I/Xe ratios are required. The mantle ^{129}Xe/^{130}Xe ratio of 6.98 ± 0.07 for Iceland was derived from a hyperbolic least-squares fit through the Ar-Xe data (**b**) corresponding to a mantle ^{40}Ar/^{36}Ar ratio of 10,745. Note that given the curvature in Ar–Xe space, the ^{129}Xe/^{130}Xe in the Iceland mantle source is not particularly sensitive to the exact choice of the mantle ^{40}Ar/^{36}Ar ratio.

contamination processes are ruled out as the reason for the lower ^{129}Xe/^{130}Xe ratios at Iceland.

The data in Fig. 3a demonstrate that the Iceland and MORB source mantles evolved with different I/Xe ratios, requiring the two mantle sources to have separated by 4.45 Gyr ago with limited subsequent mixing between the two. As atmosphere is located near the origin in this plot (Fig. 3a), and mixing in this space is linear, adding subducted atmospheric Xe to the MORB source clearly cannot produce the Iceland mantle source composition. Similarly, the Iceland source cannot supply Xe to the MORB source unless it is augmented by radiogenic ^{129}Xe produced from decay of ^{129}I. However, ^{129}I became extinct at ~4.45 Gyr ago. Therefore, the two sources must have been separated before 4.45 Gyr ago and subsequently the sources were not homogenized, as otherwise the differences in ^{129}Xe/^{130}Xe would not have been preserved in the present day mantle. Plumes, therefore, cannot have supplied Xe and all of the primordial volatiles to the MORB source (Figs 1–3), contradicting predictions of the steady-state mantle models[8,9].

The new Xe isotopic measurements also indicate a difference in ^{129}Xe/^{136}Xe ratios between MORBs and the Iceland plume that cannot be related solely to subduction of atmospheric Xe, or to addition of ^{136}Xe to MORB Xe (Fig. 4). ^{136}Xe, along with ^{131}Xe, ^{132}Xe and ^{134}Xe, is produced by fission from extinct ^{244}Pu (half-life 80 Myr) and extant ^{238}U. However, Pu and U produce the four fission Xe isotopes in different proportions, and so measurements of the fissiogenic isotopes can be used to deconvolve Pu- from U-derived Xe (ref. 23). A reservoir that has remained closed to volatile loss over Earth history should have ~97% of the fission Xe isotopes produced from ^{244}Pu. Progressive degassing of a reservoir, particularly after ^{244}Pu becomes extinct, leads to increasing proportions of U-derived fission Xe in the reservoir.

Compared to the MORB source, the Iceland plume source has a substantially higher proportion of Pu- to U-derived fission Xe. Thus, the Iceland plume source is less degassed than the MORB source. Depending on whether the initial Xe isotopic composition of the mantle was solar or chondritic[24], the MORB source has a few per cent to 43 ± 16% of fission ^{136}Xe derived from ^{244}Pu. The corresponding values for Iceland are 66 ± 19% to 99^{+1}_{-3}% (Supplementary Table 4); the latter value, based on an initial chondritic mantle Xe composition, is identical to values expected for closed system evolution. Hence, irrespective of the initial Xe isotopic composition of the mantle, the Iceland plume sample has a higher proportion of Pu- to U-derived fission Xe compared to MORB samples when both sets of data are processed in the same manner during the deconvolution (Methods, Supplementary Table 4). The requirement for a lower degree of degassing for the Iceland source, based on its higher proportion of Pu- to U-derived fission Xe, is a conclusion that is independent of the absolute concentrations of noble gases and the relative partition coefficients of the noble gases with respect to their radiogenic parents.

The combined I–Pu–Xe system has been used to constrain the closure time for volatile loss of a mantle reservoir through the ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratio[1,2,6,25], where ^{129*}Xe is the decay product of ^{129}I decay and ^{136*}Xe$_{\mathrm{Pu}}$ is ^{136}Xe produced from ^{244}Pu fission. ^{129}I has a shorter half-life than ^{244}Pu, and as a result higher ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratios are indicative of earlier closure to volatile loss[1,2,6,25]. Depending on the initial mantle Xe composition, the ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratio varies between $2.9^{+0.1}_{-0.1}$ and $5.8^{+1.1}_{-0.9}$ for the Iceland mantle and the corresponding values for MORBs are between $7.9^{+3.3}_{-2.9}$ and $64.9^{+132}_{-31.2}$ (Methods; Supplementary Table 4). If the mantle had a homogenous I/Pu ratio, the lower ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratio in the plume source would paradoxically imply that the deep mantle became closed to volatile loss after the shallow mantle. A simpler explanation is that the lower ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratio reflects a lower I/Pu ratio for the plume source compared to the MORB source. These differences would indicate that the initial phase of Earth's accretion was volatile-poor compared to the later stages of accretion, a conclusion consistent with a recent Pd–Ag study[26].

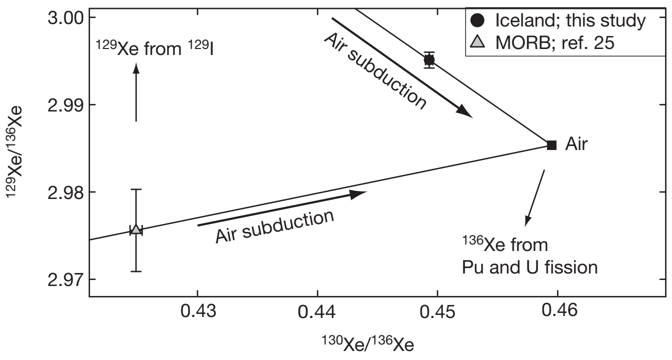


Figure 4 | Difference in the measured ^{129}Xe/^{136}Xe ratio between MORB and the Iceland plume. Unlike the traditional ^{129}Xe/^{130}Xe–^{136}Xe/^{130}Xe plot, the x and y errors are de-correlated. The arrows illustrate how the MORB and OIB source compositions evolve as subducted air is added. Error bars are 1σ. The figure demonstrates a small Xe isotopic difference between the Iceland plume and MORBs that cannot be related solely through recycling atmospheric Xe or by adding fissiogenic ^{136}Xe to MORB Xe. The data points represent the weighted means of the different step crushes for MORBs (n = 38) and Iceland (n = 51; this study). The MORB ^{129}Xe/^{136}Xe ratio was calculated from the weighted means of the ^{129}Xe/^{130}Xe and ^{136}Xe/^{130}Xe ratios[25].

The long-term separation of the plume and MORB sources inferred from the ^{129}Xe/^{130}Xe ratio (Fig. 3), and the lower degree of outgassing of the plume source inferred from fission Xe, have implications for interpreting the differences in ^{3}He/^{4}He ratios between plumes and MORBs. It suggests that the high ^{3}He/^{4}He ratios observed in many mantle plumes are from an ancient reservoir created within the first 100 Myr of Solar System history, and that the high ^{3}He/^{4}He ratios reflect a lower degree of outgassing of the plume source compared to the MORB source. However, the incorporation of noble gases into the deep mantle appears to be associated with lower abundances (compared to the MORB source) of other volatiles, such as iodine and water. Subsequent processing, through partial melting of the mantle, led to a greater degree of volatile loss from the MORB source, with differences in the MORB and plume mantle sources established by 4.45 Gyr ago.

I note that the persistence of two different mantle Xe reservoirs established by 4.45 Gyr ago may seem to conflict with the homogeneous ^{142}Nd/^{144}Nd in the present-day mantle and crust[27]. A possible explanation is the low recycling efficiency of the noble gases into the mantle compared to the lithophile elements. For example, compared to the present mantle, the Hadean mantle had a 20 p.p.m. excess in ^{142}Nd/^{144}Nd that was subsequently erased through crustal recycling and mixing[27]. In addition, the observed heterogeneity in Xe isotopic composition compared to ^{142}Nd/^{144}Nd could be related to the magnitude of I–Xe and Pu–Xe fractionations during partial melting and magmatic degassing, which led to larger Xe isotopic anomalies compared to ^{142}Nd/^{144}Nd. In any case, since the chemical differences established very early in Earth's history are still preserved in the ^{129}Xe/^{130}Xe ratios, direct mixing between the MORB and plume reservoirs since 4.45 Gyr ago must have been limited.

The preservation of an ancient mantle reservoir has important implications for Earth evolution. For example, although it has been hypothesized that the Moon-forming giant impact led to homogenization of the mantle[28], the Ne and Xe isotopic heterogeneities suggest that complete homogenization did not occur. Finally, whether the primordial noble gases that supply mantle plumes are distributed throughout the whole lower mantle or are localized in a region of the deep mantle—such as the large low-shear-wave-velocity provinces (LLSVPs) at the base of the mantle[29]—can be debated. However, if noble gases in plumes are derived from the LLSVPs[29], then based on the new Xe evidence, LLSVPs are features that have existed since the formation of the Earth and cannot exclusively be composed of subducted slabs.

METHODS SUMMARY

To minimize contamination by air adsorbed on the sample surface, single large pieces of the DICE 10 glass (~2.2–3.5 g) from Iceland were analysed. Gases were released by step-crushing under vacuum and noble gas abundances and isotopic ratios were determined using a Nu Noblesse multi-collector mass spectrometer (Supplementary Tables 5–7). Mass discrimination was monitored through sample-standard bracketing using air as a standard.

Deconvolution of U- from Pu-derived fission Xe was done using five Xe isotopes (^{130,131,132,134,136}Xe). To investigate whether the conclusion of a higher Pu- to U-derived fission Xe for Iceland is robust, three different initial mantle Xe isotopic compositions were investigated: U-Xe, solar wind Xe and chondritic Xe. Furthermore, two different techniques were used to compute the mantle source fission isotopic compositions. In the first technique, the atmospheric Xe isotopic ratios were linearly projected to a mantle source value of ^{129}Xe/^{132}Xe through the weighted averages of the measured Xe isotopic ratios (Methods). In the second technique, ^{136}Xe/^{130}Xe ratios in the individual steps were regressed against ^{129}Xe/^{130}Xe ratios. Given the mantle ^{129}Xe/^{130}Xe ratio (Fig. 3), the mantle ^{136}Xe/^{130}Xe was computed from the best-fit slopes. The mantle values for the other fission isotope ratios were then established by regressing the Xe isotopic ratios in individual steps against the ^{136}Xe/^{130}Xe ratio. The fission isotopic compositions were used to solve the system of equations $Ax = b$ with $\Sigma x_i = 1$ and $0 \leq x_i \leq 1$, where A defines the end-member compositions (recycled air, initial mantle Xe, Pu- and U-derived Xe), x the fraction of each end-member, and b the mantle source composition. To compute the uncertainties, a Monte Carlo technique was used whereby the estimated sample compositions were varied at random within the 1σ uncertainty and the least squares fit recomputed.

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

Received 7 November 2011; accepted 11 April 2012.

1. Allègre, C. J., Staudacher, T. & Sarda, P. Rare gas systematics: formation of the atmosphere, evolution and structure of the Earth's mantle. *Earth Planet. Sci. Lett.* **81,** 127–150 (1987).
2. Marty, B. Neon and xenon isotopes in MORB: implications for the earth-atmosphere evolution. *Earth Planet. Sci. Lett.* **94,** 45–56 (1989).
3. Holland, G. & Ballentine, C. J. Seawater subduction controls the heavy noble gas composition of the mantle. *Nature* **441,** 186–191 (2006).
4. Trieloff, M. & Kunz, J. Isotope systematics of noble gases in the Earth's mantle: possible sources of primordial isotopes and implications for mantle structure. *Phys. Earth Planet. Inter.* **148,** 13–38 (2005).
5. Gonnermann, H. M. & Mukhopadhyay, S. Preserving noble gases in a convecting mantle. *Nature* **459,** 560–563 (2009).
6. Yokochi, R. & Marty, B. Geochemical constraints on mantle dynamics in the Hadean. *Earth Planet. Sci. Lett.* **238,** 17–30 (2005).
7. Pepin, R. O. & Porcelli, D. Origin of noble gases in the terrestrial planets. *Rev. Mineral. Geochem.* **47,** 191–246 (2002).
8. Porcelli, D. & Wasserburg, G. J. Mass transfer of helium, neon, argon and xenon through a steady-state upper mantle. *Geochim. Cosmochim. Acta* **59,** 4921–4937 (1995).
9. Tolstikhin, I. & Hofmann, A. W. Early crust on top of the Earth's core. *Phys. Earth Planet. Inter.* **148,** 109–130 (2005).
10. Kurz, M. D., Jenkins, W. J. & Hart, S. R. Helium isotopic systematics of oceanic islands and mantle heterogeneity. *Nature* **297,** 43–47 (1982).
11. Brandenburg, J. P., Hauri, E. H., van Keken, P. E. & Ballentine, C. J. A multiple-system study of the geochemical evolution of the mantle with force-balanced plates and thermochemical effects. *Earth Planet. Sci. Lett.* **276,** 1–13 (2008).
12. van der Hilst, R. D. & Karason, H. Compositional heterogeneity in the bottom 1000 kilometers of Earth's mantle: toward a hybrid convection model. *Science* **283,** 1885–1888 (1999).
13. Class, C. & Goldstein, S. L. Evolution of helium isotopes in the Earth's mantle. *Nature* **436,** 1107–1112 (2005).
14. Albarede, F. Rogue mantle helium and neon. *Science* **319,** 943–945 (2008).
15. Parman, S. W. Helium isotopic evidence for episodic mantle melting and crustal growth. *Nature* **446,** 900–903 (2007).
16. Lee, C. T. A. *et al.* Upside-down differentiation and generation of a 'primordial' lower mantle. *Nature* **463,** 930–933 (2010).
17. Moreira, M., Kunz, J. & Allegre, C. Rare gas systematics in popping rock: isotopic and elemental compositions in the upper mantle. *Science* **279,** 1178–1181 (1998).
18. Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allegre, C. J. The nature of pristine noble gases in mantle plumes. *Science* **288,** 1036–1038 (2000).
19. Harrison, D., Burnard, P. G., Trieloff, M. & Turner, G. Resolving atmospheric contaminants in mantle noble gas analyses. *Geochem. Geophys. Geosyst.* **4,** 1023 (2003).
20. Ballentine, C. J., Marty, B., Lollar, B. S. & Cassidy, M. Neon isotopes constrain convection and volatile origin in the Earth's mantle. *Nature* **433,** 33–38 (2005).
21. Yokochi, R. & Marty, B. A determination of the neon isotopic composition of the deep mantle. *Earth Planet. Sci. Lett.* **225,** 77–88 (2004).
22. Honda, M. & McDougall, I. Primordial helium and neon in the Earth — a speculation on early degassing. *Geophys. Res. Lett.* **25,** 1951–1954 (1998).
23. Caffee, M. W. *et al.* Primordial noble cases from Earth's mantle: identification of a primitive volatile component. *Science* **285,** 2115–2118 (1999).
24. Pujol, M., Marty, B. & Burgess, R. Chondritic-like xenon trapped in Archean rocks: a possible signature of the ancient atmosphere. *Earth Planet. Sci. Lett.* **308,** 298–306 (2011).
25. Kunz, J., Staudacher, T. & Allegre, C. J. Plutonium-fission xenon found in Earth's mantle. *Science* **280,** 877–880 (1998).
26. Schonbachler, M., Carlson, R. W., Horan, M. F., Mock, T. D. & Hauri, E. H. Heterogeneous accretion and the moderately volatile element budget of Earth. *Science* **328,** 884–887 (2010).
27. Caro, G. Early silicate Earth differentiation. *Annu. Rev. Earth Planet. Sci.* **39,** 31–58 (2011).
28. Pahlevan, K. & Stevenson, D. J. Equilibration in the aftermath of the lunar-forming giant impact. *Earth Planet. Sci. Lett.* **262,** 438–449 (2007).
29. Torsvik, T. H., Burke, K., Steinberger, B., Webb, S. J. & Ashwal, L. D. Diamonds sampled by plumes from the core–mantle boundary. *Nature* **466,** 352–355 (2010).
30. Raquin, A., Moreira, M. A. & Guillon, F. He, Ne and Ar systematics in single vesicles: mantle isotopic ratios and origin of the air component in basaltic glasses. *Earth Planet. Sci. Lett.* **274,** 142–150 (2008).

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

Acknowledgements I thank D. Graham for supplying the Iceland sample, and K. Zahnle, C. Langmuir, S. Stewart and R. Parai for comments. Reviews by C. Ballentine, B. Marty and D. Porcelli helped to improve the paper. This work was supported by NSF grant EAR 0911363.

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 the author (sujoy@eps.harvard.edu).

METHODS

Mass spectrometry. Five separate analyses were performed on single large pieces (~2.2–3.5 g) of the DICE 10 basaltic glass from the Reykjanes Peninsula in Iceland[18] by step-crushing. The glass piece was loaded into the vacuum crusher, baked at ~90–100 °C for 18 h and then pumped for an additional 10–14 days until blanks were stable and low. To release magmatic volatiles, the glass was crushed in a step-wise manner with a hydraulic ram. The released gases were purified by sequential exposure to hot and cold SAES getters and a small split of the gas was let into a quadrupole mass spectrometer to determine the Ar abundance and an approximate $^{40}Ar/^{36}Ar$ ratio. The noble gases were then trapped on a cryogenic cold-finger. He was separated from Ne at 32 K and let into the Nu Noblesse mass spectrometer. The measurements were carried out at 200 µA trap current and an electron accelerating voltage of 60 eV.

After the He measurement was completed, the cryogenic trap was warmed to 74 K to separate Ne from Ar. Ne was measured in multi-collection mode using three discrete dynode multipliers. ^{21}Ne was measured on the axial multiplier fitted with an energy filter, while ^{20}Ne and ^{22}Ne were measured on the low and high mass multipliers, respectively. For ^{20}Ne beams larger than 100,000 counts per second (~3.9×10^{-10} cm^3 of ^{20}Ne), ^{20}Ne was measured on a Faraday cup while ^{21}Ne and ^{22}Ne were measured on the low mass multiplier in single collection mode. An automated liquid nitrogen trap was used to keep the Ar and CO_2 backgrounds low and isobaric interferences from doubly-charged Ar and CO_2 were corrected for. The Ar^{2+}/Ar^{+} ratios for the three sets of Ne measurements were 0.03 ± 0.002, 0.034 ± 0.003 and 0.031 ± 0.003, while the corresponding $CO_2{}^{2+}/CO_2{}^{+}$ ratios were 0.0039 ± 0.001, 0.0035 ± 0.0008 and 0.0045 ± 0.0005, respectively. No significant variations were observed in these ratios as a function of Ar, CO_2 or H_2 partial pressure in the mass spectrometer, and the Ar^{2+} and $CO_2{}^{2+}$ corrections for all step crushes were <1%.

Following the Ne measurement, the cold-finger was warmed to 185 K to release Ar, and depending upon the Ar abundance and approximate isotope ratio previously determined by the quadrupole mass spectrometer, a fraction of the gas was let into the mass spectrometer for more precise isotope ratio determination. The Ar isotopes were measured in multi-collection mode with ^{40}Ar on the Faraday cup, and ^{38}Ar and ^{36}Ar were measured on the axial and low mass multipliers, respectively.

Kr was not measured and to separate Kr from Xe, the cold-finger was warmed to 210 K to release Kr which was pumped away. The cold-finger was then warmed to 340 K to release all of the Xe. Xenon isotopes were measured in a combination of multi-collection and peak-jumping mode in the following five steps: $^{126\text{-}132}Xe$, $^{128\text{-}134}Xe$, $^{124\text{-}130\text{-}136}Xe$, ^{129}Xe, ^{131}Xe. ^{124}Xe and ^{126}Xe are the two rarest Xe isotopes, and since sufficient time was not spent counting these isotopes, they are not reported here.

The total procedural blanks for ^{4}He, ^{22}Ne, ^{36}Ar and ^{130}Xe before starting the crush were (in cm^3) $\leq 2 \times 10^{-11}$, $\leq 3 \times 10^{-14}$, $\leq 4 \times 10^{-13}$ and $\leq 1.5 \times 10^{-16}$, respectively. Blanks were monitored during the course of the step-crushes. With the exception of the He blank, which increased by up to a factor of 2–4, Ne, Ar and Xe blanks were either stable or decreased with progressive sample crushing. Ne, Ar and Xe blank isotopic ratios were statistically indistinguishable from atmosphere, and because the sample gases are a mixture of mantle and air, no blank corrections were applied to any of the step crushes. Mass discrimination for Ne, Ar and Xe isotope ratios was determined with air standards and instrumental drift was monitored through sample-standard bracketing with additional standards run overnight.

Hyperbolic fits. Least-squares hyperbolic fits were computed using the curve fitting tool in Matlab. Since the fits were constrained to go through atmospheric composition, the data were fitted with an equation of the form $ax + bxy + cy = 0$.

Deconvolving U- from Pu-derived fissiogenic Xe. Before discussing the deconvolution of Pu- to U-derived fission Xe, I note that Iceland and the MORB source have different fission isotopic compositions that cannot be related to each other solely through addition of atmospheric Xe. In ^{130}Xe-normalized fission isotope spaces (such as $^{136}Xe/^{130}Xe$ versus $^{132}Xe/^{130}Xe$), the Iceland sample always defines a steeper slope than MORBs when the regressions are done using individual steps (Supplementary Table 2) or using the weighted mean composition of the sample (Supplementary Fig. 5). Thus, irrespective of the exact proportions of the Pu- to U-derived fission Xe, Supplementary Table 3 and Supplementary Fig. 5 demonstrate a difference in Pu- to U-derived fission Xe between MORBs and the Iceland plume.

To deconvolve U- from Pu-derived fissiogenic Xe, five Xe isotopes were used ($^{130,131,132,134,136}Xe$). In mantle-derived basalts, vesicles have different degrees of post-eruptive atmospheric Xe contamination. To demonstrate that differences between Iceland and MORB fission Xe isotopes are robust, two different methods were used to determine the mantle fission Xe isotopic ratios free of post-eruptive air contamination.

First, the weighted means of the fissiogenic isotope ratios ($^{130}Xe/^{132}Xe$, $^{131}Xe/^{132}Xe$, $^{134}Xe/^{132}Xe$, $^{136}Xe/^{132}Xe$) were calculated from the 51 Icelandic measurements (Supplementary Table 7). The atmospheric $^{130,131,134,136}Xe/^{132}Xe$ ratios were then linearly projected to a mantle $^{129}Xe/^{132}Xe$ ratio of 1.032 (Supplementary Fig. 4) through the weighted average of the Iceland measurements. For MORBs, fission isotopes are reported normalized to ^{130}Xe ($^{131,132,134,136}Xe/^{130}Xe$)[4]. Hence, the weighted mean of the ^{130}Xe normalized ratios were computed. The atmospheric $^{131,132,134,136}Xe/^{130}Xe$ ratios were then linearly projected to a mantle $^{129}Xe/^{130}Xe$ ratio of 7.8 (refs 3, 17) through the weighted average of the MORB measurements, and the ^{132}Xe-normalized mantle source fission ratios were subsequently computed. These compositions represent Iceland-1 and MORB-1 in Supplementary Tables 3 and 4. Note that to process both data sets the same way, the weighted means for Iceland and MORBs were calculated without eliminating any data points.

Second, for Iceland and MORBs, the $^{136}Xe/^{130}Xe$ ratios in the individual steps were regressed against the $^{129}Xe/^{130}Xe$ ratios. Because one is interested in the slope of the best fit lines, data points were translated such that the atmospheric composition was at (0,0). The y data were then scaled by the square root of the ratio of the variance in x to the variance in y so as to put x and y on the same scale and fitted with an equation of the form $y = mx$. The x and y error-weighted best fit slope was computed by minimizing the value of χ^2. The uncertainty in the slope was calculated with a Monte Carlo method; the x and y data were varied at random, the best-fit line recomputed and the 67% confidence limit on the best fit line determined subsequently. The best fit line and the uncertainty in the slope were then scaled back to the original coordinate system.

From the slope and uncertainty in the slope, the mantle $^{136}Xe/^{130}Xe$ ratio, along with its uncertainty, were calculated for mantle $^{129}Xe/^{130}Xe$ ratios of 6.98 and 7.8 for Iceland and MORB sources, respectively (Fig. 3). $^{131}Xe/^{130}Xe$, $^{132}Xe/^{130}Xe$ and $^{134}Xe/^{130}Xe$ ratios were then calculated by regressing against $^{136}Xe/^{130}Xe$. The ^{132}Xe-normalized fission ratios were subsequently computed and errors were propagated through each step. These compositions represent Iceland-2 and MORB-2 in Supplementary Tables 3 and 4. To investigate whether inclusion of some of the less precise measurements affect the fission deconvolution, the above analyses were re-done using a filtered data set. To use the same filtering criteria for both the MORB and Iceland data set, only data points with $^{136}Xe/^{130}Xe$ distinct from the atmospheric composition at the 2σ level and with a relative error of <1% were selected. Such filtering, however, does not affect the deconvolution significantly.

Following determination of the mantle source composition, the least-squares solution to the system $Ax = b$ was found with the following additional constraints: $\Sigma x_i = 1$ and $0 \leq x_i \leq 1$ (also see ref. 23). Here, A is the matrix that defines the composition of the end-members, x defines the fraction of each component, and b is the matrix with the sample composition. To compute the uncertainties, a Monte Carlo technique was used whereby the estimated sample (MORB and Iceland) compositions were varied at random within the 1σ uncertainty and the least squares fit recomputed using the new values. For all simulations, it was verified that convergence to a minimum was achieved.

In the present day mantle, $^{131,132,134,136}Xe$ isotopes reflect a mixture of the initial mantle Xe, Pu- and U-produced fission Xe, and subducted atmospheric Xe, if any. The initial Xe isotopic composition of the mantle is subject to debate. Hence, three possible initial Xe isotopic compositions were investigated: U-Xe, solar Xe, and chondritic Xe (AVCC). U-Xe and solar Xe are almost identical, except that U-Xe has a deficiency in ^{134}Xe and ^{136}Xe.

Supplementary Table 4 indicates that for all three starting mantle compositions and for both the techniques used to calculate Xe isotopic composition of the mantle source, the Iceland plume has significantly higher Pu-derived fission Xe than the MORB source. Note that for the Iceland sample, the deconvolution yields the same results within error using the two techniques (Iceland-1 and Iceland-2). For the MORB data, the deconvolution using MORB-2 gives a higher proportion of Pu-produced Xe. However, the MORB-2 composition is determined from least squares regression and, compared to the weighted mean, least squares regression may be more prone to outliers as it depends on the residual sum of squares. Nonetheless, the results in Supplementary Table 4 demonstrate that when the Iceland and MORB data sets are processed the same way, the Iceland sample, compared to MORBs, always has approximately a factor of two or higher Pu- to U-produced fission Xe. Note that for a mantle starting composition of either U-Xe, solar Xe or AVCC Xe, substantial injection of atmospheric Xe back into the mantle is required (Supplementary Table 4). The recycling of atmospheric Xe could have happened all through Earth's history associated with plate subduction or very early on in Earth's history. The requirement of subducting atmospheric Xe, however, does not change the argument that OIB Xe composition cannot be produced solely through mixing subducted Xe with MORB Xe.

The I–Pu–Xe system can constrain rates of volatile loss of a mantle reservoir through the ^{129*}Xe/^{136*}Xe$_{\mathrm{Pu}}$ ratio, where ^{129*}Xe is the decay product of ^{129}I decay and ^{136*}Xe$_{\mathrm{Pu}}$ is ^{136}Xe produced from ^{244}Pu fission. ^{129}Xe in the mantle is a mixture of initial ^{129}Xe, recycled atmospheric ^{129}Xe and radiogenic ^{129}Xe (^{129}Xe*) produced from decay of ^{129}I. Once the fractions of initial ^{132}Xe and ^{132}Xe from recycled air are calculated, the ^{129}Xe/^{132}Xe ratio of the mantle source (Supplementary Fig. 4) defines the fraction of ^{129}Xe produced from decay of ^{129}I (^{129*}Xe). Combining this with the Pu-derived fission Xe (Supplementary Table 4) yields the ^{129}Xe*/^{136}Xe$_{\mathrm{Pu}}$* ratio. Note that the ^{129}Xe*/^{136}Xe$_{\mathrm{Pu}}$* ratio in the Iceland mantle is lower than in the MORB source for all three starting mantle compositions when the data sets are processed in the same way (Supplementary Table 4).

doi:10.1038/nature11118

A global synthesis reveals biodiversity loss as a major driver of ecosystem change

David U. Hooper[1], E. Carol Adair[2,3], Bradley J. Cardinale[4], Jarrett E. K. Byrnes[2], Bruce A. Hungate[5], Kristin L. Matulich[6], Andrew Gonzalez[7], J. Emmett Duffy[8], Lars Gamfeldt[9] & Mary I. O'Connor[2,10]

Evidence is mounting that extinctions are altering key processes important to the productivity and sustainability of Earth's ecosystems[1–4]. Further species loss will accelerate change in ecosystem processes[5–8], but it is unclear how these effects compare to the direct effects of other forms of environmental change that are both driving diversity loss and altering ecosystem function. Here we use a suite of meta-analyses of published data to show that the effects of species loss on productivity and decomposition—two processes important in all ecosystems—are of comparable magnitude to the effects of many other global environmental changes. In experiments, intermediate levels of species loss (21–40%) reduced plant production by 5–10%, comparable to previously documented effects of ultraviolet radiation and climate warming. Higher levels of extinction (41–60%) had effects rivalling those of ozone, acidification, elevated CO_2 and nutrient pollution. At intermediate levels, species loss generally had equal or greater effects on decomposition than did elevated CO_2 and nitrogen addition. The identity of species lost also had a large effect on changes in productivity and decomposition, generating a wide range of plausible outcomes for extinction. Despite the need for more studies on interactive effects of diversity loss and environmental changes, our analyses clearly show that the ecosystem consequences of local species loss are as quantitatively significant as the direct effects of several global change stressors that have mobilized major international concern and remediation efforts[9].

A variety of global changes are driving rates of species extinction that greatly outpace background rates in the fossil record[10,11]. If these trends continue, projections suggest that within 240 years Earth may face the sixth mass extinction[12]. Such projections have prompted hundreds of experiments examining how different components of biodiversity affect ecosystem processes that sustain the provisioning of goods and services to society. Syntheses of these experiments have made it clear that plant biodiversity loss will reduce plant production and alter decomposition[5,6]. However, it is uncertain how the sizes of these effects compare with the direct effects of other types of environmental change, such as changing atmospheric composition, climate warming and nutrient pollution, that also threaten ecosystem functioning[13–15]. This uncertainty has generated wide-ranging speculation about how strongly biodiversity loss might affect humanity[16,17].

Here we report the results of a large data synthesis in which we compared the effects of species loss against other drivers of environmental change. We focus on primary production and decomposition because these major biological processes influence carbon storage and other ecosystem services, and illustrate the breadth of sensitivity of ecosystem processes to changes in species richness[2,6,18]. We took two approaches in our analyses. First, we statistically summarized existing meta-analyses that have estimated the mean effect size of experimental manipulations of a variety of environmental changes on primary production (biomass production by plants) and decomposition (mass loss of plant litter) in a variety of ecosystems around the world (Tables 1 and 2). We compared these environmental effect sizes to the estimated effects of species loss derived from a database we constructed using 192 peer-reviewed publications on experiments that manipulated species richness and examined the effects on ecosystem processes (see Methods). This approach allows comparison among a wide range of environmental changes, but has the limitation that it evaluates the effects of environmental and diversity changes measured by different researchers using different organisms and ecosystems. To complement our summary of meta-analyses, we also summarized the results of 16 experiments that simultaneously manipulated plant species richness in factorial combination with some other environmental change (elevated CO_2, nutrient pollution, etc). Although a far smaller data set, analysis of factorial experiments allowed two additional comparisons: (1) effect sizes of diversity loss versus other environmental changes, within experiments focusing on identical ecosystems; and (2) effect sizes of diversity loss under current versus projected environmental conditions. We assessed a breadth of projections of local species loss because estimates vary widely for magnitudes of global species extinctions (Supplementary Table 1). Similarly, species losses at local scales most relevant to biodiversity and ecosystem functioning (BEF) experiments and ecosystem services (m^2 to watersheds) probably do not bear a one-to-one relationship with global extinctions (complete loss of a species from the planet) and may respond nonlinearly to multiple environmental changes[10,19,20].

Our analyses suggest that biodiversity loss in the 21st century could rank among the major drivers of ecosystem change. Experiments to date have shown that effects of plant species richness on biomass production are nonlinear and saturating (Fig. 1). Our analysis suggests that in areas where local species loss this century falls within the lower range of projections (1–20%), negligible effects on biomass production will result, and changes in species richness will rank low relative to the effects projected for other environmental changes (Fig. 1 and Table 1). Where actual losses fall within intermediate projections (21–40%), however, species loss is expected to reduce biomass production by 5–10% below the most diverse mixtures (based on exponentiation of log response ratios (LRR): $e^{-0.05} = 0.951$, $e^{-0.107} = 0.898$). This effect is comparable in magnitude to the effects of ultraviolet radiation and climate warming on plant production (Fig. 1 and Table 1).

Where losses fall within higher projections of extinction (41–60%), the effects of species loss rank with those of many other drivers of environmental change, such as warming, ozone and acidification (Fig. 1). The mid-point of this range, fifty per cent species loss, is a benchmark at the upper end of 21st century projections of global extinctions, but is a common estimate at the local scale in heavily-affected

[1]Department of Biology, Western Washington University, Bellingham, Washington 98225-9160, USA. [2]National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300, Santa Barbara, California 93101, USA. [3]Rubenstein School of Environment and Natural Resources, Aiken Center, University of Vermont, Burlington, Vermont 05405, USA. [4]School of Natural Resources & Environment, University of Michigan, Ann Arbor, Michigan 48109-1041, USA. [5]Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona 86011, USA. [6]Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, USA. [7]Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montréal, Québec H3A 1B1, Canada. [8]Virginia Institute of Marine Science, College of William and Mary, Gloucester Point, Virginia 23062, USA. [9]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Göteborg, Sweden. [10]Department of Zoology, University of British Columbia, 2370-6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada.

Table 1 | Effects of species richness and environmental changes on primary productivity for the broad meta-analysis and factorial diversity crossed with environment experiments.

	Broad meta-analysis				Factorial experiments			
Factor	N_S, N_{obs}	LRR	LCI	UCI	N_S, N_{obs}	LRR	LCI	UCI
Primary producer diversity								
50% loss	60, 145	−0.144	−0.175	−0.112	10, 15	−0.168	0.104	−0.439
Avg. mono.	73, 299	−0.332	−0.378	−0.285	16, 30	−0.458	−0.259	−0.658
Best mono.	62, 241	0.159	0.116	0.203	13, 29	−0.136	0.067	−0.338
Other factors								
Acidif.	1, 12	−0.186	−0.342	−0.020				
+Ca	1, 31	0.351	−0.105	0.820	1, 1	0.256	−0.781	1.293
+CO_2	6, 3076	**0.217**	**0.207**	**0.227**	3, 5	0.070	−0.400	0.539
Drought	1, 20	−0.616	−0.892	−0.342	3, 5	0.215	−0.255	0.686
+N	6, 2895	**0.310**	**0.192**	**0.428**	+N-med 3, 13	0.155	−0.155	0.466
					+N-high 2, 21	0.434	0.165	0.703
+N +CO_2	1, 252	**0.694**	**0.622**	**0.766**				
+N +P	1, 941	**0.964**	**0.894**	**1.034**	5, 8	0.586	0.215	0.958
+Ozone	4, 2162	**−0.149**	**−0.161**	**−0.137**				
+P	2, 766	**0.239**	**0.175**	**0.302**	1, 1	1.216	0.177	2.254
Plant inv.	1, 144	**0.514**	**0.447**	**0.581**				
+Ultraviolet	2, 432	**−0.082**	**−0.107**	**−0.057**				
Warming	1, 1064	0.116	0.078	0.154	1, 1	−0.529	−1.568	0.510

N_S, total number of studies—references listed in Supplementary Table 2, except for diversity effects for the broad meta-analysis which come from the database on biodiversity and ecosystem functioning (BEF) experiments[6]; N_{obs}, total number of observations across all meta-analyses or experiments; LRR, log response ratio; LCI, lower 95% confidence interval; UCI, upper 95% confidence interval. Bold values indicate bootstrapped mean LRRs and confidence intervals in the broad meta-analysis (see Supplementary Fig. 2). Treatment factors: 50% loss; Avg. mono., average monoculture; Best mono., best monoculture (see Methods for calculation of LRRs). In all of these, negative values indicate that species loss causes a decline in productivity rates. Acidif., acidification; Ca, calcium; CO_2, carbon dioxide; N, nitrogen (for terrestrial N addition, 'low' rates were $\leq 3\,g\,m^{-2}$ (no such factorial experiments), medium (med.) were > 3 and $\leq 15\,g\,m^{-2}$, and high were $> 15\,g\,m^{-2}$); P, phosphorus; Plant inv., plant invasion.

landscapes that have experienced >90% habitat loss[21]. A 50% species loss is expected to reduce biomass production by an average of 13% ($e^{-0.144}$; Table 1), an effect consistent across terrestrial, freshwater and marine ecosystems (Supplementary Fig. 1). For comparison, elevated CO_2 experiments have produced greater overall magnitudes of changes in biomass (+24%). This average, however, combines studies performed in diverse natural systems as well as in agricultural monocultures. Experiments performed in multi-species communities have shown the effects of CO_2 on production of +12–13% (ref. 22, Supplementary Fig. 2, Tot_{multi} under Elevated CO_2)—on par with projected effects of 50% species loss. Similarly, the average effect of nitrogen (N) on plant biomass production depended on N addition rates. Rates of N addition similar to intensive agricultural fertilization had effects on production (+54% for factorial experiments) that were greater than those of intermediate or high levels of species loss. However, the magnitude of effects of high species loss on production was comparable to those of intermediate (+17%; Table 1) or low (rare in terrestrial experiments, Supplementary Fig. 3)[23] rates of N addition. Thus, the magnitude of the effects of high species loss on production also seems to be comparable to those of increased nitrogen deposition, a well-recognized environmental problem[24,25].

To rival the environmental changes that have had the greatest documented effect on primary production (for example, heavy nutrient pollution, invasive species, drought), species loss would need to exceed that of prior mass extinctions (≥75% species loss). This scenario is unlikely to be realized globally in the coming century, but could occur for some types of organisms (for example, vertebrates) within 240–540 years if current rates of extinction continue[12]. It may also occur for a variety of organisms at local scales where human activities heavily affect land use. If such a scenario were realized for plants, biomass production in natural systems would be expected to decline by an average of one third, exceeding the effects of all other environmental changes except invasive species, drought and interactions among multiple pollutants (N, P, CO_2) applied in combination (Fig. 1). However, uncertainty around the effects of extinction grows large as the fraction of species loss increases, in part because the identity and biological traits of surviving species have an increasingly large effect on biomass production. The importance of species identity is most apparent from examining extreme cases where experiments reduce diverse communities to single species. Whereas the average effect of reducing diversity to a monoculture is a 28% loss in production, the distribution ranged from −68% to +62% of production compared to the most diverse mixture. Similarly, if one can conserve the most productive monocultures, these outperformed the most diverse mixtures by an average 17%, but ranged from −40% to +132% (Table 1; Supplementary Discussion, Productivity section). These values span the range of plausible effects for nearly all other environmental changes, and underscore the large variation in possible outcomes of extinction that can result from functional differences among species and ecosystems[4].

Table 2 | Effects of species richness and environmental changes on decomposition from the broad meta-analysis.

Factor	N_S, N_{obs}	LRR	LCI	UCI	Response variable
Litter diversity					
50% loss	24, 39	0.023	−0.062	0.108	Mixed
Avg. mono.	31, 67	0.034	−0.130	0.199	Mixed
Best mono.	21, 39	0.266	0.153	0.378	Mixed
Consumer diversity					
50% loss	22, 52	−0.074	−0.155	0.008	Mixed
Avg. mono.	22, 55	−0.235	−0.359	−0.111	Mixed
Best mono.	19, 49	−0.056	−0.190	0.077	Mixed
Other factors					
+CO_2	1, 101	−0.020	−0.041	0.010	Mass loss
Eutrophication	1, 6	0.250	−0.180	0.660	Microbial breakdown rate
Plant inv.	2, 62	**0.729**	**0.677**	**0.782**	Decomposition rate, microbial breakdown rate
Acidification	1, 5	−0.830	−1.600	−0.520	Microbial breakdown rate
+N	2, 520	**−0.023**	**−0.046**	**0.000**	Decomposition rate, mass loss

Abbreviations as in Table 1. Response variable, biomass types and response variables used to calculate LRR value. Mixed: microbial respiration rate, decomposition rate, mass loss rate, feeding rate. Bold values indicate bootstrapped mean LRRs and confidence intervals.

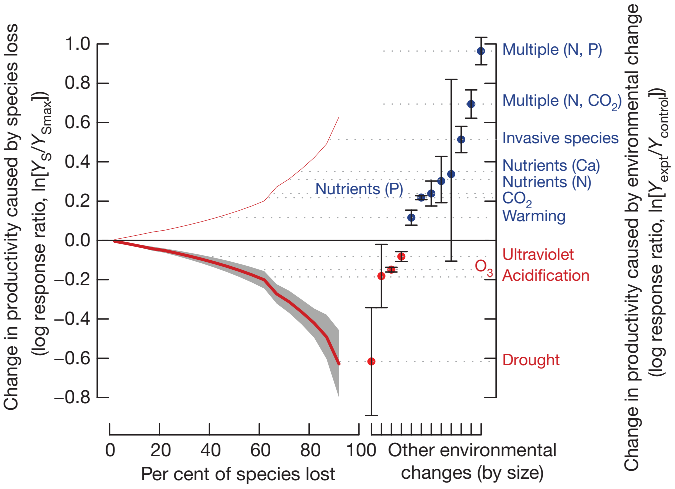


Figure 1 | Changes in primary production as a function of per cent local species loss. Effects of species loss on primary production from 62 studies (379 observations). Thick red line, lower productivity as species richness decreases; grey bands and black error bars, 95% confidence intervals. The thin red line shows the inverse of the thick red line to allow comparison of effect magnitudes with environmental changes with positive effects. Dotted grey lines show the mean effect of each environmental change for comparison with the effect of richness. Right axis, effects of other environmental changes. Blue is for increases and red for decreases in productivity (Table 1 and Supplementary Table 2).

Analysis of experiments that manipulated species richness in factorial combination with other environmental changes generally reinforced our conclusions from the broader meta-analysis (Table 1; Supplementary Discussion, Productivity section). Within intermediate projections of species loss (21–40%), the effects of species loss on plant biomass production equalled or exceeded the effects of elevated CO_2, and rivalled the effect of drought (Table 1 and Supplementary Fig. 4). Interactions between species loss and environmental changes are important for understanding net effects on ecosystem processes, because both will often occur simultaneously (environmental changes rank among the major drivers of species loss[13,26,27]). We compared the effects of species loss (average monoculture metric) under experimental conditions with the effects of species loss under control environmental conditions to investigate potential interactions between these drivers of ecosystem processes (Supplementary Figs 5 and 6). The available evidence indicates that diversity effects were independent of many environmental changes (interactions were not detectably different from zero). The exception was N addition, which led to smaller average effects of diversity under elevated than control conditions ($P = 0.043$, when weighted by n; Supplementary Fig. 5). However, the scarcity of studies meant we found three or fewer experiments for any change other than fertilization. Clearly, this is a critical topic for future research.

Both environmental changes and species loss had smaller effects on decomposition than on production. However, the effects of consumer species loss on decomposition were comparable to the effects of some major forms of environmental change. Loss of litter consumer richness reduced decomposition rates by ~8% for mid-ranges of projected extinction, giving rise to effects that were comparable in magnitude to elevated CO_2 (−2%) and nitrogen pollution (−2%), although smaller than the effects of multiple nutrient addition in aquatic systems, acidification and plant invasion (Fig. 2 and Table 2). The effects of consumer loss were more pronounced and consistent in freshwater, where the majority of experiments have taken place[6], than in terrestrial ecosystems (−12% versus −7%, respectively, for a 50% loss scenario, Supplementary Fig. 1). In contrast to the effects of consumer species loss, loss of litter diversity did not alter average rates of decomposition (Fig. 2 and Table 2; Supplementary Discussion, Decomposition section). Because species loss reduced primary productivity more than decomposition, future species loss could limit the capacity for carbon uptake and storage in the biosphere[18].

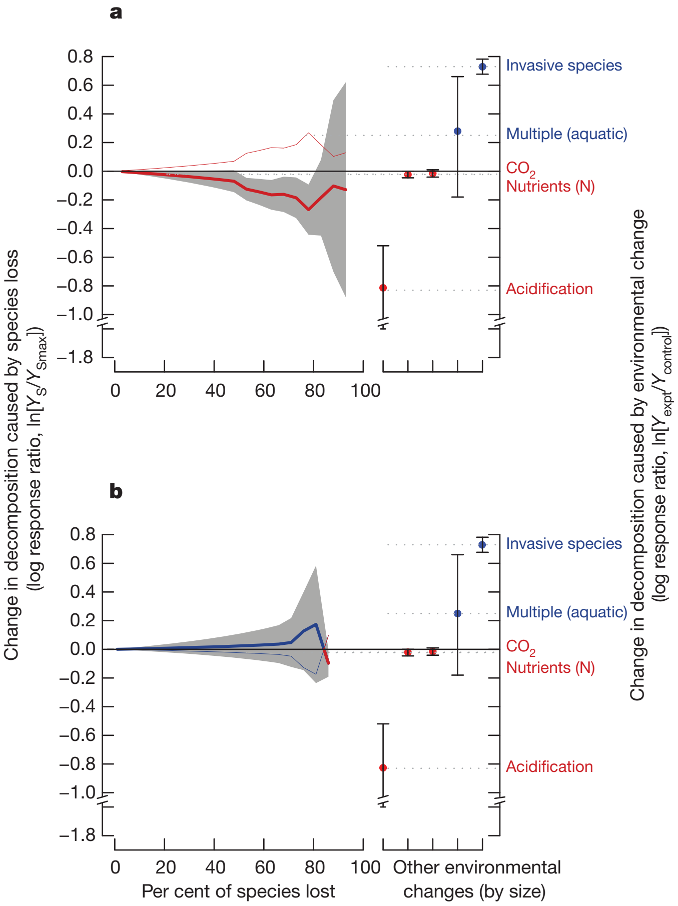


Figure 2 | Changes in decomposition as a function of per cent local species loss. a, Effects of detrital consumer diversity on decomposition from 19 studies (54 observations). **b**, Effects of plant litter diversity on decomposition from 22 studies (60 observations). Thick red lines, slower decomposition rates as species richness decreases; thick blue lines, higher decomposition rate as species richness decreases; grey bands and black error bars, 95% confidence intervals. Thin coloured lines, dotted grey lines, axes and colour coding as in Fig. 1. See also Table 2 and Supplementary Table 3.

In summary, we have shown that species loss ranks among the major drivers of primary production and decomposition—key processes involved in the carbon cycle and the provisioning of many ecosystem services[11,18]. Refining these estimates for key ecosystem services will require a better understanding of how realistic extinction scenarios interact with other forms of environmental change in influencing multiple ecosystem processes[4,16,26]. Even so, the range of effects caused by species loss spanned the range of plausible outcomes for nearly all other drivers of environmental change. And the average effects of local extinction were comparable in magnitude to numerous other global change stressors that have already mobilized major international concern and remediation efforts. As such, our study provides a quantitative basis for integrating consequences of species loss into assessments to be conducted by the Intergovernmental Science Policy Platform for Biodiversity and Ecosystem Services[28].

METHODS SUMMARY

To quantify how species loss affects primary production and decomposition, we used the database of ref. 6 that summarized 192 studies (574 experiments) through 2009 that manipulated species richness and measured the effects on ecosystem processes. We extracted experiments describing (1) how species richness of primary producers influenced producer biomass and (2) how primary producer or consumer richness affected decomposition of litter. We then calculated two log response ratios (LRRs) for each experiment: $\ln(Y_{\mathrm{Avemono}}/Y_{\mathrm{Smax}})$ and $\ln(Y_{\mathrm{Bestmono}}/Y_{\mathrm{Smax}})$ where Y_{Smax} was production or decomposition in the most diverse mixture in an experiment, Y_{Avemono} is the average value of the monocultures, and Y_{Bestmono} is the value in the most productive or fastest decomposing monoculture (for considerations, see Methods and Supplementary Discussion[29]). When possible, we also fit data from each study to a power function $\ln(Y_S/Y_{\mathrm{Smax}}) = \mathrm{a} + \mathrm{b} \times \ln(S)$. Parameter estimates were used to produce the nonlinear species loss curves in Figs 1, 2 and Supplementary Fig. 4, and the 50% loss scenario in Tables 1 and 2.

To gather data on how other forms of environmental change affect production and decomposition, we searched the ISI Web of Science for published meta-analyses (see Methods). From each paper or publicly available data set[23], we extracted response ratios ($\mathrm{RR} = Y_{\mathrm{expt}}/Y_{\mathrm{control}}$, where Y_{expt} is the response variable in the experimental treatment), number of studies, and estimates of variance (Data Thief III, Version 1.5). We calculated the overall mean LRR and 95% confidence interval for each treatment via bootstrapping using skew normal distributions[30].

We also identified 16 factorial experiments to directly compare productivity LRRs for diversity to other forms of environmental change in the same experiment. Where manipulations either reduced or enhanced resources, we changed the sign of the LRRs to allow comparison to the broader meta-analysis. We analysed LRRs using mixed models in SYSTAT v.12, with environmental change as a fixed effect and study as a random effect.

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

Received 9 January; accepted 13 April 2012.
Published online 2 May 2012.

1. Loreau, M., Naeem, S. & Inchausti, P. *Biodiversity and Ecosystem Functioning: Synthesis and perspectives* (Oxford Univ. Press, 2002).
2. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75,** 3–35 (2005).
3. Tilman, D. Ecological consequences of biodiversity: a search for general principles. *Ecology* **80,** 1455–1474 (1999).
4. Wardle, D. A., Bardgett, R. D., Callaway, R. M. & Van der Putten, W. H. Terrestrial ecosystem responses to species gains and losses. *Science* **332,** 1273–1277 (2011).
5. Balvanera, P. *et al.* Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol. Lett.* **9,** 1146–1156 (2006).
6. Cardinale, B. J. *et al.* The functional role of producer diversity in ecosystems. *Am. J. Bot.* **98,** 572–592 (2011).
7. Stachowicz, J. J., Bruno, J. F. & Duffy, J. E. Understanding the effects of marine biodiversity on communities and ecosystems. *Annu. Rev. Ecol. Evol. Syst.* **38,** 739–766 (2007).
8. Perrings, C. *et al.* Ecosystem services, targets, and indicators for the conservation and sustainable use of biodiversity. *Front. Ecol. Environ* **9,** 512–520 (2011).
9. IPCC. *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Core Writing Team, Pachauri, R. K. & Reisinger, A.) (IPCC, 2007).
10. Sala, O. E. *et al.* Global biodiversity scenarios for the year 2100. *Science* **287,** 1770–1774 (2000).
11. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis* (World Resources Institute, 2005).
12. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471,** 51–57 (2011).
13. Chapin, F. S. III *et al.* Consequences of changing biodiversity. *Nature* **405,** 234–242 (2000).
14. Grace, J. B. *et al.* Does species diversity limit productivity in natural grassland communities? *Ecol. Lett.* **10,** 680–689 (2007).
15. Paquette, A. & Messier, C. The effect of biodiversity on tree productivity: from temperate to boreal forests. *Glob. Ecol. Biogeogr.* **20,** 170–180 (2011).
16. Srivastava, D. S. & Vellend, M. Biodiversity-ecosystem function research: is it relevant to conservation? *Annu. Rev. Ecol. Evol. Syst.* **36,** 267–294 (2005).
17. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).
18. Díaz, S., Wardle, D. A. & Hector, A. in *Biodiversity, Ecosystem Functioning, and Human Wellbeing: An Ecological and Economic Perspective* (eds Naeem, S. *et al.*) Ch. 11 149–166 (Oxford Univ. Press, 2009).
19. Pereira, H. M. *et al.* Scenarios for global biodiversity in the 21st century. *Science* **330,** 1496–1501 (2010).
20. Brook, B. W., Sodhi, N. S. & Bradshaw, C. J. A. Synergies among extinction drivers under global change. *Trends Ecol. Evol.* **23,** 453–460 (2008).
21. Ewers, R. M. & Didham, R. K. Confounding factors in the detection of species responses to habitat fragmentation. *Biol. Rev. Camb. Philos. Soc.* **81,** 117–142 (2006).
22. Wang, X. Effects of species richness and elevated carbon dioxide on biomass accumulation: a synthesis using meta-analysis. *Oecologia* **152,** 595–605 (2007).
23. Elser, J. J. *et al.* Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol. Lett.* **10,** 1135–1142 (2007).
24. Vitousek, P. M. *et al.* Human alteration of the global nitrogen cycle: sources and consequences. *Ecol. Appl.* **7,** 737–750 (1997).
25. Carpenter, S. R. *et al.* Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl.* **8,** 559–568 (1998).
26. Tylianakis, J. M., Didham, R. K., Bascompte, J. & Wardle, D. A. Global change and species interactions in terrestrial ecosystems. *Ecol. Lett.* **11,** 1351–1363 (2008).
27. Suding, K. N. *et al.* Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. *Glob. Change Biol.* **14,** 1125–1140 (2008).
28. Larigauderie, A. & Mooney, H. A. The Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services: moving a step closer to an IPCC-like mechanism for biodiversity. *Curr. Opin. Environ. Sust.* **2,** 9–14 (2010).
29. Schmid, B., Hector, A., Saha, P. & Loreau, M. Biodiversity effects and transgressive overyielding. *J. Plant Ecol.* **1,** 95–102 (2008).
30. Johnson, N. J. Modified *t* tests and confidence intervals for asymmetrical populations. *J. Am. Stat. Assoc.* **73,** 536–544 (1978).

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

Acknowledgements We thank the authors of the original data sets, both published and unpublished, who shared data with us. We thank the National Center for Ecological Analysis and Synthesis (NCEAS) for funding the workshop ‘‘Biodiversity and the Functioning of Ecosystems: Translating Results from Model Experiments into Functional Reality’’. Support for NCEAS comes from University of California Santa Barbara and the National Science Foundation. J.E.K.B., E.C.A. and M.I.O. had NCEAS post-doctoral fellowships, J.E.D. had support from NSF OCE-1031061; B.J.C. had support from NSF DEB-1046121; L.G. was supported by grant 621-2009-5457 from the Swedish Research Council VR; A.G. is supported by the Canada Research Chair Program and NSERC.

Author Contributions All authors contributed to the design of the study, data interpretation and manuscript editing; B.J.C. and K.L.M. developed the database of biodiversity and ecosystem functioning experiments; D.U.H., E.C.A., J.E.K.B., B.J.C. and K.L.M. collected additional data and performed statistical analyses. E.C.A., J.E.K.B., B.J.C., B.A.H. and D.U.H. drafted the figures and D.U.H. wrote the initial draft.

Author Information The biodiversity and ecosystem functioning database is deposited with the National Center for Ecological Analysis and Synthesis (http://knb.ecoinformatics.org/knb/metacat/nceas.984/nceas). 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 D.U.H. (hooper@biol.wwu.edu).

METHODS

Biodiversity and ecosystem functioning database. To quantify the effects of species loss on biomass production and decomposition, we used the data set of ref. 6. This data set summarizes 192 peer-reviewed papers published through 2009 reporting results from 574 experiments that manipulated species richness and measured the effects on various ecosystem processes. We extracted the subset of experiments that examined (1) how species richness of primary producers influenced producer biomass accumulation and (2) how richness of producer litter, or richness of litter consumers, affected decomposition rates (Tables 1 and 2).For multi-year studies, we used only data from the last year as this was least likely to be influenced by transient responses. For each experiment, we calculated two log response ratios (LRRs): $\ln(Y_{\mathrm{Avemono}}/Y_{\mathrm{Smax}})$ and $\ln(Y_{\mathrm{Bestmono}}/Y_{\mathrm{Smax}})$ where Y_{Smax} was production or decomposition in the most diverse mixture in an experiment, Y_{Avemono} is the average value of the monocultures, and Y_{Bestmono} is the value in the most productive or fastest decomposing monoculture. Both LRRs quantify the net effect of species loss going from the most to least diverse communities, but differ in their assumptions about the sequence of extinction. For productivity only, best monoculture values were restricted to communities where increasing plant diversity increased production (that is, where the average monoculture LRR was positive)[6]. Statistical issues may bias the effects of best monocultures[29]; because this topic is controversial, however, we use the best monoculture metric primarily to illustrate the range of potential process responses, particularly in heavily managed ecosystems.

Log ratios like those described above are frequently used to summarize diversity effect sizes[6], in part because they can be calculated for most experiments. However, these metrics represent extreme scenarios of local diversity loss that are not likely to be realized in many natural communities. Therefore, we also ran a more comprehensive analysis on the subset of experiments that included at least three levels of species richness. For these experiments, we fit the mean value of the response at each level of richness S to the power function: $\ln(Y_S/Y_{\mathrm{Smax}}) = \mathrm{a} + \mathrm{b} \times \ln(S)$. Prior meta-analyses lend much stronger support to saturating models of diversity effects (for example, Michaelis–Menten or power functions) compared to linear or exponential fits[6]. We used the power function here because it gave a good fit and provided a balance between simplicity and generality (mean $R^2 = 0.71$ for productivity and 0.30 for decomposition, compared to mean $R^2 = 0.73$ and 0.29 for Michaelis–Menten). After obtaining parameter estimates for each experiment, we calculated the effect of species loss on production and decomposition across all levels of per cent loss that we could interpolate within an individual experiment. We calculated the log response ratio $\ln(Y_S/Y_{\mathrm{Smax}})$ at 5% increments of species loss, where Y_S is the value at S species ($<S_{\mathrm{max}}$). The distribution of log ratios was estimated by bootstrapping, and means and 95% confidence intervals were plotted in Fig. 1. The 50% loss scenario in Tables 1 and 2 came from these estimates.

Meta-analysis comparison. To quantify how other forms of environmental change have an effect on production and decomposition, we collated data published in past syntheses and data analyses. These studies typically manipulated abiotic conditions consistent with accepted scenarios of environmental change for the factors at hand (for example, doubling of atmospheric CO_2, ref. 9; Supplementary Fig. 2).

Productivity: we searched ISI Web of Science for meta-analyses that examined the effects of global change factors on biomass production (search terms: [product* AND meta-analysis] OR [biomass AND meta-analysis]). Because more recent meta-analyses often have extensive reference overlap with earlier meta-analyses of the same environmental effect, we only used meta-analyses published after 2005 to maximize independence across studies. We found 18 meta-analyses summarizing 67 LRRs showing how various aspects of environmental change influence primary production in marine, freshwater and terrestrial ecosystems (Supplementary Table 2). LRRs were calculated as ln(mean treatment biomass/mean control biomass). From each meta-analysis, we extracted the LRR, number of observations, and the associated error measurement from text, tables or digitized figures (Data Thief III, Version 1.5), or calculated them directly where data were freely available[23].

We found LRR values for 12 forms of environmental change: acidification, calcium (Ca) additions, elevated CO_2, drought, plant invasion, nitrogen (N) additions, phosphorous (P) additions, N + P additions, N additions + elevated CO_2, elevated ozone, elevated ultraviolet radiation, and warming (Table 1 and Supplementary Table 2). If a treatment was represented by only one LRR value, then the reported LRR and associated confidence intervals were used in our analysis (Table 1). If a treatment was represented by more than one LRR value, we calculated the overall mean LRR and confidence interval for each treatment via bootstrapping from skew normal distributions[30]. Distributions were resampled 10,000 times to generate an overall mean and lower/upper confidence interval using the 'fGarch' package of R version 2.12.2. Bootstrapped means and confidence intervals are compared with the means and confidence intervals from the original data sources in Supplementary Fig. 2.

Decomposition: to limit reference overlap, we searched for decomposition meta-analyses published after 2000, using the search terms [decomp* AND meta-analysis] in the ISI Web of Science. We found five meta-analyses, resulting in seven LRRs of a treatment effect on decomposition in freshwater and/or terrestrial ecosystems (Table 2 and Supplementary Table 3). LRR values were available for five different treatments: acidification, elevated CO_2, plant invasion, N additions and eutrophication (multiple nutrient additions in aquatic ecosystems). We extracted data and calculated mean LRR and associated confidence intervals as described above for productivity.

Environment crossed with species richness manipulations. We complemented our summary of meta-analyses with a more focused analysis that compared the effects of species richness to the effects of other forms of environmental change when both were manipulated simultaneously in the same experiment. To do this, we extracted records from the ref. 6 database for experiments that manipulated species richness and some component of environmental change in factorial combination. We only had sufficient data to assess effects of diversity and environmental manipulations on biomass production (16 studies, Supplementary Table 2): +calcium[31], +CO_2 (refs 32–34), water availability ("drought")[35–37], nitrogen addition[31,32,38], phosphorus addition[39], multiple nutrient addition[40–44], and warming[45]. In our statistical analyses, we also included effects of light manipulation[46], although the explicit link to global environmental change is less clear for this factor, so it is not shown in figures. For each study, we calculated the suite of diversity LRRs previously described, as well as the effect of the environmental manipulations at maximum species richness (Table 1). For experiments where manipulations either reduced or enhanced resources (for example, nutrient or water availability), we changed the sign of the LRRs appropriately so that magnitudes of effects could be compared on a scale similar to environmental changes assessed in the broader meta-analysis. We analysed LRRs using mixed models in SYSTAT v.12 (SYSTAT, Inc.) with environmental change as a fixed effect and study as a random effect. We compared equally weighted results to analyses where we weighted LRRs by sample size $(n_1 \times n_2)/(n_1 + n_2)$; results were qualitatively similar, unless otherwise noted.

31. Rixen, C., Huovinen, C., Huovinen, K., Stöckli, V. & Schmid, B. A plant diversity×water chemistry experiment in subalpine grassland. *Perspect. Plant Ecol.* **10,** 51–61 (2008).
32. Reich, P. B. *et al.* Plant diversity enhances ecosystem responses to elevated CO_2 and nitrogen deposition. *Nature* **410,** 809–810 (2001).
33. Maestre, F. T. & Reynolds, J. F. Biomass responses to elevated CO_2, soil heterogeneity and diversity: an experimental assessment with grassland assemblages. *Oecologia* **151,** 512–520 (2007).
34. Stocker, R., Körner, C., Schmid, B., Niklaus, P. A. & Leadley, P. W. A field study of the effects of elevated CO_2 and plant species diversity on ecosystem-level gas exchange in a planted calcareous grassland. *Glob. Change Biol.* **5,** 95–105 (1999).
35. Mulder, C. P. H., Uliassi, D. D. & Doak, D. F. Physical stress and diversity-productivity relationships: the role of positive species interactions. *Proc. Natl Acad. Sci. USA* **98,** 6704–6708 (2001).
36. Rixen, C. & Mulder, C. P. H. Improved water retention links high species richness with increased productivity in arctic tundra moss communities. *Oecologia* **146,** 287–299 (2005).
37. Wenninger, E. J. & Inouye, R. S. Insect community response to plant diversity and productivity in a sagebrush–steppe ecosystem. *J. Arid Environ.* **72,** 24–33 (2008).
38. Wacker, L., Baudois, O., Eichenberger-Glinz, S. & Schmid, B. Diversity effects in early- and mid-successional species pools along a nitrogen gradient. *Ecology* **90,** 637–648 (2009).
39. Striebel, M., Behl, S. & Stibor, H. The coupling of biodiversity and productivity in phytoplankton communities: consequences for biomass stoichiometry. *Ecology* **90,** 2025–2031 (2009).
40. Fridley, J. D. Resource availability dominates and alters the relationship between species diversity and ecosystem productivity in experimental plant communities. *Oecologia* **132,** 271–277 (2002).
41. Lanta, V. & Leps, J. Effect of functional group richness and species richness in manipulated productivity–diversity studies: a glasshouse pot experiment. *Acta Oecol.* **29,** 85–96 (2006).
42. Smith, A. & Allcock, P. J. The influence of species diversity on sward yield and quality. *J. Appl. Ecol.* **22,** 185–198 (1985).
43. Boyer, K. E., Kertesz, J. S. & Bruno, J. F. Biodiversity effects on productivity and stability of marine macroalgal communities: the role of environmental context. *Oikos* **118,** 1062–1072 (2009).
44. von Felten, S. & Schmid, B. Complementarity among species in horizontal versus vertical rooting space. *J. Plant Ecol.* **1,** 33–41 (2008).
45. De Boeck, H. J. *et al.* Biomass production in experimental grasslands of different species richness during three years of climate warming. *Biogeosciences* **5,** 585–594 (2008).
46. Fridley, J. D. Diversity effects on production in different light and fertility environments: an experiment with communities of annual plants. *J. Ecol.* **91,** 396–406 (2003).

LETTER

doi:10.1038/nature11145

International trade drives biodiversity threats in developing nations

M. Lenzen[1], D. Moran[1], K. Kanemoto[1,2], B. Foran[1,3], L. Lobefaro[1,4] & A. Geschke[1]

Human activities are causing Earth's sixth major extinction event[1]—an accelerating decline of the world's stocks of biological diversity at rates 100 to 1,000 times pre-human levels[2]. Historically, low-impact intrusion into species habitats arose from local demands for food, fuel and living space[3]. However, in today's increasingly globalized economy, international trade chains accelerate habitat degradation far removed from the place of consumption. Although adverse effects of economic prosperity and economic inequality have been confirmed[4,5], the importance of international trade as a driver of threats to species is poorly understood. Here we show that a significant number of species are threatened as a result of international trade along complex routes, and that, in particular, consumers in developed countries cause threats to species through their demand of commodities that are ultimately produced in developing countries. We linked 25,000 Animalia species threat records from the International Union for Conservation of Nature Red List to more than 15,000 commodities produced in 187 countries and evaluated more than 5 billion supply chains in terms of their biodiversity impacts. Excluding invasive species, we found that 30% of global species threats are due to international trade. In many developed countries, the consumption of imported coffee, tea, sugar, textiles, fish and other manufactured items causes a biodiversity footprint that is larger abroad than at home. Our results emphasize the importance of examining biodiversity loss as a global systemic phenomenon, instead of looking at the degrading or polluting producers in isolation. We anticipate that our findings will facilitate better regulation, sustainable supply-chain certification and consumer product labelling.

Many studies have linked export-intensive industries with biodiversity threats, for example, coffee growing in Mexico[6] and Latin America[7], soya[8] and beef[9] production in Brazil, forestry[10] and fishing[11] in Papua New Guinea, palm oil plantations in Indonesia and Malaysia[12], and ornamental fish catching in Vietnam[13], to name but a few. However, such studies are neither systematic nor comprehensive in their coverage of international trade. They also do not link exports to consuming countries, and miss threats more difficult to connect to specific exports, such as agricultural and industrial pollution.

Our approach provides a comprehensive view of the commercial causes of biodiversity threats. Using information from the International Union for Conservation of Nature (IUCN) Red List on threat causes, we associated threatened species with implicated commodities; for example, *Ateles geoffroyi* (spider monkey) is endangered and threatened by habitat loss linked to coffee and cocoa plantations in Mexico and Central America. Using a high-resolution global trade input–output table, we traced the implicated commodities from the country of their production, often through several intermediate trade and transformation steps, to the country of final consumption (Methods). This is the first time, to our knowledge, that the important role of international trade and foreign consumption as a driver of threats to species has been comprehensively quantified.

We calculated the net trade balances of 187 countries (Supplementary Information section 1) in terms of implicated commodities (Supplementary Information section 2). Countries that export more implicated commodities than they import are net biodiversity exporters, and importers vice versa. A striking division exists between the world's top ten net exporters and net importers of biodiversity (Fig. 1 and Supplementary Information section 3). Developed countries tend to be relatively minor net exporters, but major net importers of implicated commodities. This is probably due to environmental policies that effectively protect remaining domestic species and that force impacting industries to locate elsewhere. Among the net importers a total of 44% of their biodiversity footprint is linked to imports produced outside their boundaries. In stark contrast, developing countries find themselves degrading habitat and threatening biodiversity for the sake of producing exports. Among the net exporters a total of 35% of domestically recorded species threats are linked to production for export. In Madagascar, Papua New Guinea, Sri Lanka and Honduras, this proportion is approximately 50–60%.

Examining exporters and importers in unison shows that primarily the USA, the European Union and Japan are the main final destinations of biodiversity-implicated commodities. For the five selected exporting countries shown in Fig. 2, export activities are linked to between 50 and 60% of all domestically recorded biodiversity threats.

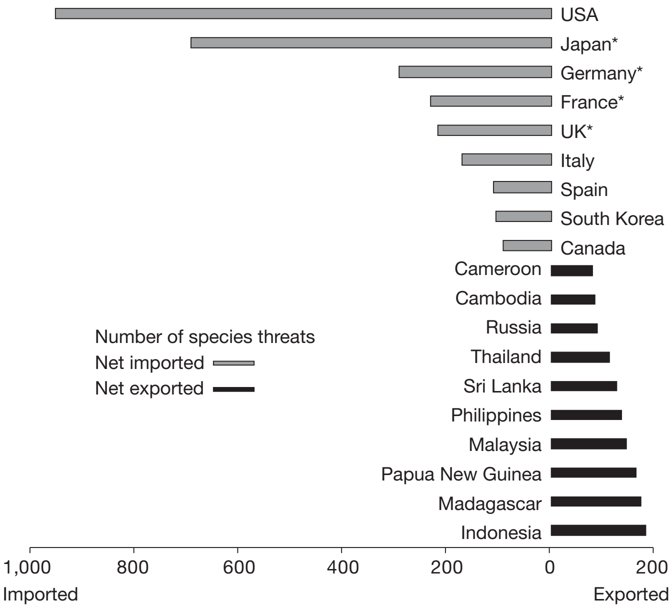


Figure 1 | Top net importers and exporters of biodiversity threats. In importer countries marked with an asterisk, the biodiversity footprint rests more abroad then domestically; that is, more species are threatened by implicated imports than are threatened by domestic production.

[1]ISA, School of Physics A28, The University of Sydney, New South Wales 2006, Australia. [2]Graduate School of Environmental Studies, Tohoku University, Sendai 980-8579, Japan. [3]Institute of Land Water and Society, Charles Sturt University, Albury, New South Wales 2640, Australia. [4]Department of Business and Law Studies, I Faculty of Economics, University of Bari Aldo Moro, 70124 Bari, Italy.

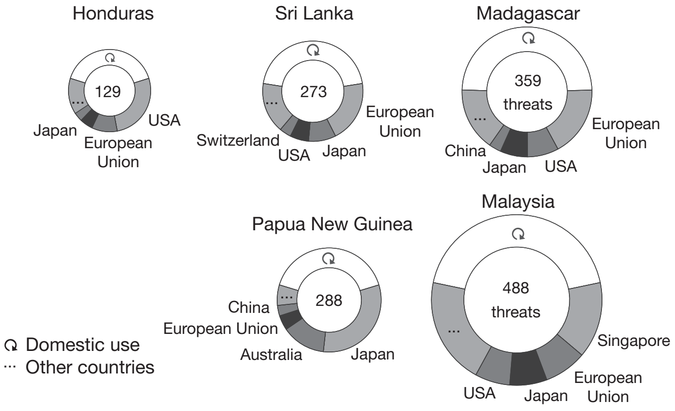


Figure 2 | Selected net exporters. Selected net exporters and final destinations of biodiversity-implicated commodities.

Further examination of the commodity content of these trade activities shows that threats to species are often facilitated by supply chains involving more than two countries or producers, and that major supply chains originate in developing countries rich in biodiversity and with export-oriented agricultural, fishing and forestry industries (Supplementary Information section 4). Coffee, a top-ranking commodity, is threatening species in Mexico, Colombia and Indonesia. Agriculture also affects habitat in Papua New Guinea (where coffee, cocoa, palm oil and coconut growing are linked to nine critically endangered species including the northern glider, *Petaurus abidi*, the black-spotted cuscus, *Spilocuscus rufoniger*, and the eastern long-beaked echidna, *Zaglossus bartoni*), Malaysia (the main export products are palm oil, rubber and cocoa; 135 species are affected by agriculture) and Indonesia (the main crops are rubber, coffee, cocoa and palm oil, affecting 294 species including *Panthera tigris*, the Sumatran serow, *Capricornis sumatraensis*, and Sir David's long-beaked echidna, *Zaglossus attenboroughi*). Fishing and forestry industries cause biodiversity loss directly through excessive and illegal resource use and indirectly through bycatch and habitat loss. Such impacts occur not only in developing countries such as the Philippines (affecting 420 species, 28 of which are critically endangered) and Thailand (affecting 352 species, 28 critically) but also in the United States (affecting 450 species, 63 critically). Biological resource use is not the only threat. In China, pollution is responsible for one-fifth (304 out of 1,526) of all threats. Consumers in the United States and Japan are the largest beneficiaries of these trade flows. Finally, most species on the Red List suffer several different threats. For example, the vulnerable round whipray, *Himantura pastinacoides*, is under threat in Indonesia owing to chemical pollution and loss of its native mangrove habitat to shrimp aquaculture, logging and coastal development.

The international trade in biodiversity-implicated commodities can be visualized using global trade-flow maps. Figure 3 illustrates the flows of implicated commodities for two countries: exports from Malaysia, and imports by Germany (this figure is available in higher resolution in the Supplementary Information, and an interactive version is available online at http://www.worldmrio.com/biodivmap/). German imports are linked to 395 species threats, and Malaysian exports to 276 species threats. Further details supporting Fig. 3 are given in Supplementary Information section 7. In Papua New Guinea, 171 listed species are threatened by exports to fewer, but larger, trade partners. Half of Papua New Guinea's implicated exports are destined for Japan. These are mostly timber and agricultural products that undergo intermediate processing in Malaysia and Indonesia (wood machining), and Hong Kong, Taiwan, Australia and Thailand (food processing). Countries producing implicated goods bound for Germany are diverse, such as Madagascar (twine, rattan, sisal, cocoa, vanilla, cloves and processed food prepared in France, Austria and the Netherlands; 18 species), Democratic Republic of the Congo and Ghana (mining inputs to Finnish metal products used in German passenger-car production; 3 and 5 species, respectively), Sri Lanka (tea, latex gloves, rubber products for automobiles and cotton clothing; 14 species), Colombia (coffee, bananas, tobacco and cocoa made into chocolate; 3 species) and Cameroon (coffee, rubber, wool, lumber and cargo pallets; 6 species).

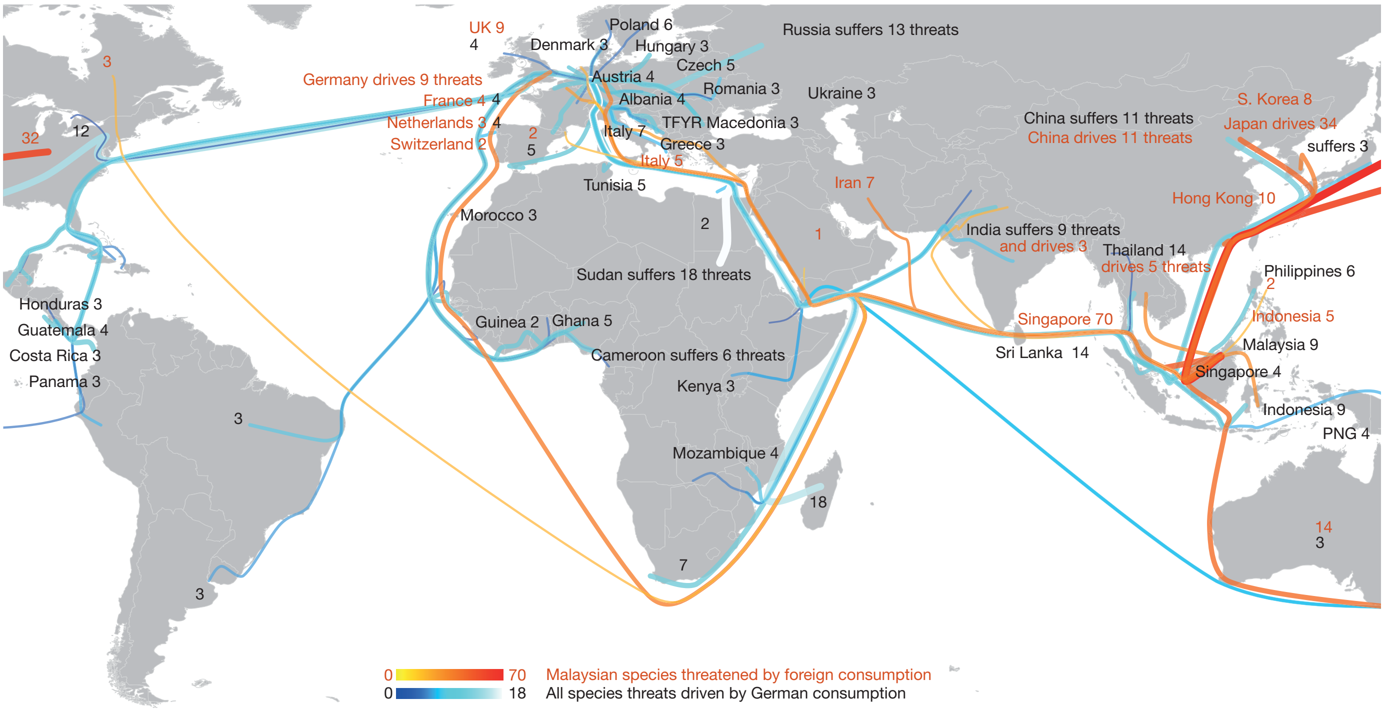


Figure 3 | Flow map of threats to species. Flow map of threats to species caused by exports from Malaysia (reds) and imports into Germany (blues). Note that the lines directly link the producing countries, where threats are recorded, and final consumer countries. Supply-chain links in intermediate countries are accounted for but not explicitly visualized. An interactive version is available at http://www.worldmrio.com/biodivmap/.

Our findings clearly show that local threats to species are driven by economic activity and consumer demand across the world. Consequently, policy aimed at reducing local threats to species should be designed from a global perspective, taking into account not just the local producers who directly degrade and destroy habitat but also the consumers who benefit from the degradation and destruction.

Allocating responsibility between producers and consumers is not straightforward, even as an academic exercise. Producers exert the impacts and control production methods, but consumer choice and demand drives production, so that responsibility may lie with both camps, and may hence have to be shared between them[14]. Notwithstanding its theoretical challenges, the consumer responsibility principle is now receiving ample attention in the climate change debate. Its political relevance is demonstrated by China's official stance that final consumer countries should be held accountable for the greenhouse gases emitted during the production of China's export goods[14]. To inform this debate, countries' carbon footprints are now being calculated using global multi-region input–output models[15]. The biodiversity footprints introduced here use identical concepts and methods. Therefore, policies to mitigate climate change and biodiversity loss may share analytical approaches and implementation protocols on the basis of supply-chain quantification.

Starting with the producers, regulating polluting and degrading industries in developing countries may be difficult if these industries have limited means and alternatives, and are vital to income and employment. These limits may not apply to multi-national producers that operate in the developing world but are controlled from a developed country. The emigration of industries as a result of tightening domestic environmental or work standard laws is well known. Such migration can be countered by extending domestic jurisdiction to producers abroad. Similar processes may be behind the stark division between net importing and net exporting countries shown in Fig. 1. Harmonizing regulation and standards among trade partners may stem the migration of habitat-intensive producers. Producer-side sustainability initiatives such as the developing[16] Roundtable for Sustainable Palm Oil can further reduce the impacts of production.

Moving from producers to traders, the 1977 Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) today protects more than 30,000 species[17]. CITES is exclusively concerned with the international trade of endangered species, be it as live specimens, parts or derived products. Indeed, trade in marine,[18] sylvan[19] and iconic endangered species[20] can be, to a degree, constrained by certification, quota and regulatory regimes. We argue that there is no practical difference in terms of imperilment between trading specimens and trading commodities whose production leads to their imperilment. The motivation for banning the first kind of trade equally applies to the second kind, and, consequently, trade in biodiversity-implicated commodities should be governed by the same control and licensing procedures.

Ending with consumers, environmental labels such as advertising dolphin-safe canned tuna, organic produce and fair trade coffee have been a well-known sight for decades. Although these examples refer to relative short, intuitively traceable supply chains, there is in principle no obstacle to extending such labelling and certification to more complex international trade routes. This is demonstrated by the United Kingdom's Carbon Reduction Label, which despite methodological shortcomings[21] requires the quantification of a product's full carbon footprint. Given the complete equivalence of carbon and biodiversity footprinting methodologies (Supplementary Information sections 8 and 9), our integration of species Red Lists with global trade databases could provide a starting point for more comprehensive biodiversity labelling schemes. However, whether sustainability-minded consumers and shareholders can be a force in mitigating the impacts they drive will depend on whether sustainability certification schemes will be able to overcome their current limited efficacy[22].

To combat biodiversity loss, policies aimed at producers, traders and consumers must be implemented in parallel. This is reflected in the reinterpretation of the wedge approach[23] for biodiversity stabilization, which considers wide-ranging measures on human population, consumption, endowment funds to underpin permanence of habitat refuges, economic accounting of habitat degradation, reclamation of degraded lands, empowerment of local peoples and transformation of human attitudes to nature[24]. We suggest a new wedge: suppressing trade in at-risk commodities. Granted, such a policy reform would be difficult to implement given the importance of international trade. However, Article XX of the General Agreement on Tariffs and Trade (GATT) allows "measures relating to the conservation of exhaustible natural resources", thus providing a framework to support measures regulating biodiversity-implicated goods[25]. Reducing the volume of trade in implicated commodities and implementing protective policies at the production, trade and consumption points in the supply chain could have a significantly higher impact in preventing biodiversity loss than the CITES controls. Raising consumers' awareness of the biodiversity footprint of the products they buy also helps with most of these measures. Mexico's spider monkey, *Ateles geoffroyi*, is listed in the CITES as a protected species, but its survival would be more certain if consumers could see that the coffee encroaching on its home were listed as a biodiversity-implicated commodity as well.

METHODS SUMMARY

We integrated the IUCN Red List of Threatened Species[26] plus a compatible list of threatened bird species from Bird Life International[27] with a new high-resolution global multi-region input–output database[28]. The combined threat lists (excluding natural disaster, intrinsic factors and invasive species) provide country-wise information on 166 anthropogenic threat causes. We considered only endangered, critically endangered and vulnerable species. This data set covered 6,964 Animalia species and 171,825 country, species and cause records.

We linked these threats to a multi-region input–output table containing the domestic and international monetary transactions between 15,909 industry sectors across 187 countries. Using a binary concordance matrix, we attributed each threat cause to one or more industry sectors that exert the respective threat. We could not distinguish legal from illegal activities (for example, fishing, forestry and hunting), as data were unavailable. For species threatened by climate change, responsibility was allocated to all sectors worldwide. We then normalized the concordance matrix by weighting threat assignments by the gross industrial output of sectors for all causes except for climate change, where the weights are based on the sectors' greenhouse gas emissions. This normalization ensured that threat causes were not double-counted. We weighted all threat causes equally as there are no data with which to weight threat severity. Finally we determined biodiversity footprints using Leontief's standard input–output calculus[29]. These biodiversity footprints quantify threats caused directly and indirectly as a consequence of the expenditure of a final consumer. For example, the United States' biodiversity footprint contains the number of species threatened in Mexico caused indirectly by consumer spending on Mexican coffee beans in the USA. Such international indirect threats are facilitated by complex, multi-stage, global supply chains, which can be traced, extracted and ranked using structural path analysis. Further details are available in Supplementary Information sections 8, 9, 10 and 11.

Received 18 December 2011; accepted 16 April 2012.

1. Chapin, F. S. *et al.* Consequences of changing biodiversity. *Nature* **405,** 234–242 (2000).
2. Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. The future of biodiversity. *Science* **269,** 347–350 (1995).
3. Donald, P. F. Biodiversity impacts of some agricultural commodity production systems. *Conserv. Biol.* **18,** 17–38 (2004).
4. Naidoo, R. & Adamowicz, W. L. Effects of economic prosperity on numbers of threatened species. *Conserv. Biol.* **15,** 1021–1029 (2001).
5. Mikkelson, G. M., Gonzalez, A. & Peterson, G. D. Economic inequality predicts biodiversity loss. *PLoS ONE* **2,** e444 (2007).
6. Perfecto, I., Mas, A., Dietsch, T. & Vandermeer, J. Conservation of biodiversity in coffee agroecosystems: a tri-taxa comparison in southern Mexico. *Biodivers. Conserv.* **12,** 1239–1252 (2003).
7. Philpott, S. M. *et al.* Biodiversity loss in Latin American coffee landscapes: review of the evidence on ants, birds, and trees. *Conserv. Biol.* **22,** 1093–1105 (2008).
8. Fearnside, P. M. Soybean cultivation as a threat to the environment in Brazil. *Environ. Conserv.* **28,** 23–38 (2001).

9. Nepstad, D. C., Stickler, C. M. & Almeida, O. T. Globalization of the Amazon soy and beef industries: opportunities for conservation. *Conserv. Biol.* **20,** 1595–1603 (2006).
10. Shearman, P. L., Ash, J., Mackey, B., Bryan, J. E. & Lokes, B. Forest conversion and degradation in Papua New Guinea 1972–2002. *Biotropica* **41,** 379–390 (2009).
11. Michael E, H. An assessment of the status of the coral reefs of Papua New Guinea. *Mar. Poll. Bull.* **29,** 69–73 (1994).
12. Koh, L. P. & Wilcove, D. S. Cashing in palm oil for conservation. *Nature* **448,** 993–994 (2007).
13. Giles, B. G., Ky, T. S., Hoang, H. & Vincent, A. C. J. in *Topics in Biodiversity and Conservation* Vol. 3 (eds Hawksworth, D. L. & Bull, A. T.) 157–173 (Springer Netherlands, 2006).
14. Lenzen, M., Murray, J., Sack, F. & Wiedmann, T. Shared producer and consumer responsibility – theory and practice. *Ecol. Econ.* **61,** 27–42 (2007).
15. Peters, G. P., Minx, J. C., Weber, C. L. & Edenhofer, O. Growth in emission transfers via international trade from 1990 to 2008. *Proc. Natl Acad. Sci. USA,* (2011).
16. Edwards, D. P., Fisher, B. & Wilcove, D. S. High conservation value or high confusion value? Sustainable agriculture and biodiversity conservation in the tropics. *Conserv. Lett.* **5,** 20–27 (2012).
17. Convention on International Trade in Endangered Species of Wild Fauna and Flora. http://www.cites.org (1979).
18. Villasante, S., Rodríguez, D., Antelo, M., Quaas, M. & Österblom, H. The Global Seafood Market Performance Index: a theoretical proposal and potential empirical applications. *Mar. Policy* **36,** 142–152 (2012).
19. Rotherham, T. Forest management certification around the world — progress and problems. *For. Chron.* **87,** 603–611 (2011).
20. Parsons, E. C. M. & Cornick, L. A. Sweeping scientific data under a polar bear skin rug: The IUCN and the proposed listing of polar bears under CITES Appendix I. *Mar. Policy* **35,** 729–731 (2011).
21. Huang, A. Y., Lenzen, M., Weber, C., Murray, J. & Matthews, H. S. The role of input-output analysis for the screening of corporate carbon footprints. *Econ. Syst. Res.* **21,** 217–242 (2009).
22. Blackman, A. & Rivera, J. Producer-level benefits of sustainability certification. *Conserv. Biol.* **26,** 1176–1185 (2011).
23. Pacala, S. & Socolow, R. Stabilization wedges: Solving the climate problem for the next 50 years with current technologies. *Science* **305,** 968–972 (2004).
24. Ehrlich, P. R. & Pringle, R. M. Where does biodiversity go from here? A grim business-as-usual forecast and a hopeful portfolio of partial solutions. *Proc. Natl Acad. Sci. USA* **105,** 11579–11586 (2008).
25. World Trade Organization. WTO Rules and Environmental Policies: GATT Exceptions. http://www.wto.org/english/tratop_e/envir_e/envt_rules_exceptions_e.htm (2012).
26. International Union for Conservation of Nature. The IUCN Red List of Threatened Species. Version 2011.2. http://www.iucnredlist.org (2011).
27. BirdLife International. Threatened Birds of the World. http://www.birdlife.org (2011).
28. Lenzen, M., Kanemoto, K., Moran, D. & Geschke, A. The Eora Global Multi-Region Input-Output Tables. ISA, Univ. Sydney, Australia http://www.worldmrio.com (2011).
29. Leontief, W. & Ford, D. Environmental repercussions and the economic structure: an input-output approach. *Rev. Econ. Stat.* **52,** 262–271 (1970).

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

Acknowledgements We thank C. Pollock and M. Hoffmann from the IUCN and A. Symes from BirdLife International for advice on using the Red Lists. The work described in this paper was financially supported by the Australian Research Council through its Discovery Project DP0985522 and its Linkage Project LP0669290.

Author Contributions M.L. and D.M. conducted the analysis and prepared the figures. M.L., D.M. and B.F. wrote the paper. K.K., L.L. and A.G. prepared the data.

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.L. (m.lenzen@physics.usyd.edu.au).

doi:10.1038/nature10958

Preferential electrical coupling regulates neocortical lineage-dependent microcircuit assembly

Yong-Chun Yu[1]*, Shuijin He[2]*, She Chen[2], Yinghui Fu[1], Keith N. Brown[2,3], Xing-Hua Yao[1], Jian Ma[1], Kate P. Gao[2,3], Gina E. Sosinsky[4], Kun Huang[5] & Song-Hai Shi[2,3]

Radial glial cells are the primary neural progenitor cells in the developing neocortex[1]. Consecutive asymmetric divisions of individual radial glial progenitor cells produce a number of sister excitatory neurons that migrate along the elongated radial glial fibre, resulting in the formation of ontogenetic columns[2–4]. Moreover, sister excitatory neurons in ontogenetic columns preferentially develop specific chemical synapses with each other rather than with nearby non-siblings[5]. Although these findings provide crucial insight into the emergence of functional columns in the neocortex, little is known about the basis of this lineage-dependent assembly of excitatory neuron microcircuits at single-cell resolution. Here we show that transient electrical coupling between radially aligned sister excitatory neurons regulates the subsequent formation of specific chemical synapses in the neocortex. Multiple-electrode whole-cell recordings showed that sister excitatory neurons preferentially form strong electrical coupling with each other rather than with adjacent non-sister excitatory neurons during early postnatal stages. This preferential coupling allows selective electrical communication between sister excitatory neurons, promoting their action potential generation and synchronous firing. Interestingly, although this electrical communication largely disappears before the appearance of chemical synapses, blockade of the electrical communication impairs the subsequent formation of specific chemical synapses between sister excitatory neurons in ontogenetic columns. These results suggest a strong link between lineage-dependent transient electrical coupling and the assembly of precise excitatory neuron microcircuits in the neocortex.

Sister excitatory neurons in individual ontogenetic columns preferentially develop specific chemical synapses with each other rather than with adjacent non-sister excitatory neurons[5]. Given the almost complete overlap of the dendritic fields of neighbouring excitatory neurons, it is unclear how this lineage-dependent assembly of precise columnar microcircuits is controlled at the individual cell level. Some studies have suggested that gap-junction-mediated neuronal communication is involved in the formation of local connectivity in the developing neocortex[6–8], even though direct evidence of electrically coupled neocortical neurons at early developmental stages is lacking. In this study, we set out to investigate whether gap-junction-mediated electrical coupling exists between sister excitatory neurons in ontogenetic columns and, if so, whether this coupling regulates the preferential formation of chemical synapses between sister excitatory neurons.

Gap junctions are composed of two membrane-docked hexameric hemi-channels that consist of connexin proteins from two adjacent cells. There are ~20 genes encoding connexins in rodents, and the corresponding protein symbols are denoted as CX plus the calculated molecular mass of the protein[9]. Of these proteins, CX26 and CX43 have been shown to be abundantly expressed in the developing neocortex at embryonic and neonatal stages[10,11]. Consistent with this, we found that developing neurons in the neonatal neocortex expressed CX26 (Supplementary Fig. 1a). Moreover, CX26-positive puncta were present at the dendrodendritic and dendrosomatic contacts of radially aligned sister excitatory neurons that were labelled by *in utero* intraventricular injection of low-titre enhanced green fluorescence protein (eGFP)-expressing retrovirus at embryonic day 12 to 13 (E12–13) (Fig. 1a–e and Supplementary Movie 1), indicating the existence of gap junctions between sister excitatory neurons in ontogenetic columns.

Gap junctions mediate intercellular adhesion and the exchange of small molecules (typically less than 1 kDa), including low-molecular-mass dyes and ions that can be detected experimentally[12,13]. Previous dye injection experiments have suggested the presence of gap junctions between progenitor cells in the embryonic neocortex[2,14] and between neurons in the neonatal neocortex[6,15]. Although these studies have provided important insight, the accuracy of dye coupling in revealing the presence of gap junctions has been debated[12,13]. To circumvent this issue and to quantitatively examine gap junction channel activity, we performed whole-cell patch-clamp recording experiments to study gap junctions between sister excitatory neurons in ontogenetic columns.

We prepared acute neocortical slices from postnatal mice (postnatal day 1 to 28, P1–28) that had received *in utero* intraventricular injection of eGFP-expressing retrovirus at E12–13. Guided by infrared differential interference contrast (DIC) and epifluorescence illumination, we simultaneously recorded from two radially aligned eGFP-expressing sister excitatory neurons (Fig. 1f and Supplementary Fig. 1b, top (green)). We identified excitatory neurons on the basis of their morphological characteristics, including a pyramid-shaped cell body and a long apical dendrite. After the recordings were established, we tested the electrical coupling between the neurons under current-clamp conditions (Fig. 1f and Supplementary Fig. 1b, bottom) or voltage-clamp conditions (Supplementary Fig. 2a). Under current-clamp conditions, hyperpolarization (V_1) of one of the neurons (neuron 1, the driver) by current injection (I_1) produced a simultaneous hyperpolarization (V_2) of the non-injected neuron (neuron 2, the receiver) (Fig. 1f and Supplementary Fig. 1b, bottom). Similarly, depolarization of the driver neuron (neuron 1) produced a simultaneous depolarization of the receiver neuron (neuron 2). As expected for electrotonic propagation, voltage deflections recorded in the non-injected receiver neuron (Fig. 1f, red, and Supplementary Fig. 1b, bottom) had a smaller amplitude and a slower time course than those in the injected driver neuron (Fig. 1f, black, and Supplementary Fig. 1b, bottom). In all cases, electrical transmission between the two neurons was found to be reciprocal (Fig. 1f and Supplementary Fig. 1b, bottom). Similar bidirectional electrotonic propagation was recorded under voltage-clamp conditions (Supplementary Fig. 2a).

To confirm that the electrical coupling between sister excitatory neurons is mediated by gap junctions, we exposed electrically coupled sister excitatory neuron pairs to the gap junction blocker meclofenamic

[1] Institute of Neurobiology, Institutes of Brain Science and State Key Laboratory of Medical Neurobiology, Fudan University, 138 Yixueyuan Road, Shanghai 200032, China. [2] Developmental Biology Program, Memorial Sloan-Kettering Cancer Centre, 1275 York Avenue, New York, New York 10065, USA. [3] Neuroscience Graduate Program, Weill Cornell Medical College, 1230 York Avenue, New York, New York 10065, USA. [4] National Centre for Microscopy and Imaging Research and Department of Neurosciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0608, USA. [5] Department of Biomedical Informatics, Comprehensive Cancer Center Biomedical Informatics Shared Resource, The Ohio State University, 333 West 10th Avenue, Columbus, Ohio 43210, USA.
*These authors contributed equally to this work.

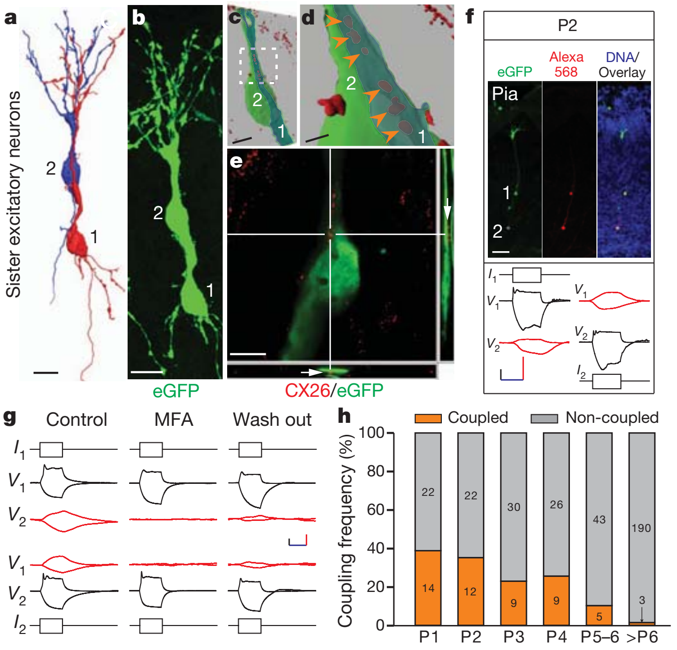


Figure 1 | Gap-junction-mediated electrical coupling between sister excitatory neurons in neonatal neocortical ontogenetic columns. **a**–**e**, Images and three-dimensional reconstruction of two radially aligned eGFP-expressing sister excitatory neurons (1 and 2) (green) immunostained with antibody specific for CX26 (red). (A morphological reconstruction is shown in **a**, in which neuron 1 is shown in red and neuron 2 in blue.) Note the presence of CX26 puncta at the dendrosomatic and dendrodendritic contacts of sister excitatory neurons in the three-dimensional reconstruction (**c** and indicated by arrowheads in **d**, which shows a magnification of the boxed region in **c**) and in the *Z*-axis cross section (**e**, arrows). Scale bars, 10 μm (**a**, **b**); 5 μm (**c**); 1 μm (**d**); and 5 μm (**e**). **f**, Dual whole-cell recordings of sister excitatory neurons in ontogenetic columns at the neonatal stage P2. Top, images of sister excitatory neurons expressing eGFP (green) in an ontogenetic column; the neurons have been filled with Alexa Fluor 568 hydrazide (red) through the recording pipette and stained with the DNA-binding dye 4′,6-diamidino-2-phenylindole (DAPI; blue). Scale bar, 50 μm. Bottom, simultaneous depolarization or hyperpolarization of two sister excitatory neurons (V_1 and V_2) when a positive or negative current (+100 pA or −100 pA) was injected into only one of the neurons (I_1 or I_2) under current-clamp conditions, indicating electrical coupling. Subsequent figures use a similar panel layout. Scale bars, 50 mV (black), 10 mV (red) and 200 ms (blue). Pia, pial surface. **g**, Blockade of electrical coupling by the gap junction blocker MFA (100 μM). Scale bars, 25 mV (black), 5 mV (red) and 200 ms (blue). **h**, Summary of the frequency of electrical coupling between sister excitatory neurons in ontogenetic columns at different postnatal stages.

acid (MFA, 100 μM), the glutamate receptor antagonists 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX, 10 μM) and 3-(2-carboxypiperazin-4-yl)propyl-1-phosphonic acid (CPP, 20 μM), or the GABA_A (γ-aminobutyric acid A) receptor antagonist bicuculline methiodide (BMI, 10 μM). Neither BMI nor CNQX plus CPP had any effect ($n = 3$; Supplementary Fig. 2b); however, MFA completely eliminated electrical transmission in all cases ($n = 9$; Fig. 1g). Moreover, electrical transmission recovered partially after MFA was washed out (Fig. 1g).

The frequency of observed gap-junction-mediated electrical coupling between sister excitatory neurons in ontogenetic columns at P1 and P2 was 38.9% (occurring in 14 of the 36 pairs tested) and 35.3% (12 of 34 pairs tested), respectively (Fig. 1h). As development proceeded, this coupling frequency progressively decreased to ~20–25% at P3–P4 (9 of 39 pairs tested at P3 and 9 of 35 pairs tested at P4) and to ~10% at P5–P6 (5 of 48 pairs tested) (Fig. 1h). After P6, electrical coupling between sister excitatory neurons was rare; only 3 of 193 pairs were coupled (Fig. 1h); this finding is consistent with previous observations of very sparse electrical coupling among more mature excitatory neurons in the neocortex[16–18]. No chemical synapses were detected between sister excitatory neurons at early postnatal stages (P1–P6), as previously reported[5].

To test whether sister excitatory neurons preferentially form gap junctions with each other, we performed quadruple whole-cell recordings on two eGFP-expressing sister excitatory neurons in individual ontogenetic columns and on two non-eGFP-expressing excitatory neurons adjacent to the eGFP-expressing neurons on the same side in the developing neocortex (Fig. 2a). We carried out this experiment at P1–P6, when electrical coupling is prominent (Fig. 1h). Once all four recordings were established, hyperpolarizing and depolarizing currents were injected sequentially into one of the four neurons, and the voltage changes were monitored in all four neurons to probe gap-junction-mediated electrical coupling. When eGFP-expressing neuron 1 was hyperpolarized or depolarized, only its sister neuron

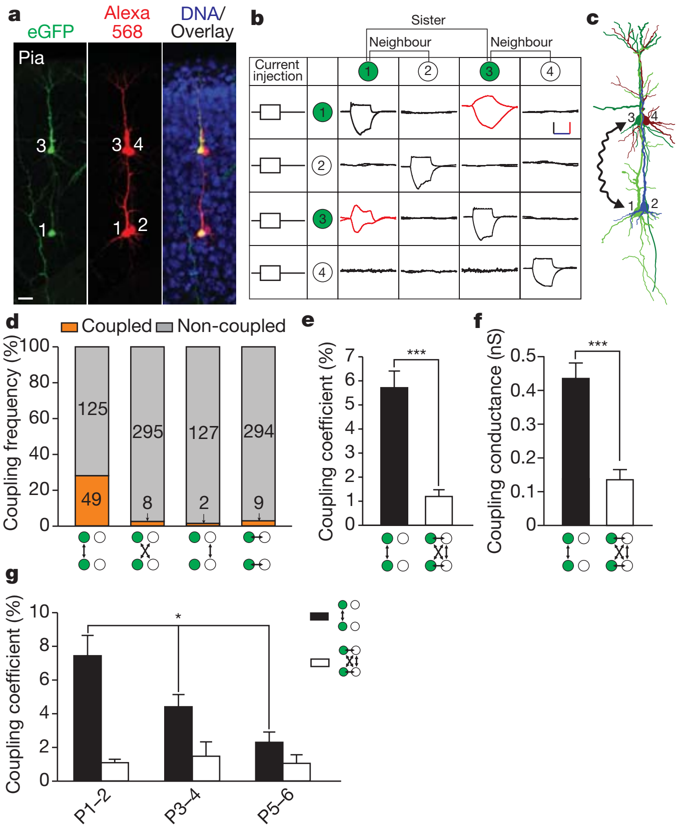


Figure 2 | Preferential formation of strong electrical coupling between sister excitatory neurons in neonatal neocortical ontogenetic columns. **a**, Image of a quadruple whole-cell recording of two eGFP-expressing sister excitatory neurons (neurons 1 and 3, green) in an ontogenetic column and two non-eGFP-expressing excitatory neurons (neurons 2 and 4) adjacent to the sisters on the same side, filled with Alexa Fluor 568 hydrazide (red) and stained with DAPI (blue). Scale bar, 20 μm. **b**, Sample traces of voltage changes in the four neurons in response to sequential current injection into one of the four neurons. Green circles indicate eGFP-expressing sister excitatory neurons, and white circles indicate non-eGFP-expressing neighbouring excitatory neurons. The average traces are shown in each table cell. Scale bars, 50 mV (black), 5 mV (red) and 200 ms (blue). **c**, A morphological reconstruction of the four neurons in the quadruple recording. The wavy arrow indicates reciprocal electrical coupling between sister excitatory neurons 1 and 3. **d**, Summary of the frequency of electrical coupling observed between sister excitatory neurons in ontogenetic columns and their adjacent non-sister excitatory neurons at P1–P6. **e**, **f**, Summary of the coupling coefficient (**e**) (sisters, $n = 74$; non-sisters, $n = 24$; $^{***}P < 1 \times 10^{-8}$; data are mean ± s.e.m.) and conductance (**f**) (sisters, $n = 85$; non-sisters, $n = 24$; $^{***}P < 5 \times 10^{-7}$; data are mean ± s.e.m.) at P1–P6. **g**, Progressive decrease in the coupling coefficient between sister excitatory neurons in ontogenetic columns as development proceeds (sisters, P1–P2, $n = 36$; P3–P4, $n = 32$; P5–P6, $n = 6$; $^{*}P < 0.01$; non-sisters, P1–P2, $n = 11$; P3–P4, $n = 7$; P5–P6, $n = 6$; $P = 0.7$; data are mean ± s.e.m.).

(neuron 3) showed simultaneous hyperpolarization or depolarization (Fig. 2b), despite the almost complete overlap of the neurite arbours of the adjacent neurons, neurons 3 and 4 (Fig. 2c). Similarly, when eGFP-expressing neuron 3 was hyperpolarized or depolarized, only its sister neuron (neuron 1) showed simultaneous hyperpolarization or depolarization (Fig. 2b). The hyperpolarization or depolarization of neuron 2 or 4 failed to trigger simultaneous voltage deflection in any of the other three neurons (Fig. 2b).

We analysed a total of 174 pairs of radially aligned eGFP-expressing sister excitatory neurons and their neighbouring non-sibling excitatory neurons at P1–P6 (Fig. 2d). Of the sister excitatory neuron pairs in an ontogenetic column, 28.2% (49 of 174 pairs) were electrically coupled. By contrast, only 2.6% (8 of 303 pairs) of radially situated non-sister excitatory neuron pairs (one eGFP-expressing and one non-eGFP-expressing) were electrically coupled (Fig. 2d). In addition, only 1.6% (2 of 129 pairs) of similarly radially situated non-eGFP expressing excitatory neuron pairs were coupled, and only 3.0% (9 of 303) of nearby eGFP-expressing and non-eGFP-expressing excitatory neuron pairs were coupled (Fig. 2d). These findings are consistent with a recent study showing almost no electrical coupling between randomly selected excitatory neurons in the developing neocortex[19]. Our results clearly demonstrate that sister excitatory neurons in ontogenetic columns have a strong preference for gap-junction-mediated electrical coupling with each other rather than with adjacent non-sister excitatory neurons and that the frequency of electrical coupling between non-sister excitatory neurons in the developing neocortex is low.

We also compared the coupling coefficient between coupled sister excitatory neuron pairs and rarely coupled non-sister excitatory neuron pairs. The coupling coefficient, estimated as the ratio of the amplitude of the low-frequency voltage change in the receiver neuron to that in the driver neuron, reflects the strength of the electrical coupling. We found that the coupling coefficient between coupled sister excitatory neurons was substantially higher than that between coupled non-sister excitatory neurons (sisters, 5.7 ± 0.7%, $n = 74$; non-sisters, 1.2 ± 0.3%, $n = 24$; $P < 1 \times 10^{-8}$; Fig. 2e). We also estimated the coupling conductance under voltage-clamp conditions and found that the coupling conductance between sister excitatory neurons was significantly larger than that between non-sister excitatory neurons (sisters, 0.43 ± 0.05 nS, range 0.06–2.16 nS, $n = 85$; non-sisters, 0.13 ± 0.03 nS, range 0.06–0.81 nS, $n = 24$; $P < 5 \times 10^{-7}$; Fig. 2f). These results suggest that the electrical coupling between sister excitatory neurons is much stronger than that between non-sisters. Furthermore, we observed a progressive decrease in the coupling coefficient of coupled sister excitatory neuron pairs as development proceeded (P1–P2, 7.4 ± 1.2%, $n = 36$; P3–P4, 4.4 ± 0.7%, $n = 32$; P5–P6, 2.3 ± 0.6%, $n = 6$; $P < 0.01$), whereas the coupling coefficient of rarely coupled non-sister pairs did not change significantly during this period (P1–P2, 1.1 ± 0.2%, $n = 11$; P3–P4, 1.5 ± 0.8%, $n = 7$; P5–P6, 1.1 ± 0.5%, $n = 6$; $P = 0.7$; Fig. 2g).

Having found that sister excitatory neurons in ontogenetic columns preferentially form strong gap-junction-mediated electrical coupling with each other, we then examined the properties of this electrical transmission and tested whether this selective electrical coupling modulates the neuronal activity of sister excitatory neurons at neonatal stages. We injected a series of sinusoidal current waveforms of the same amplitude but different frequencies into one excitatory neuron and measured the response in its coupled sister excitatory neuron (Supplementary Fig. 3). We found that responses to low-frequency sine waves showed higher coupling coefficients and smaller phase lags than those triggered by high-frequency sine waves (Supplementary Fig. 3). These results suggest that the efficacy of signal transmission through electrical coupling between sister excitatory neurons is frequency dependent, similarly to previous observations of electrical coupling between inhibitory interneurons in the more mature neocortex[16,17].

To determine whether this selective electrical coupling modulates neuronal activity, we injected subthreshold depolarizing current pulses at the same time or at different times into two eGFP-labelled, electrically coupled sister excitatory neurons in ontogenetic columns (Fig. 3a). In most cases, the asynchronous pulses did not generate an action potential in either neuron (Fig. 3a, arrowheads, and Fig. 3b, open bars; neuron 1, 0.1 ± 0.1 spikes per pulse; neuron 2, 0.3 ± 0.2 spikes per pulse). However, when the same current pulses were synchronously injected into both neurons, the two sister neurons reached action potential threshold and generated spikes (Fig. 3a, arrows, and Fig. 3b, filled bars; neuron 1, 1.5 ± 0.3 spikes per pulse; neuron 2, 1.8 ± 0.4 spikes per pulse). Similar observations were made in seven pairs of electrically coupled sister excitatory neurons but not in non-coupled sister neuron pairs (Fig. 3c; $P < 0.001$). These results show that selective electrical coupling can strongly facilitate the generation of action potentials in coupled sister excitatory neurons in ontogenetic columns.

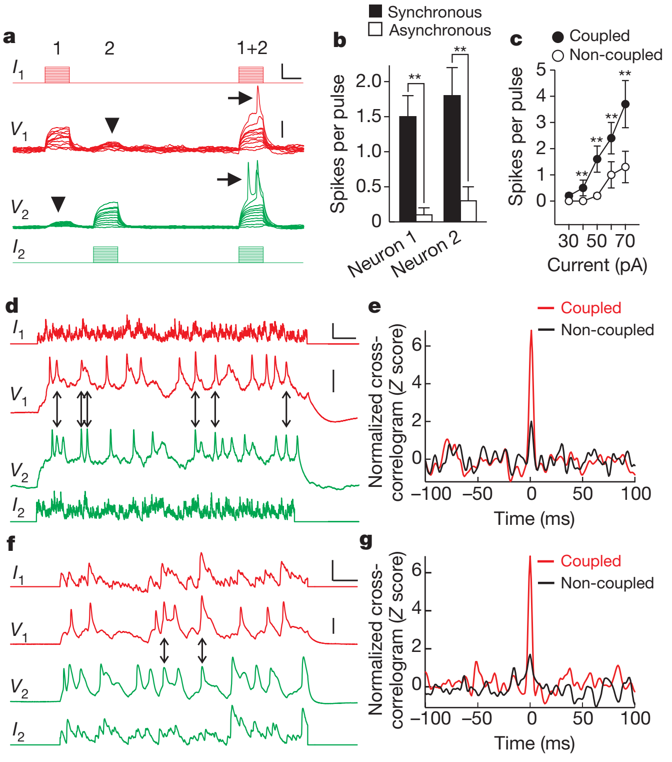


Figure 3 | Electrical coupling promotes action potential generation and synchronous firing of sister excitatory neurons in neonatal neocortical ontogenetic columns. **a–c**, Electrical coupling facilitates action potential generation in sister excitatory neurons. **a**, Sample traces of synchronous (1 + 2) or asynchronous (1 or 2) injection of subthreshold current pulses into electrically coupled sister excitatory neurons in an ontogenetic column. Arrowheads indicate voltage deflection due to the electrical transmission. Note that synchronous injection, but not asynchronous injection, results in action potential generation (arrows). Scale bars, 50 pA, 20 mV and 100 ms. **b**, Summary of the firing rate of the two sister excitatory neurons in **a** responding to a 50-pA current injection (**$P < 0.001$; data are mean ± s.e.m.). **c**, Summary of the firing rate in electrically coupled or non-coupled sister excitatory neurons responding to simultaneous current injections ($n = 7$; **$P < 0.001$; data are mean ± s.e.m.). **d–g**, Electrical coupling promoting synchronous firing of sister excitatory neurons in ontogenetic columns in response to uncorrelated simulated neuronal activity (**d**, **e**) or uncorrelated native neuronal activity (**f**, **g**). **d**, **f**, Sample traces of voltage changes in electrically coupled sister excitatory neurons. Arrows indicate the spikes that occur in both neurons within a 1-ms window. Scale bars, 200 pA, 30 mV and 100 ms (**d**); and 100 pA, 25 mV and 100 ms (**f**). **e**, **g**, Normalized cross-correlogram analysis. The bin size is 1 ms. Note that the firing frequency is significantly increased near 0 ms for coupled sister excitatory neuron pairs (red) but not for non-coupled sister excitatory neuron pairs (black), indicating synchronous firing.

We next determined whether this selective electrical coupling could facilitate synchronous spiking of radially aligned sister excitatory neurons in response to natural stimuli. First, we simulated natural activity by injecting two electrically coupled or non-coupled sister excitatory neurons with uncorrelated random current signals obtained by convolving the Poisson trains of recorded spontaneous excitatory postsynaptic current waveforms[17,20] (I_1 and I_2; Fig. 3d). Second, we directly recorded the spontaneous subthreshold activity of neurons in the neonatal neocortex and then injected two electrically coupled or non-coupled sister excitatory neurons with signals corresponding to the recorded uncorrelated native neuronal activity (I_1 and I_2; Fig. 3f). We then analysed the cross-correlogram of firing induced in these sister excitatory neurons and found that there was a significant increase in firing centred at a time of 0 ms in coupled pairs compared with non-coupled ones (Fig. 3e, coupled, $n = 4$; non-coupled, $n = 5$; $P < 0.005$; and Fig. 3g, coupled, $n = 10$; non-coupled, $n = 9$; $P < 0.05$). These results suggest that gap-junction-mediated electrical transmission between sister excitatory neurons in ontogenetic columns can effectively promote precise (within 1 ms) synchronous firing in response to uncorrelated neuronal activity.

It has long been postulated that correlated neuronal activity facilitates chemical synapse formation and neuronal circuit assembly[21,22]. Given that sister excitatory neurons in ontogenetic columns preferentially form strong gap-junction-mediated electrical coupling, which promotes action potential generation and synchronous firing, this process could provide a mechanism by which chemical synapses form preferentially between sister excitatory neurons in ontogenetic columns. We therefore tested whether the selective electrical coupling between sister excitatory neurons is required for preferential formation of chemical synapses within individual ontogenetic columns.

Our immunohistochemistry experiments suggest that CX26 is a major connexin isoform that mediates electrical coupling between sister excitatory neurons in the developing neocortex (Fig. 1a–e and Supplementary Fig. 1a). Previous studies have shown that mutation of a conserved threonine residue (Thr 135) in the third transmembrane helix of CX26 to an alanine residue (CX26(Thr135Ala)) creates dominant-negative closed gap junction channels without affecting the synthesis, assembly or trafficking of the channels[23]. Mutant CX26 also exerts a *trans*-dominant-negative effect on other connexins[24], providing a useful tool for eliminating broad gap junction channel functionality[23]. Notably, this point mutation should not interfere with gap-junction-mediated adhesion that is required for the proper radial migration of newly generated excitatory neurons in the developing neocortex[10]. We therefore engineered retroviruses expressing either wild-type CX26 or the dominant-negative closed channel mutant CX26(Thr135Ala), together with eGFP, using an internal ribosomal entry site (IRES) sequence, and performed *in utero* intraventricular injection of the retroviruses at E12–E13.

We found that expression of wild-type CX26 had no discernible effect, but expression of CX26(Thr135Ala) largely eliminated the electrical coupling between sister excitatory neurons at P1–P5 (CX26, coupling frequency 26.4%, 14 of 53 pairs; CX26(Thr135Ala), 9.8%, 4 of 41 pairs; Supplementary Fig. 4). More importantly, when we examined the chemical synapses formed between sister excitatory neurons at P10–P21, we found that only ~7.0% (6 of 86) of the sister pairs expressing the dominant-negative closed channel mutant CX26(Thr135Ala) were connected by chemical synapses (Fig. 4d–g), compared to 33–36% of the sister pairs expressing either eGFP alone (12 of 33 pairs tested) or wild-type CX26 (34 of 102 pairs tested) (Fig. 4a–c, g). The expression of CX26(Thr135Ala) did not affect neocortical excitatory neuron migration[10] (Supplementary Fig. 5), maturation (Supplementary Fig. 6) or chemical synapse formation in general (Supplementary Figs 7 and 8). Taken together, these results suggest that the blockade of electrical communication between sister excitatory neurons impairs the subsequent formation of chemical synapses between them.

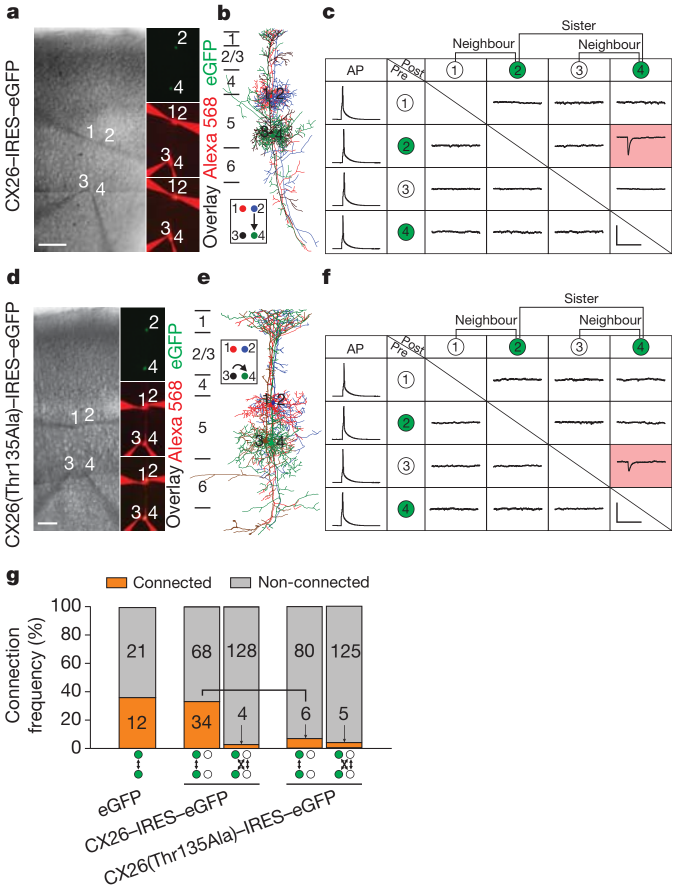


Figure 4 | Preferential electrical coupling is required for chemical synapse formation between sister excitatory neurons in neocortical ontogenetic columns. **a**–**f**, Quadruple whole-cell recordings of two sister excitatory neurons (2 and 4) in ontogenetic columns expressing CX26–IRES–eGFP (**a**–**c**) or CX26(Thr135Ala)–IRES–eGFP (**d**–**f**) and of two adjacent non-sibling excitatory neurons (1 and 3). DIC and fluorescence images (**a**, **d**) and morphological reconstruction (**b**, **e**) of the respective quadruple whole-cell recordings are shown. The numbers 1 to 6 (left in **b** and **e**) indicate layers 1–6. Scale bars, 100 μm (**a**, **d**). A summary of the chemical synaptic connections detected in the quadruple recordings is also shown (**c**, **f**). Pink shading indicates the existence of chemical synapses. Scale bars, 10 pA and 200 ms (**c**, **f**). AP, action potential; post, postsynaptic neuron; pre, presynaptic neuron. **g**, Summary of the frequency of chemical synapse formation between sister excitatory neurons in ontogenetic columns expressing eGFP, CX26–IRES–eGFP or CX26(Thr135Ala)–IRES–eGFP and their adjacent non-eGFP-expressing excitatory neurons. The bracket highlights the difference in the connectivity of CX26–IRES–eGFP-expressing sister neuron pairs and CX26(Thr135Ala)–IRES–eGFP-expressing sister neuron pairs.

Transient electrical coupling occurs before the establishment of mature patterns of synaptic connectivity in many developing nervous systems, and in some cases, this electrical coupling is crucial for the development of chemical synapses[25–27]. Previously, dye coupling was observed in the developing neocortex[2,6,14,15]. In addition, it has been suggested that the formation of electrically coupled neuronal domains might help to guide the emergence of chemically transmitting neuronal circuits[7,8]. However, electrically coupled neuronal pairs have not been reported in the neocortex at early developmental stages[19]. By performing dual and quadruple whole-cell recording experiments, which allow the detection of gap-junction-mediated electrical coupling with high sensitivity and spatial precision[12,18], we demonstrated the electrical coupling of excitatory neurons in the early postnatal neocortex. Moreover, we revealed that neocortical excitatory neurons show a high preference for forming strong electrical coupling with their sister excitatory neurons but not with nearby non-sister excitatory neurons.

Furthermore, we found that strong electrical coupling between sister excitatory neurons in ontogenetic columns promotes their action potential generation and synchronous firing. Although previous studies have suggested that electrical coupling is crucial for robust synchronous activity in the neonatal neocortex[8,28,29], the precise function of electrical coupling has been elusive. Our results show that electrical transmission between sister excitatory neurons in ontogenetic columns is required for the development of precise chemical synapses between these neurons. These findings provide clear evidence of the role of gap junctions in regulating precise neuronal circuit assembly in the neocortex.

METHODS SUMMARY

Replication-incompetent Moloney murine leukaemia retroviruses expressing eGFP (obtained from F. H. Gage) or avian RCAS (replication-competent ASLV long terminal repeat with a splice acceptor) retroviruses expressing eGFP, CX26–IRES–eGFP or CX26(Thr135Ala)–IRES–eGFP were intraventricularly injected into E12–E13 CD-1 (Charles River Laboratories) or nestin-TVA-transgenic mouse embryos, respectively. Acute cortical slices were prepared at various postnatal stages, and multiple-electrode whole-cell recordings were performed on eGFP-expressing excitatory neurons in individual radial clones and on their excitatory neuron neighbours. Recordings were collected and analysed using two Axon Multiclamp 700B amplifiers and pCLAMP 10 (Molecular Devices) and IGOR 5 (WaveMetrics) software. Images were collected by confocal laser scanning microscopy (FluoView FV1000, Olympus) and analysed using FluoView (Olympus), Neurolucida (MicroBrightField), Imaris (Andor Technology) and Photoshop (Adobe). Data are presented as mean ± s.e.m., and statistical differences were determined using non-parametric statistical tests: the Mann–Whitney–Wilcoxon and Kruskal–Wallis tests.

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

Received 22 September 2011; accepted 14 February 2012.
Published online 2 May 2012.

1. Kriegstein, A. & Alvarez-Buylla, A. The glial nature of embryonic and adult neural stem cells. *Annu. Rev. Neurosci.* **32,** 149–184 (2009).
2. Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S. & Kriegstein, A. R. Neurons derived from radial glial cells establish radial units in neocortex. *Nature* **409,** 714–720 (2001).
3. Rakic, P. Specification of cerebral cortical areas. *Science* **241,** 170–176 (1988).
4. Luskin, M. B., Pearlman, A. L. & Sanes, J. R. Cell lineage in the cerebral cortex of the mouse studied *in vivo* and *in vitro* with a recombinant retrovirus. *Neuron* **1,** 635–647 (1988).
5. Yu, Y. C., Bultje, R. S., Wang, X. & Shi, S. H. Specific synapses develop preferentially among sister excitatory neurons in the neocortex. *Nature* **458,** 501–504 (2009).
6. Peinado, A., Yuste, R. & Katz, L. C. Extensive dye coupling between rat neocortical neurons during the period of circuit formation. *Neuron* **10,** 103–114 (1993).
7. Yuste, R., Peinado, A. & Katz, L. C. Neuronal domains in developing neocortex. *Science* **257,** 665–669 (1992).
8. Yuste, R., Nelson, D. A., Rubin, W. W. & Katz, L. C. Neuronal domains in developing neocortex: mechanisms of coactivation. *Neuron* **14,** 7–17 (1995).
9. Willecke, K. *et al.* Structural and functional diversity of connexin genes in the mouse and human genome. *Biol. Chem.* **383,** 725–737 (2002).
10. Elias, L. A., Wang, D. D. & Kriegstein, A. R. Gap junction adhesion is necessary for radial migration in the neocortex. *Nature* **448,** 901–907 (2007).
11. Nadarajah, B., Jones, A. M., Evans, W. H. & Parnavelas, J. G. Differential expression of connexins during neocortical development and neuronal circuit formation. *J. Neurosci.* **17,** 3096–3111 (1997).
12. Connors, B. W. & Long, M. A. Electrical synapses in the mammalian brain. *Annu. Rev. Neurosci.* **27,** 393–418 (2004).
13. Bennett, M. V. & Zukin, R. S. Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron* **41,** 495–511 (2004).
14. Lo Turco, J. J. & Kriegstein, A. R. Clusters of coupled neuroblasts in embryonic neocortex. *Science* **252,** 563–566 (1991).
15. Connors, B. W., Benardo, L. S. & Prince, D. A. Coupling between neurons of the developing rat neocortex. *J. Neurosci.* **3,** 773–782 (1983).
16. Gibson, J. R., Beierlein, M. & Connors, B. W. Two networks of electrically coupled inhibitory neurons in neocortex. *Nature* **402,** 75–79 (1999).
17. Galarreta, M. & Hestrin, S. A network of fast-spiking cells in the neocortex connected by electrical synapses. *Nature* **402,** 72–75 (1999).
18. Wang, Y., Barakat, A. & Zhou, H. Electrotonic coupling between pyramidal neurons in the neocortex. *PLoS ONE* **5,** e10253 (2010).
19. Pangratz-Fuehrer, S. & Hestrin, S. Synaptogenesis of electrical and GABAergic synapses of fast-spiking inhibitory neurons in the neocortex. *J. Neurosci.* **31,** 10767–10775 (2011).
20. Stevens, C. F. & Zador, A. M. Input synchrony and the irregular firing of cortical neurons. *Nature Neurosci.* **1,** 210–217 (1998).
21. Hebb, D. O. *The Organization of Behavior* (Wiley, 1949).
22. Katz, L. C. & Shatz, C. J. Synaptic activity and the construction of cortical circuits. *Science* **274,** 1133–1138 (1996).
23. Beahm, D. L. *et al.* Mutation of a conserved threonine in the third transmembrane helix of α- and β-connexins creates a dominant-negative closed gap junction channel. *J. Biol. Chem.* **281,** 7994–8009 (2006).
24. Rouan, F. *et al. Trans*-dominant inhibition of connexin-43 by mutant connexin-26: implications for dominant connexin disorders affecting epidermal differentiation. *J. Cell Sci.* **114,** 2105–2113 (2001).
25. Personius, K. E. & Balice-Gordon, R. J. Loss of correlated motor neuron activity during synaptic competition at developing neuromuscular synapses. *Neuron* **31,** 395–408 (2001).
26. Chuang, C. F., Vanhoven, M. K., Fetter, R. D., Verselis, V. K. & Bargmann, C. I. An innexin-dependent cell network establishes left–right neuronal asymmetry in *C. elegans*. *Cell* **129,** 787–799 (2007).
27. Curtin, K. D., Zhang, Z. & Wyman, R. J. Gap junction proteins expressed during development are required for adult neural function in the *Drosophila* optic lamina. *J. Neurosci.* **22,** 7088–7096 (2002).
28. Dupont, E., Hanganu, I. L., Kilb, W., Hirsch, S. & Luhmann, H. J. Rapid developmental switch in the mechanisms driving early cortical columnar networks. *Nature* **439,** 79–83 (2006).
29. Kandler, K. & Katz, L. C. Coordination of neuronal activity in developing visual cortex by gap junction-mediated biochemical communication. *J. Neurosci.* **18,** 1419–1427 (1998).

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

Acknowledgements We thank the following: C. I. Bargmann, Y. Dan, A. L. Joyner, K. M. Hively and Y. Chin for comments on the manuscript; F. H. Gage for the 293gp NIT–GFP package cell line; E. C. Holland for nestin-TVA transgenic mice; and members of the Shi and Yu laboratories for their input. This work was supported by grants from the Ministry of Science and Technology of China (2012CB966300), the Natural Science Foundation of China (31121061 and 31070947), the Pujiang Talent Project of the Shanghai Science and Technology Committee (10PJ1400700), the Foundation of the Ministry of Education of China (20100071120061) (Y.-C.Y.), the National Institutes of Health (R01DA024681 and R21NS072483 (S.-H.S.), R21MH083624 (S.-H.S. and K.H.) and R01GM065947 (G.E.S.)), the McKnight Foundation and the March of Dimes Foundation (S.-H.S.).

Author Contributions Y.-C.Y., S.H. and S.-H.S. conceived the project. Y.-C.Y. and S.H. conducted the electrophysiology and morphology reconstruction experiments. S.C., Y.F. and K.N.B. generated viruses and performed *in utero* virus injections and morphological reconstructions. X.-H.Y., J.M. and K.P.G. performed immunohistochemistry and morphological reconstruction experiments. G.E.S. helped with CX26-carrying retrovirus engineering. K.H. performed cross-correlogram analysis. Y.-C.Y., S. H. and S.-H.S. analysed the data, interpreted the results and wrote the manuscript. All authors edited 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.-C.Y. (ycyu@fudan.edu.cn), S.H. (hes@mskcc.org) or S.-H.S. (shis@mskcc.org).

METHODS

Retrovirus production and *in utero* intraventricular injection. Replication-incompetent eGFP-expressing Moloney murine leukaemia virus was produced from a stably transfected packaging cell line (293gp NIT–GFP, obtained from F. H. Gage) as previously reported[2]. Animals were maintained according to protocols approved by the Institutional Animal Care and Use Committee at the Sloan-Kettering Institute for Cancer Research and by Fudan University. *In utero* intraventricular injection was performed as previously described[5]. In brief, the uterine horns of pregnant CD-1 mice (Charles River Laboratories) at the E12–E13 stage of gestation were exposed in a clean environment. Retrovirus ($\sim$1.0 µl) with Fast Green (2.5 mg ml^{-1}, Sigma) was injected into the embryonic cerebral ventricle through a bevelled, calibrated glass micropipette (Drummond Scientific). After injection, the peritoneal cavity was lavaged with $\sim$10 ml warm PBS (pH 7.4) containing antibiotics; the uterine horns were replaced; and the wound was closed. Avian RCAS (replication-competent ASLV long terminal repeat with a splice acceptor) retroviruses expressing CX26–IRES–eGFP or CX26(Thr135Ala)–IRES–eGFP were generated as previously described[30]. Similar *in utero* intraventricular injection of RCAS retroviruses was performed on the embryos of nestin-TVA-transgenic mice that were generated previously[31].

Immunohistochemistry and confocal imaging. For CX26 and TUJ1 immunohistochemistry, after intracardial perfusion with cold PBS (pH 7.4) and 4% paraformaldehyde (PFA) in PBS (pH 7.4), the brains were removed and embedded in the tissue freezing medium OCT compound (Electron Microscopy Sciences) after sucrose treatment. Coronal sections (20 µm) were prepared using a cryostat (Leica Microsystems) and incubated for 2 h at room temperature in a blocking solution (10% normal goat serum and 0.1% Triton X-100 in PBS), followed by incubation with primary antibodies, including rabbit anti-CX26 antibody (Invitrogen) and mouse anti-TUJ1 antibody (Covance) for 2 days at 4 °C. Sections were then washed in 0.1% Triton X-100 in PBS and incubated with the appropriate secondary antibodies overnight at 4 °C. For three-dimensional reconstruction, *Z*-series images were taken at 0.1-µm steps using a 100× objective lens in an Olympus FluoView FV1000 confocal laser scanning microscope and were analysed using Neurolucida (MicroBrightField) and Imaris (Andor Technology).

Slice preparation, electrophysiological recording and data analysis. Embryos that received retroviral injections were delivered naturally. Brains were removed at various times after birth, and acute cortical slices (200–300 µm thick) were prepared in artificial cerebrospinal fluid (ACSF) containing 126 mM NaCl, 3 mM KCl, 1.25 mM KH_2PO4, 1.3 mM $MgSO_4$, 3.2 mM $CaCl_2$, 26 mM $NaHCO_3$ and 10 mM glucose, bubbled with 95% O_2 and 5% CO_2, with a vibratome (Leica Microsystems) at 4 °C. Slices were allowed to recover in an interface chamber at 35 °C for at least 1 h and were then kept at room temperature before being transferred to a recording chamber containing ACSF at 34 °C. An infrared DIC microscope (Olympus) equipped with epifluorescence illumination, a charge-coupled device camera and two water immersion lenses (10× and 60×) was used to visualize and target the recording electrodes to eGFP-expressing sister cells in ontogenetic columns and to their nearby control cells. Glass recording electrodes (20–30 MΩ resistance) were filled with an intracellular solution consisting of 130 mM potassium gluconate, 6 mM KCl, 2 mM $MgCl_2$, 0.2 mM EGTA, 10 mM HEPES, 2.5 mM Na_2ATP, 0.5 mM Na_2GTP, 10 mM potassium phosphocreatine and 0.3% Alexa Fluor 568 hydrazide (Invitrogen) (pH 7.25 and 295 mOsmol per kg solution). Recordings were collected and analysed using an Axon Multiclamp 700B amplifier and pCLAMP 10 software (Molecular Devices). In all dual and quadruple recordings, electrical coupling was assessed by injecting currents to trigger hyperpolarization or depolarization under current-clamp mode or by voltage steps under voltage-clamp mode. In some experiments, 100 µM MFA (Sigma) was added to the bath to block gap junctions, and 10 µM BMI, 20 µM CPP and 10 µM CNQX (Tocris Bioscience) were used to block $GABA_A$, NMDA and AMPA receptors, respectively. MFA was applied for 10–30 min to block gap junctions and was later washed out for 20–30 min to test the recovery of the gap junctions. The gap junction electrical conductance was estimated under voltage-clamp conditions[32]. Native neuronal activity stimuli were simulated as previously described[20].

Chemical connections between neuron pairs were assessed by injecting current to induce action potentials in one of the neurons kept in current clamp while testing for postsynaptic responses in other neurons under the voltage-clamp recording condition at −70 mV. Glass recording electrodes (8–12 MΩ resistance) were filled with an intracellular solution consisting of 130 mM potassium gluconate, 6 mM KCl, 2 mM $MgCl_2$, 0.2 mM EGTA, 10 mM HEPES, 2.5 mM Na_2ATP, 0.5 mM Na_2GTP, 10 mM potassium phosphocreatine and 0.5% neurobiotin. For every possible pair, the connections were tested in both directions for at least 20 trials, with both single action potentials and trains of action potentials being generated in each presynaptic neuron. For miniature excitatory postsynaptic current (mEPSC) and miniature inhibitory postsynaptic current (mIPSC) analysis, cells were clamped at −70 mV, and recordings were performed in the presence of 50 µM BMI or10 µM CNQX/20 µM D-AP5 together with 1 µM tetrodotoxin, respectively. The decay time constant, τ, was estimated by single-exponential function fitting: ($f(t) = A{*}\exp(-t/\tau) + C$), where A and C are constants and t is time. In whole-cell patch-clamp recording experiments, slices were fixed in 4% PFA in PBS (pH 7.4) after the recordings were completed, and the morphology of recorded neurons that had been loaded with Alexa Fluor 568 hydrazide through the recording pipette was visualized using an Olympus FluoView FV1000 confocal laser scanning microscope. *Z*-series images were taken at 1–3-µm steps and analysed using FluoView (Olympus), Neurolucida (MicroBrightField) and Photoshop (Adobe). In some recording experiments, neurobiotin was later visualized with Alexa-Fluor-647- or Alexa-Fluor 568-conjugated streptavidin (Invitrogen).

Normalized cross-correlograms of firing patterns were analysed as previously described[33]. In brief, the number of times that neuron 1 fired within a time interval ($n\Delta t$, $(n+1)\Delta t$) from spikes fired by neuron 2 was calculated (and is denoted y_n, which is the number of counts per bin, where the bin width is $\Delta t = 1$ ms). The cross-correlogram y_n was normalized to standard scores: $Z = (y_n - \gamma_E)/s_y$, where $\gamma_E = f_1{*}f_2{*}T{*}\Delta t$, and $f_{1,2}$ is the average firing rate of neurons 1 and 2, T is the recording time and s_y is the standard deviation of y_n. Data are presented as mean ± s.e.m., and statistical differences were determined using non-parametric statistical tests: the Mann–Whitney–Wilcoxon and Kruskal–Wallis tests. Peaks in the cross-correlogram were considered significant if individual bins exceeded the expected value by three standard deviations (that is, if the Z score was >3).

30. Du, Z. *et al.* Introduction of oncogenes into mammary glands *in vivo* with an avian retroviral vector initiates and promotes carcinogenesis in mouse models. *Proc. Natl Acad. Sci. USA* **103,** 17396–17401 (2006).
31. Holland, E. C., Hively, W. P., DePinho, R. A. & Varmus, H. E. A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes Dev.* **12,** 3675–3685 (1998).
32. Neyton, J. & Trautmann, A. Single-channel currents of an intercellular junction. *Nature* **317,** 331–335 (1985).
33. Vos, B. P., Maex, R., Volny-Luraghi, A. & De Schutter, E. Parallel fibers synchronize spontaneous activity in cerebellar Golgi cells. *J. Neurosci.* **19,** RC6 (1999).

doi:10.1038/nature11110

Clonally related visual cortical neurons show similar stimulus feature selectivity

Ye Li[1,2]*, Hui Lu[1,2]*, Pei-lin Cheng[1], Shaoyu Ge[3], Huatai Xu[4], Song-Hai Shi[4] & Yang Dan[1,2]

A fundamental feature of the mammalian neocortex is its columnar organization[1]. In the visual cortex, functional columns consisting of neurons with similar orientation preferences have been characterized extensively[2–4], but how these columns are constructed during development remains unclear[5]. The radial unit hypothesis[6] posits that the ontogenetic columns formed by clonally related neurons migrating along the same radial glial fibre during corticogenesis[7] provide the basis for functional columns in adult neocortex[1]. However, a direct correspondence between the ontogenetic and functional columns has not been demonstrated[8]. Here we show that, despite the lack of a discernible orientation map in mouse visual cortex[4,9,10], sister neurons in the same radial clone exhibit similar orientation preferences. Using a retroviral vector encoding green fluorescent protein to label radial clones of excitatory neurons, and *in vivo* two-photon calcium imaging to measure neuronal response properties, we found that sister neurons preferred similar orientations whereas nearby non-sister neurons showed no such relationship. Interestingly, disruption of gap junction coupling by viral expression of a dominant-negative mutant of Cx26 (also known as Gjb2) or by daily administration of a gap junction blocker, carbenoxolone, during the first postnatal week greatly diminished the functional similarity between sister neurons, suggesting that the maturation of ontogenetic into functional columns requires intercellular communication through gap junctions. Together with the recent finding of preferential excitatory connections among sister neurons[11], our results support the radial unit hypothesis and unify the ontogenetic and functional columns in the visual cortex.

To identify clonally related sister cells, we used a green fluorescent protein (GFP)-expressing retrovirus, previously shown to label isolated ontogenetic columns of excitatory neurons[7,11,12]. The retrovirus was injected into the right ventricle *in utero* at embryonic day 15–17 (E15–17; see Methods), the beginning of neurogenesis in cortical layer 2/3 (ref. 13). At postnatal day 12–17 (P12–17, soon after eye opening), *in vivo* two-photon imaging[14,15] was performed in the primary visual cortex (V1) of injected mice under anaesthesia. A low density of GFP-labelled neurons was observed in layer 2/3 (1.1 ± 0.9 (standard deviation) per animal, within an imaging window ∼500 μm in diameter at cortical depths up to 400 μm, $n = 181$ neurons, 161 mice). In some cases ($n = 52$), we found a pair of GFP-labelled neurons aligned nearly vertically (Fig. 1a, b), with no other GFP neurons nearby, suggesting that they were clonally related sister cells. Although large tangential dispersion has been observed in some clonally related cells[16], here we focused on GFP-labelled neuronal pairs with <120 μm horizontal separation (see Methods; Supplementary Fig. 1).

To examine the functional properties of layer 2/3 neurons, we injected the calcium indicator dye Oregon Green BAPTA-1 AM (OGB-1) into a region encompassing the GFP-labelled cell pair. Orientation and direction selectivity of OGB-1-loaded neurons was measured with drifting grating stimuli (100% contrast, spatial frequency 0.02–0.03 cycles per degree, temporal frequency, 1–2 Hz) presented through the contralateral eye. The mapping was made at two cortical depths to include both GFP-labelled neurons in the sister pair. We found that 875 of the 2,286 OGB-1-loaded neurons (38%) showed significant increases in intracellular calcium in response to the grating stimuli

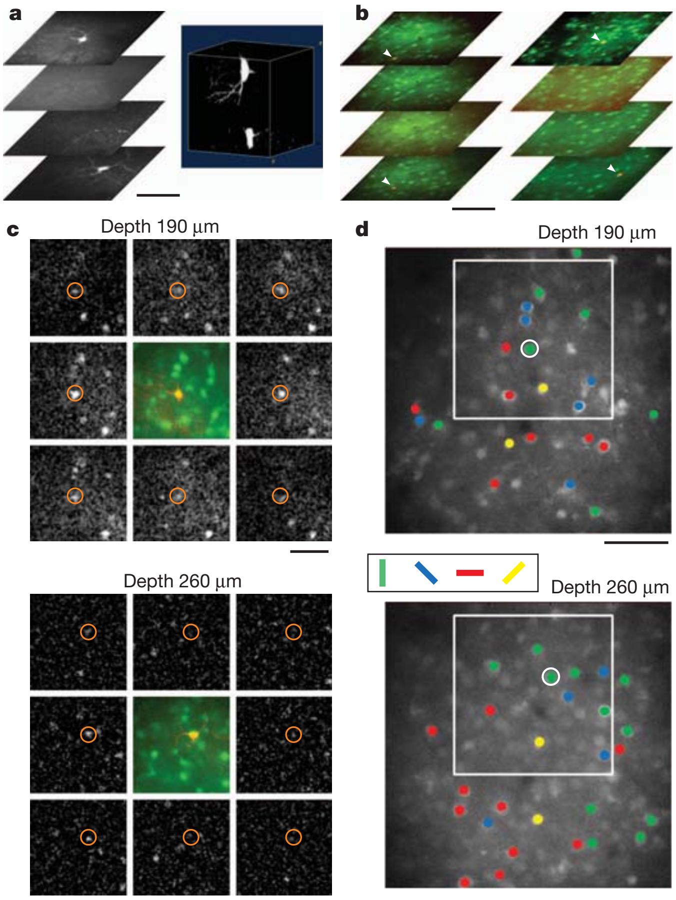


Figure 1 | Two-photon imaging of clonally related sister cells and nearby layer 2/3 neurons. **a**, Left, fluorescence images at 190–260 μm from pia, showing two GFP-expressing cells that were nearly vertically aligned. Scale bar, 50 μm. Right, three-dimensional reconstruction of the GFP pair. **b**, Two more examples of GFP cell pairs (arrowheads). Red, GFP; green, OGB-1. Left, 250–310 μm from pia; right, 150–240 μm. Scale bar, 100 μm. **c**, Single condition maps of fluorescence change (ΔF), computed by averaging the images during each stimulus and subtracting baseline (grey screen) for the experiment in **a**. Central panels, GFP (red) and OGB-1 (green) labelling. Red circles, GFP-labelled cells. Scale bar, 50 μm. **d**, Orientation maps with visually responsive cells coloured according to their preferred orientations for a larger imaging area. White box, region shown in **c**. Scale bar, 50 μm.

[1] Division of Neurobiology, Department of Molecular and Cell Biology, Helen Wills Neuroscience Institute, University of California, Berkeley, California 94720, USA. [2] Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [3] Department of Neurobiology & Behavior, State University of New York at Stony Brook, Stony Brook, New York 11794, USA. [4] Developmental Biology Program, Memorial Sloan-Kettering Cancer Centre, 1275 York Avenue, New York, New York 10065, USA.
*These authors contributed equally to this work.

(see Methods; Fig. 1c). Among these visually driven neurons, 75% (657/875) showed significant orientation selectivity ($P < 0.05$, Hotelling's T-squared test), comparable to previous studies in rodent visual cortex[4,17]. Notably, nearby neurons often preferred different orientations with no apparent spatial organization (Fig. 1c, d), consistent with previous findings of a 'salt-and-pepper' arrangement of orientation preferences in rodent visual cortex[4,10].

Comparing the response properties of the sister cells, however, we found that they often preferred similar orientations (Fig. 1c, d, circles). To quantify this relationship, we fitted the tuning curve of each visually driven neuron with a double Gaussian function (Fig. 2a and Supplementary Fig. 2) to identify its preferred orientation (θ). The functional similarity between each cell pair was measured by the difference between their preferred orientations ($\Delta\theta$, varying between 0° and 90°). Of the 34 sister pairs in which both neurons were visually driven, 20 pairs (59%) preferred similar orientations ($0° < \Delta\theta < 30°$), and only 5 pairs (15%) preferred near orthogonal orientations ($60° < \Delta\theta < 90°$; Fig. 2b). The distribution of $\Delta\theta$ was significantly non-uniform ($P = 0.0071$, Kolmogorov–Smirnov test), with a strong bias towards 0. In contrast, for pairs of non-sister neurons with horizontal distance <120 μm, the distribution of $\Delta\theta$ was largely flat (Fig. 2c), significantly different from the sister pairs (Fig. 2d; $P = 0.018$). The slight biases of the distribution for non-sister pairs towards both 0° and 90° were caused by the overrepresentation of cardinal orientations (horizontal and vertical) in mouse V1 soon after eye opening[17].

When we restricted the analysis to cell pairs with significant orientation selectivity ($n = 23$), the distribution of $\Delta\theta$ for sister pairs showed a more marked bias towards 0 (Fig. 2b, filled bars), significantly different from both the uniform distribution ($P = 0.0038$) and the distribution for orientation-selective non-sister pairs (Fig. 2c, filled bars; $P = 0.034$). Furthermore, even with the sparse labelling, our population of sister pairs could still be contaminated by GFP-labelled neurons from separate but neighbouring radial clones. Thus, the similarity between true sisters could be even stronger than that observed here.

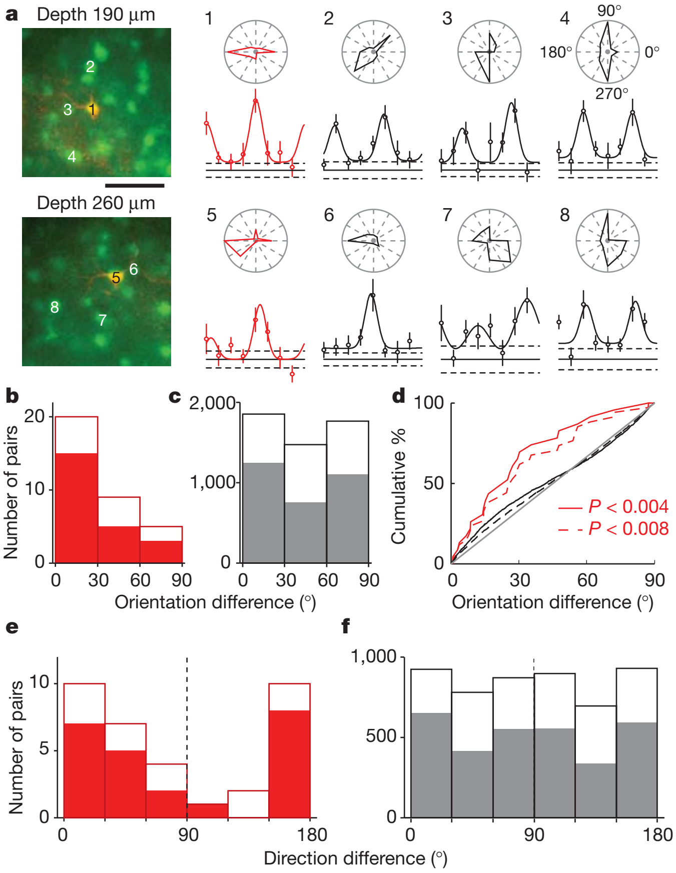


Figure 2 | Orientation and direction preference of sister neurons. **a**, Orientation tuning of the GFP pair (red) and several nearby cells (indicated by numbers) in the experiment shown in Fig. 1a, c, d. Circle, mean; vertical line, ± s.e.m.; curve, fitted double Gaussian function. Horizontal lines, baseline (solid) ± s.e.m. (dashed). Inset above, polar plot of orientation tuning. Scale bar, 50 μm. **b**, Histogram distribution of difference in preferred orientation ($\Delta\theta$) between sister cells, for all visually driven (open bars) and orientation selective (filled bars) pairs. **c**, Histogram of $\Delta\theta$ between non-sister (GFP/non-GFP and non-GFP/non-GFP) pairs, for all visually driven (open) and orientation-selective (filled) pairs. **d**, Cumulative distribution of $\Delta\theta$ for sister and non-sister pairs. Red, sisters (P values, difference from uniform distribution); black, non-sisters. Dashed, all visually driven pairs; solid, orientation-selective pairs. Grey, diagonal line. **e**, **f**, Histograms of difference in preferred direction between sister cells (**e**) and non-sister cells (**f**).

In addition to similar orientation tuning, the sister neurons also showed a modest tendency to prefer similar directions. When their difference in preferred direction was plotted over the range 0–180°, we found more pairs falling between 0° and 90° (21/34 visually driven, 14/23 orientation-selective pairs) than between 90° and 180° (Fig. 2e), although the difference was not statistically significant ($P = 0.11$ for visually driven, $P = 0.26$ for significantly tuned pairs, bootstrap). For the non-sister pairs, the distribution was largely symmetrical (Fig. 2f).

We next explored the mechanism that confers sister neurons with similar functional properties. Previous studies in developing cortical slices have revealed spontaneous co-activation of neurons within discrete, radially oriented domains spanning multiple cortical layers, which is mediated by gap junction coupling between the neurons[18]. These domains are comparable to the radial clones in shape and size, and they could provide a blueprint for the functional columns by influencing the formation and fine tuning of chemical synapses. To test this idea, we examined the effect of disrupting gap junction coupling between cortical neurons on orientation tuning. Among all the genes encoding the gap junction protein connexin, *Cx26* was found to be highly expressed in developing neocortex and strongly associated with interneuronal coupling[19]. We thus injected *in utero* a retrovirus expressing a mutant Cx26, with a threonine in the third transmembrane helix (T135) replaced by alanine[20] (Cx26(T135A)–T2A–EGFP). Because the mutant Cx26 forms closed gap junction channels and exerts a trans-dominant-negative effect on other connexins[21], it provides a useful tool for selective disruption of gap junction communication in a small number of cortical neurons (which in this case were labelled with enhanced (E)GFP). When we measured orientation tuning at P12–17, we found that 35% (585/1,674) of the OGB-1-loaded neurons were visually responsive, among which 72% (423/585) showed significant orientation selectivity ($P < 0.05$, Hotelling's T-squared test), similar to the mice injected with retrovirus expressing GFP alone. Thus, expression of the mutant Cx26 in a small number of neurons caused no global disruption of V1 responses. However, in contrast to the GFP control (Fig. 2b, d), the distribution of $\Delta\theta$ for sister pairs showed little bias towards 0 (Fig. 3b, d), not significantly different from uniform or the distribution for non-sister pairs (Fig. 3c; $P > 0.6$, Kolmogorov–Smirnov test). We also found no clear tendency of the sister cells to prefer similar directions (Fig. 3e, f). Thus, in addition to their roles in prenatal neuronal proliferation and migration[22], gap junctions may also be required for coordinating postnatal functional development of sister cells, through either electrical coupling or intercellular exchanges of small molecules[23].

In addition to the selective disruption of gap junction coupling in a small number of neurons with retrovirus, we also tested the effect of systemic application of a gap junction blocker, carbenoxolone (CBX), in the mice injected with retrovirus expressing GFP alone. As gap junction coupling between cortical neurons declines rapidly in the second postnatal week[24], we injected CBX (intraperitoneally, 10–20 mg kg^{-1}) daily for the first postnatal week to disrupt interneuronal communication during early postnatal development but not at the time of imaging. We found that CBX injection similarly disrupted the functional

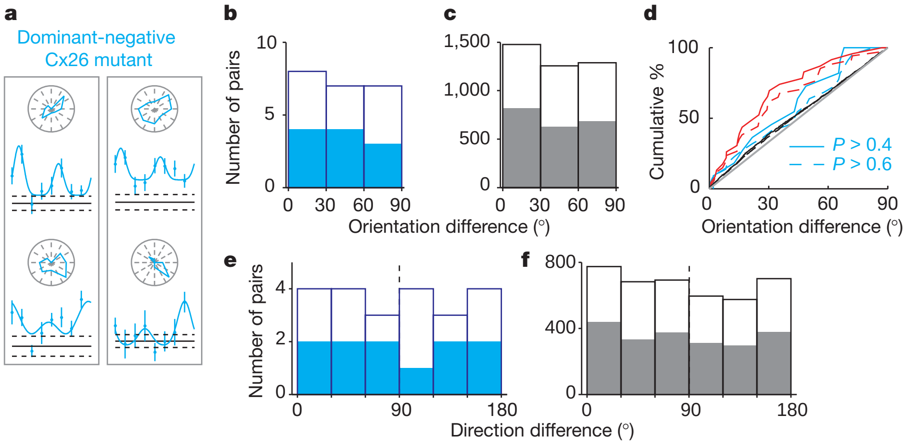


Figure 3 | Effect of expressing dominant-negative mutant of Cx26 in sister neurons. **a**, Two example experiments showing tuning curves of sister pairs (each pair in a box) in mice injected with retrovirus containing Cx26(T135A)–T2A–EGFP. **b**, Histogram of difference in preferred orientation ($\Delta\theta$) between sister neurons for all visually driven (open) and orientation-selective (filled) pairs in mice expressing mutant Cx26. **c**, Histogram of $\Delta\theta$ between non-sisters in these mice. **d**, Cumulative distribution of $\Delta\theta$ for sister pairs expressing mutant Cx26 (cyan; *P* values, comparison with uniform distribution), those expressing GFP only (red, same as Fig. 2d), and for non-sister pairs in mice injected with retrovirus containing Cx26(T135A)–T2A–EGFP (black). Dashed, all visually driven pairs; solid, orientation-selective pairs. **e**, **f**, Histograms of difference in preferred direction between sisters (**e**) and non-sisters (**f**).

similarity between sister neurons (Supplementary Fig. 3). However, CBX injection also caused an overall reduction in the percentage of visually driven neurons (225/1,121, 20%). This suggests that gap junctions may have an important role in early postnatal development of normal visual responses, although with systemic application of CBX it is difficult to rule out its potential non-specific effects on cortical development.

Intracortical excitatory connections are highly non-random[25], organizing the neurons into fine-scale subnetworks[26]. A recent study showed that sister neurons in the same radial clone are much more connected to each other than to nearby non-sister neurons[11], suggesting that the radial clones may provide a basis for subnetwork organization. The high connectivity between sister cells should also contribute to their functional similarity as observed in our study. However, inputs from the sister cells alone are likely to be insufficient to determine stimulus selectivity, as each neuron receives inputs from ~1,000 other neurons, whereas each radial clone only consists of tens of neurons[6]. Other factors, such as common inputs to the sister cells, may also have important roles. In mouse V1, layer 2/3 neurons with similar orientation tuning are shown to be preferentially interconnected[10]. A significant fraction of these neurons may be sister cells, exhibiting similar orientation tuning (Fig. 2) and preferential connectivity[11].

Although the columnar structure has long been thought to be a fundamental organizational principle of the neocortex, the existence of a basic processing unit has remained controversial[8]. Although the anatomical minicolumns observed in adult cortex[27] are believed to arise from ontogenetic columns, the relationship between the functional columns and mini/ontogenetic columns remained speculative[1]. Our results demonstrate a direct correspondence between them in V1, at least in superficial layers where neurons are most orientation selective[28]. Contrary to the notion of random organization, our study shows that orientation tuning is organized in columns even in rodent visual cortex. The fine spatial scale of ontogenetic columns may also explain the extraordinary precision of the orientation map in cat visual cortex[4,29]. The interspecies difference in macroscopic cortical organization may be due to differences in the horizontal connections between ontogenetic columns[1], which can lead to either a smoothly varying map or apparent salt-and-pepper organization[30]. Thus, our results support the view that the ontogenetic columns, rather than the macroscopic functional columns, constitute the basic units of cortical processing.

METHODS SUMMARY

Retrovirus was injected into the right ventricle of each mouse embryo *in utero* at E15–17. At P12–17, the injected mice were anaesthetized with urethane ($1\,\mathrm{g\,kg^{-1}}$) and chlorprothixene ($5\,\mathrm{mg\,kg^{-1}}$), sometimes supplemented with isoflurane (0.5–1% in O_2). A 1.5-mm-diameter craniotomy was made above V1 for two-photon imaging. After the GFP-expressing cells were identified, nearby layer 2/3 neurons were labelled with OGB-1 via bolus loading. For measuring orientation tuning, 8–12 repeats of drifting sinusoidal gratings were presented in 8 directions spanning 0–360° in a pseudo-random sequence. Neuronal responses were measured at the two depths where the GFP-expressing neurons were found. To quantify orientation and direction preference, we fitted each measured tuning curve by a double Gaussian function. To block gap junctions, $10\text{–}20\,\mathrm{mg\,kg^{-1}}$ CBX was injected daily (intraperitoneally) over the first postnatal week.

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

Received 13 October 2011; accepted 5 April 2012.
Published online 2 May 2012.

1. Mountcastle, V. B. The columnar organization of the neocortex. *Brain* **120,** 701–722 (1997).
2. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160,** 106–154 (1962).
3. Bonhoeffer, T. & Grinvald, A. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* **353,** 429–431 (1991).
4. Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P. & Reid, R. C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433,** 597–603 (2005).
5. White, L. E. & Fitzpatrick, D. Vision and cortical map development. *Neuron* **56,** 327–338 (2007).
6. Rakic, P. Specification of cerebral cortical areas. *Science* **241,** 170–176 (1988).
7. Noctor, S. C., Flint, A. C., Weissman, T. A., Dammerman, R. S. & Kriegstein, A. R. Neurons derived from radial glial cells establish radial units in neocortex. *Nature* **409,** 714–720 (2001).
8. Horton, J. C. & Adams, D. L. The cortical column: a structure without a function. *Phil. Trans. R. Soc. Lond. B* **360,** 837–862 (2005).
9. Schuett, S., Bonhoeffer, T. & Hubener, M. Mapping retinotopic structure in mouse visual cortex with optical imaging. *J. Neurosci.* **22,** 6549–6559 (2002).
10. Ko, H. *et al.* Functional specificity of local synaptic connections in neocortical networks. *Nature* **473,** 87–91 (2011).
11. Yu, Y. C., Bultje, R. S., Wang, X. & Shi, S. H. Specific synapses develop preferentially among sister excitatory neurons in the neocortex. *Nature* **458,** 501–504 (2009).
12. Cepko, C. L. *et al.* Studies of cortical development using retrovirus vectors. *Cold Spring Harb. Symp. Quant. Biol.* **55,** 265–278 (1990).
13. Polleux, F., Dehay, C. & Kennedy, H. The timetable of laminar neurogenesis contributes to the specification of cortical areas in mouse isocortex. *J. Comp. Neurol.* **385,** 95–116 (1997).
14. Denk, W., Strickler, J. H. & Webb, W. W. 2-photon laser scanning fluorescence microscopy. *Science* **248,** 73–76 (1990).
15. Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. *In vivo* two-photon calcium imaging of neuronal networks. *Proc. Natl Acad. Sci. USA* **100,** 7319–7324 (2003).
16. Walsh, C. & Cepko, C. L. Clonal dispersion in proliferative layers of developing cerebral cortex. *Nature* **362,** 632–635 (1993).
17. Rochefort, N. L. *et al.* Development of direction selectivity in mouse cortical neurons. *Neuron* **71,** 425–432 (2011).
18. Yuste, R., Peinado, A. & Katz, L. C. Neuronal domains in developing neocortex. *Science* **257,** 665–669 (1992).

19. Nadarajah, B., Jones, A. M., Evans, W. H. & Parnavelas, J. G. Differential expression of connexins during neocortical development and neuronal circuit formation. *J. Neurosci.* **17,** 3096–3111 (1997).
20. Beahm, D. L. *et al.* Mutation of a conserved threonine in the third transmembrane helix of α- and β-connexins creates a dominant-negative closed gap junction channel. *J. Biol. Chem.* **281,** 7994–8009 (2006).
21. Rouan, F. *et al.* Trans-dominant inhibition of connexin-43 by mutant connexin-26: implications for dominant connexin disorders affecting epidermal differentiation. *J. Cell Sci.* **114,** 2105–2113 (2001).
22. Elias, L. A. & Kriegstein, A. R. Gap junctions: multifaceted regulators of embryonic cortical development. *Trends Neurosci.* **31,** 243–250 (2008).
23. Kandler, K. & Katz, L. C. Coordination of neuronal activity in developing visual cortex by gap junction-mediated biochemical communication. *J. Neurosci.* **18,** 1419–1427 (1998).
24. Peinado, A., Yuste, R. & Katz, L. C. Extensive dye coupling between rat neocortical neurons during the period of circuit formation. *Neuron* **10,** 103–114 (1993).
25. Song, S., Sjostrom, P. J., Reigl, M., Nelson, S. & Chklovskii, D. B. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.* **3,** e68 (2005).
26. Yoshimura, Y., Dantzker, J. L. & Callaway, E. M. Excitatory cortical neurons form fine-scale functional networks. *Nature* **433,** 868–873 (2005).
27. Rockel, A. J., Hiorns, R. W. & Powell, T. P. The basic uniformity in structure of the neocortex. *Brain* **103,** 221–244 (1980).
28. Niell, C. M. & Stryker, M. P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28,** 7520–7536 (2008).
29. Ohki, K. *et al.* Highly ordered arrangement of single neurons in orientation pinwheels. *Nature* **442,** 925–928 (2006).
30. Koulakov, A. A. & Chklovskii, D. B. Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. *Neuron* **29,** 519–527 (2001).

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

Acknowledgements This work was supported by National Institutes of Health (NIH) grant R01 EY018861 and NSF grant 22250400-42533 (to Y.D.), and NIH grants R01 DA024681 and R21NS072483 (to S.-H.S.). We thank L. E. White and S. D. Van Hooser for comments on the manuscript, A. Kwan and S. D. Van Hooser for help with two-photon imaging techniques and analysis and Y. Gu for help in making retrovirus.

Author Contributions Y.L. performed the two-photon imaging experiments and data analysis. H.L., Y.L. and P.-I.C. performed *in utero* virus injection. S.G., H.X. and S.-H.S. provided the viral vectors. Y.L., H.L. and Y.D. designed the experiments and 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 Y.D. (ydan@berkeley.edu).

METHODS

Retroviral infection. Replication-incompetent retrovirus expressing either GFP alone or a loss-of-function mutant of connexin 26 and EGFP (Cx26(T135A)–T2A–EGFP)[20] was produced as previously described[31]. Uterine horns of E15–17 gestation stage pregnant female C57BL/6 mice (Charles River Laboratories) were exposed in a clean environment. Retrovirus (0.5–1 μl) with fast green (2.5 mg ml^{-1}; Sigma) was injected into the right embryonic cerebral ventricle at a speed of 150 nl s^{-1}, controlled by a microinjection pump (WPI). After injection, the peritoneal cavity was washed with warm saline solution containing antibiotics, the uterine horns were replaced, and the wound was closed. Both male and female mice were used in the experiments. All experimental procedures were approved by the Animal Care and Use Committee at the University of California, Berkeley.

Two-photon imaging. Mice were anaesthetized by intraperitoneal injection of urethane (1 g kg^{-1}) and chlorprothixene (5 mg kg^{-1}), in some cases supplemented with isoflurane (0.5–1% in O_2). Body temperature was maintained at 37 °C using a feedback heating pad. The head was secured using a stainless steel head plate affixed onto the skull using super glue and dental cement. A 1.5-mm-diameter craniotomy was performed at the location of the primary visual cortex (0–1 mm anterior to the lambda suture, 2–2.5 mm lateral of the midline). The dura was left intact. The cortical surface was constantly irrigated with an extracellular solution (in mM: 135 NaCl, 5 KCl, 5 HEPES, 1.8 $CaCl_2$ and 1 $MgCl_2$, at pH 7.3 and ~285 mOsm).

The neocortical neurons were labelled with calcium indicator dye via bolus loading[32]. The dye solution consists of 1 mM OGB-1, 10% dimethylsulphoxide (DMSO), 2% (wt/vol) Pluronic F-127 in HEPES-buffered saline (in mM: 150 NaCl, 2.5 KCl and 10 HEPES). Multiple injections of the dye solution were made in adjacent regions at a depth of ~200 μm. The experiment began 1 h after the dye injection. The two-photon microscope (Movable Objective Microscope, Sutter Instrument) was controlled using the ScanImage software[33]. The intensity of the excitation from a tunable femtosecond laser (Wideband, Tsunami Mode-Locked Ti: Sapphire Laser, Spectra-Physics) was controlled by a Pockels cell (350-80-LA-02; Conoptics). The excitation was focused using a 40×/0.8 NA infrared objective (LUMPLFLN, Olympus). Fluorescence was collected after a dichroic mirror (670DCXXR, Chroma) using a pentagon-style detector into green and red channels, with respective emission filters (FF01-510/84-25, Semrock; HQ610/75, Chroma) and photomultiplier tubes (GaAsP H10770PA-40 and multi-alkali R6357, Hamamatsu). Different excitation wavelengths were used to measure fluorescence of OGB-1 (800 nm) and GFP (900 nm). To image OGB-1 during visual stimulation, frames of 512 × 512 pixels were acquired continuously every 1.5–1.8 s.

Identification of sister neurons. As shown in Supplementary Fig. 1, the distribution of the horizontal distance between each pair of GFP-expressing neurons showed a prominent peak at <100 μm and a long tail. We chose a relatively conservative criterion of 120 μm (dashed line, Supplementary Fig. 1) for the pair of neurons to be considered sisters. In a previous study, the mean horizontal spread of the newly born neurons at P14–16 was found to be >500 μm (ref. 11). Thus, it is possible that some GFP-expressing sister neurons with large horizontal separation were excluded from our analysis. However, because increasing the criterion distance is likely to increase the probability of false positives (misclassification of non-sister pairs as sister pairs), in this study we chose to focus on cell pairs with small horizontal separations.

Of course, even with this relatively conservative criterion, one cannot exclude the possibility that some GFP-expressing neurons arising from different radial clones were close to each other simply by chance and were thus misclassified as sisters. As the distribution of $\Delta\theta$ for non-sisters was largely flat (Fig. 2c, f), the contamination of the sister-pair population by non-sister pairs should cause broadening of the observed distribution. Thus the similarity between true sisters in orientation and direction preference may be even stronger than that shown in Fig. 2b, e.

Visual stimulation. Visual stimuli were generated with a PC computer containing a NVIDIA GeForce 6600 graphics board and presented with a XENARC 700V LCD monitor (19.7 cm × 12.1 cm, 960 × 600 pixels, 75 Hz refresh rate, 300 cd m^{-2} maximum luminance, gamma corrected with custom software) located 14 cm from the left eye, positioned such that the receptive fields of the imaged neurons were at the centre of the monitor. For measuring orientation tuning and direction selectivity of V1 neurons, full-field drifting gratings (100% contrast, 1–2 Hz, 0.02–0.03 cycles per degree) were presented in 8 directions (separated by 45°) in a pseudorandom sequence. Each stimulus was 5 s in duration with a 5 s interstimulus interval. After each block of 8 drifting gratings, 5 s of blank stimulus (grey screen) was presented to measure the baseline activity. A total of 8–12 blocks were presented in each experiment.

Data analysis. Images were analysed with custom software written in Matlab. Small horizontal drift over time was corrected by measuring correlation at different pixel offsets and realigning the images according to the best match. Cells were identified by the experimenter. For each frame, the fluorescence value of each cell was computed by averaging all pixels in a circle (radius, 12 pixels, 5.5 μm) centred on the soma.

The response to each stimulus was calculated as $\frac{\Delta F}{F}=\frac{(F_{\mathrm{STIM}})-F_0}{F_0}$, where F_{STIM} is the average response across all frames when the stimulus is on, and F_0 is the average response during the final 3 s of the interstimulus period before stimulus onset. Cells were considered visually responsive if the responses to visual stimuli were different from the response to blank by ANOVA test (at $P<0.05$). Among cells that were visually responsive, cells with significant orientation selectivity were identified by plotting the response in each trial as a point in orientation space: $S(\theta_i)e^{\frac{2\pi i\theta_i}{180^\circ}}$, where $S(\theta_i)$ is the raw $\frac{\Delta F}{F}$ at direction θ_i. Cells were considered orientation selective if the mean of the cloud of points was significantly different from (0,0) by Hotelling's T-squared test (at $P<0.05$).

To identify the preferred orientation and direction of each cell, the responses to drifting gratings were fit with a 2-peak Gaussian function:

$$R(\theta)=R_{\mathrm{OFFSET}}+R_{\mathrm{PREF}}e^{-\frac{\mathrm{ang}(\theta-\theta_{\mathrm{PREF}})^2}{2\sigma^2}}+R_{\mathrm{OPP}}e^{-\frac{\mathrm{ang}(\theta+180-\theta_{\mathrm{PREF}})^2}{2\sigma^2}}$$

where R_{OFFSET} is a constant offset, θ_{PREF} is the preferred direction, R_{PREF} is the above-offset response to the preferred direction, R_{OPP} is the above-offset response to the opposite direction, σ is the tuning width and $\mathrm{ang}(x)=\min(x,x-360,x+360)$, which wraps angular difference values onto the interval 10° to 180°.

31. van Praag, H. *et al.* Functional neurogenesis in the adult hippocampus. *Nature* **415,** 1030–1034 (2002).
32. Garaschuk, O., Milos, R. I. & Konnerth, A. Targeted bulk-loading of fluorescent indicators for two-photon brain imaging *in vivo*. *Nature Protocols* **1,** 380–386 (2006).
33. Pologruto, T., Sabatini, B. & Svoboda, K. ScanImage: Flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2,** 13 (2003).

doi:10.1038/nature11033

α2δ expression sets presynaptic calcium channel abundance and release probability

Michael B. Hoppa[1], Beatrice Lana[2], Wojciech Margas[2], Annette C. Dolphin[2] & Timothy A. Ryan[1]

Synaptic neurotransmitter release is driven by Ca^{2+} influx through active zone voltage-gated calcium channels (VGCCs)[1,2]. Control of active zone VGCC abundance and function remains poorly understood. Here we show that a trafficking step probably sets synaptic VGCC levels in rats, because overexpression of the pore-forming $\alpha 1_A$ VGCC subunit fails to change synaptic VGCC abundance or function. α2δs are a family of glycosylphosphatidylinositol (GPI)-anchored VGCC-associated subunits[3] that, in addition to being the target of the potent neuropathic analgesics gabapentin and pregabalin (α2δ-1 and α2δ-2)[4,5], were also identified in a forward genetic screen for pain genes (α2δ-3)[6]. We show that these proteins confer powerful modulation of presynaptic function through two distinct molecular mechanisms. First, α2δ subunits set synaptic VGCC abundance, as predicted from their chaperone-like function when expressed in non-neuronal cells[3,7]. Second, α2δs configure synaptic VGCCs to drive exocytosis through an extracellular metal ion-dependent adhesion site (MIDAS), a conserved set of amino acids within the predicted von Willebrand A domain of α2δ. Expression of α2δ with an intact MIDAS motif leads to an 80% increase in release probability, while simultaneously protecting exocytosis from blockade by an intracellular Ca^{2+} chelator. α2δs harbouring MIDAS site mutations still drive synaptic accumulation of VGCCs; however, they no longer change release probability or sensitivity to intracellular Ca^{2+} chelators. Our data reveal dual functionality of these clinically important VGCC subunits, allowing synapses to make more efficient use of Ca^{2+} entry to drive neurotransmitter release.

VGCCs are composed of pore-forming α1 and auxiliary β and α2δ subunits[8,9]. In central synapses neurotransmitter release is generally driven by P/Q-type ($\alpha 1_A$) and/or N-type ($\alpha 1_B$)[10] VGCCs. On the basis of the failure of $\alpha 1_A$ overexpression to increase synaptic strength, it had been suggested that VGCCs functionally coupled to presynaptic release machinery is limited by a fixed number of available "slots" where channels can insert into the synaptic membrane[11]. We examined the existence of such a bottleneck by expressing $\alpha 1_A$ conjugated with enhanced green fluorescent protein (EGFP–$\alpha 1_A$, ref. 12) together with a reporter of presynaptic exocytosis (vGlut1 with luminal tag mOrange2, vGmOr2) and carried out retrospective immunocytochemistry to probe the abundance of $\alpha 1_A$ in transfected compared to control neurons. EGFP–$\alpha 1_A$ correctly trafficked to nerve terminals as it co-localized well with the vesicle-targeted reporter (Fig. 1a). To ensure EGFP–$\alpha 1_A$ functionally integrated with endogenous channels to drive neurotransmitter release, we introduced a point mutation (E1656K), rendering this channel insensitive to the antagonist ω-agatoxin IVA[13]. Under control conditions a combination of ω-agatoxin IVA and the $\alpha 1_B$ inhibitor ω-conotoxin GVIA completely blocked vGmOr2 responses to action potential firing; however, in the presence of EGFP–$\alpha 1_A^{E1656K}$, a significant fraction of the response remains (Fig. 1b). Measurements of single action potential responses showed that expression of this exogenous $\alpha 1_A$ did not alter exocytosis efficiency compared to controls (Fig. 1 c, d), consistent with the "slot" hypothesis[11]. However, retrospective immunocytochemistry using an anti-$\alpha 1_A$ antibody whose specificity was verified using short hairpin RNA-mediated $\alpha 1_A$ knockdown (Supplementary Fig. 1) showed that transfected and control nerve terminals had similar immunoreactivity (Fig. 1e, f) while at the cell soma it had doubled (Supplementary Fig. 2).

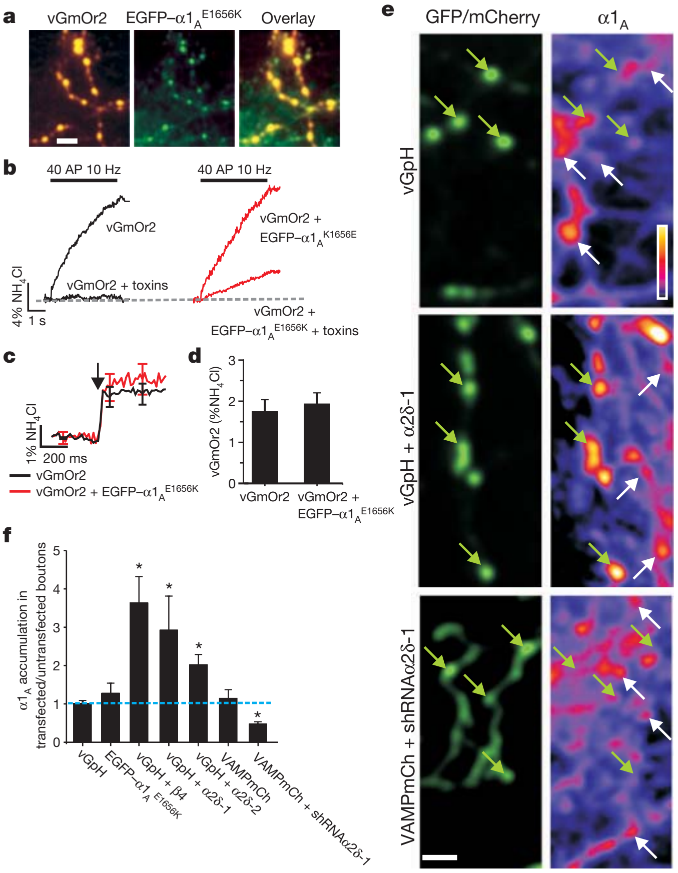


Figure 1 | Increased expression of α2δ and β subunits leads to increased P/Q Ca^{2+} channel accumulation at synapses. **a**, Coexpression of vGmOr2 (left), EGFP−$\alpha 1_A^{E1656K}$ (middle), overlay (right). **b**, Exocytic response for vGmOr2 alone (left; $n = 9$) and vGmOr2 coexpressed with EGFP–$\alpha 1_A^{E1656K}$ (right; $n = 6$) shows a significant toxin-resistant response (20 ± 7.1% of pre-toxin). **c**, Average traces of single action potential vGmOr2 responses with or without EGFP–$\alpha 1_A^{E1656K}$ ($n > 8$). Arrow indicates stimulation with one action potential. **d**, Average ΔF response for data in **c** (control 1.76 ± 0.28; $n = 14$; +EGFP−$\alpha 1_A^{E1656K}$ 1.91 ± 0.35; $n = 9$; $P > 0.1$). **e**, Presynaptic $\alpha 1_A$ abundance. Green arrows indicate transfected boutons, white arrows indicate non-transfected immunopositive $\alpha 1_A$ channel puncta. Inset, linear pseudo colour look up table (LUT) scale (yellow, high density; black, low density). **f**, Ratio of $\alpha 1_A$ immunofluorescence intensity in transfected puncta compared to untransfected puncta ($n \geq 8$ cells for all conditions). All stated values are mean ± s.e.m. Scale bar for all images, 4 μm.

[1] Department of Biochemistry, Weill Cornell Medical College, New York, New York 10023, USA. [2] Department of Neuroscience, Physiology and Pharmacology, Laboratory of Cellular and Molecular Neuroscience, University College London, London WC1E6BT, UK.

These results demonstrate that synaptic VGCC abundance is probably limited by trafficking from the cell soma, and that failure to increase synaptic performance does not result from a fixed number of active zone insertion sites. α2δ and β auxiliary VGCC subunits are both strong candidates for modulating such trafficking as they control functional expression of α1 subunits when coexpressed in non-neuronal cells[14,15]. We coexpressed individual auxiliary subunits with the reporter vGlut1–pHluorin (vGpH) in neurons and carried out measurements of exocytosis and immunocytochemistry as described above. These experiments demonstrated that expression of either α2δ-1 or β4 subunits led to a significant increase (~threefold, $P < 0.05$) in synaptic abundance of $\alpha1_A$ (Fig. 1e, f). Similar results were obtained with overexpression of α2δ-2 (Fig. 1f). Furthermore, introduction of shRNA targeting α2δ-1 caused depletion of $\alpha1_A$ at nerve terminals (Fig. 1e, f and Supplementary Fig. 3), while leaving the somatic concentration unaltered (data not shown). These results demonstrate that synaptic $\alpha1_A$ levels are titrated by expression of auxiliary VGCC subunits.

To examine whether changes in VGCC accumulation alters synaptic release properties, we measured single action-potential-stimulated exocytosis in neurons with altered VGCC levels. Overexpression and depletion of α2δ-1 led to much larger and much smaller single action potential responses, respectively, compared to control (Fig. 2a). Similar increases in exocytosis were observed following expression of all three isoforms of α2δ tested (Fig. 2b). In contrast, expression of β4 did not change exocytosis, despite the synaptic accumulation of $\alpha1_A$ (Fig. 2b). Quantitative estimates of α2δ-1 synaptic expression levels indicate a stoichiometric relationship between α2δ and $\alpha1_A$ (Supplementary Fig. 4). These results demonstrate that increasing VGCC abundance does not necessarily lead to increased function, and identify α2δ expression as a key rate-limiting parameter in determining presynaptic function.

Measurements of presynaptic strength can be parsed into two biophysical parameters: the number of vesicles available for rapid release upon stimulation, known as the readily-releasable pool (RRP), and the probability that a vesicle in the RRP will undergo fusion with a single action potential stimulus (Pv)[16]. We recently developed a rapid depletion protocol to measure RRP sizes using optical methods[17]. High frequency stimulation leads to rapid exhaustion of exocytosis in the first 8–15 action potentials. The fraction of the total pool corresponding to this rapid depletion phase is taken as the RRP[17] (Fig. 2c, d). The RRP size in neurons overexpressing α2δ was no different than controls (Fig. 2e). Thus α2δ overexpression changes Pv (Fig. 2f).

α2δ-driven increases in Pv might arise from increasing total Ca^{2+} influx (from changes in VGCC gating and/or surface abundance) and/or changing VGCC proximity to release sites[18]. To examine this question, we measured intracellular calcium concentration $[Ca^{2+}]$ at synaptic boutons in response to single action potentials using the fast fluorescent indicator of Ca^{2+}, Fluo5F-AM, visualized by expression of VAMP–mCherry (VAMPmCh; Fig. 3a left panel) with or without α2δ. Single action potentials resulted in robust Ca^{2+} signals (Fig. 3a right panel) that peaked within 1 ms, but were reduced by ~40% in synapses overexpressing α2δ isoforms compared to controls (Fig. 3b, c). We verified that the peak signal was not dominated by Ca^{2+} clearance mechanisms (for example, endogenous buffers or extrusion) in experiments where the signal decay was set by high concentrations of intracellular ethylene glycol tetra-acetic acid (EGTA). This treatment led to an approximately 50% decrease in peak signal and a decay time of about 10 ms in controls as well as α2δ-overexpressing synapses. Measurements of Ca^{2+} signals using a genetically encoded Ca^{2+} indicator GCaMP3 (ref. 19), co-expressed with or without α2δ-1, gave very similar results (Supplementary Fig. 5). This reduction in Ca^{2+} was surprising, given that α2δ overexpression increases the total number of synaptic VGCCs (surface and intracellular), indicating that α2δ might additionally control Ca^{2+} influx. Measurements of somatic action potential waveforms revealed that α2δ expression led to an approximately 30% decrease in action potential duration (Supplementary Fig. 6), providing a possible explanation for the drop in synaptic Ca^{2+} entry. Given that exocytosis at nerve terminals is steeply dependent on Ca^{2+} influx[20], the proximity of sites of Ca^{2+} influx to sites of exocytosis can, in principle, have a powerful influence over neurotransmitter release[2]. The increase in Pv with a commensurate decrease in Ca^{2+} influx strongly indicates that overexpression of α2δ subunits results in a tighter spatial relationship between sites of Ca^{2+} entry and exocytosis. We tested this hypothesis by measuring the sensitivity of exocytosis to the presence of a Ca^{2+} chelator EGTA-acetoxymethyl ester (EGTA-AM). The efficiency of the chelator in reducing exocytosis depends on its ability to buffer Ca^{2+} before it binds the calcium-sensor for exocytosis following VGCC opening, a process determined by chelator concentration and Ca^{2+} binding kinetics[21]. We chose incubation conditions for EGTA-AM that led to an approximately 50% reduction in single action potential exocytosis responses, compared to the pre-EGTA condition in control neurons (Fig. 3d). In neurons transfected with α2δ, however, EGTA application led to much smaller decreases in exocytosis (Fig. 3e), indicating that in conditions of α2δ overexpression Ca^{2+} must bind the calcium sensor more rapidly than in control conditions. Therefore, the Ca^{2+} sensor

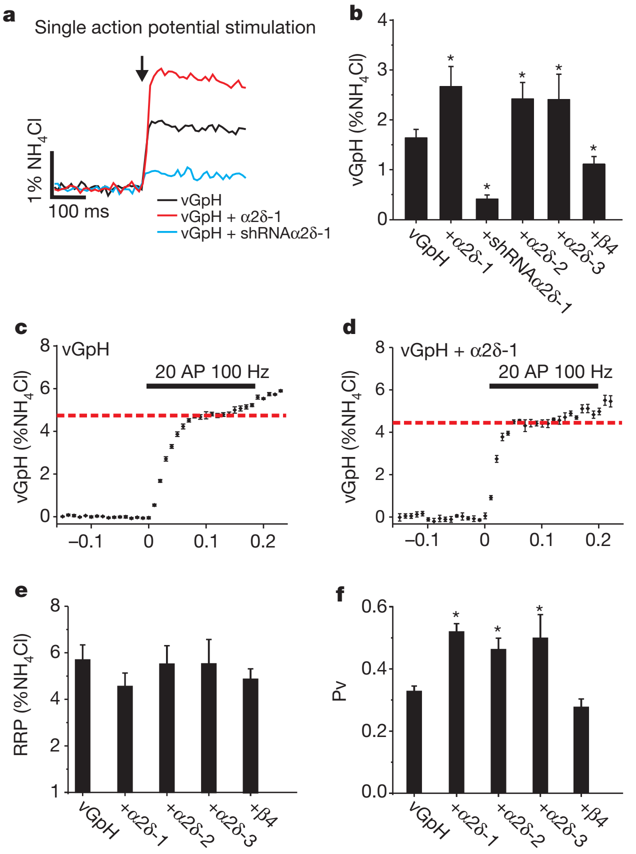


Figure 2 | Exocytosis is increased in neurons expressing α2δ. **a**, Representative single action potential vGpH responses (10 trial average, ~25 boutons). Arrow indicates stimulation with one action potential. **b**, Summary of single action potential response (%NH$_4$Cl): vGpH = 1.65 ± 0.17; vGpH + α2δ-1 = 2.71 ± 0.40; vGpH+shRNAα2δ-1 = 0.4 ± 0.08; vGpH + α2δ-2 = 2.44 ± 0.36; vGpH + α2δ-3, 2.42 ± 0.51; vGpH + β4 = 1.12 ± 0.15, $*P < 0.01$, $n \geq 7$. **c**, **d**, vGpH-based RRP measurements, dotted line identifies RRP. **e**, Summary of RRP size (%NH$_4$Cl): vGpH = 5.67 ± 0.64, vGpH + α2δ-1 = 4.6 ± 0.58, vGpH + α2δ-2 = 5.57 ± 0.78, vGpH + α2δ-3 = 5.58 ± 1.05, vGpH + β4 = 4.92 ± 0.45; ($P > 0.1$; $n \geq 7$). **f**, Pv measurements (vesicle fusion probability measured as ratio of (ΔF of one action potential)/(ΔF RRP)) ($*P < 0.01$): vGpH = 0.33 ± 0.02, vGpH + α2δ-1 0.52 ± 0.03, vGpH + α2δ-2 = 0.46 ± 0.04, vGpH + α2δ-3 = 0.50 ± 0.08, vGpH + β4 = 0.28 ± 0.03; $n \geq 7$. All stated values are mean ± s.e.m.

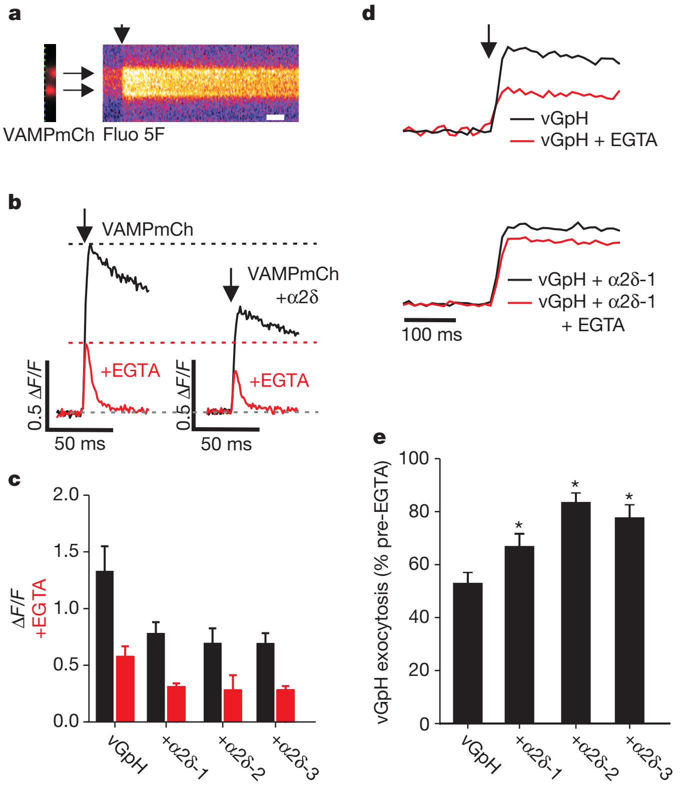


Figure 3 | α2δ leads to reduced Ca^{2+} influx and tighter coupling of calcium channels to exocytosis. **a**, Ca^{2+} influx stimulated by one action potential (vertical arrow) from boutons identified by VAMP-mCh (left), and visualized by Fluo5F (kymograph, right). Scale bar, 20 ms. **b**, Single traces of Ca^{2+} influx. **c**, Peak Fluo5F signal versus control ($*P < 0.05$): VAMP-mCh 1.3 ± 0.2; +α2δ-1 0.78 ± 0.09; +α2δ-2 0.69 ± 0.13; +α2δ-3 0.69 ± 0.13 ($n > 8$). $\Delta F/F$ of Ca^{2+} transient post EGTA (red): VAMP-mCh 0.58 ± 0.09; +α2δ-1 0.32 ± 0.03; +α2δ-2 0.28 ± 0.13; +α2δ-3 0.29 ± 0.03. **d**, Representative traces vGpH ± α2δ-2 stimulated by one action potential (vertical arrow). Traces normalized to response pre-EGTA treatment. **e**, Resistance to EGTA treatment (% exocytosis pretreatment): vGpH 53 ± 4; +α2δ-1 68 ± 6; +α2δ-2 83 ± 4; +α2δ-3 77 ± 5 $*P < 0.05$, $n \geq 7$ for all conditions. All stated values are mean ± s.e.m.

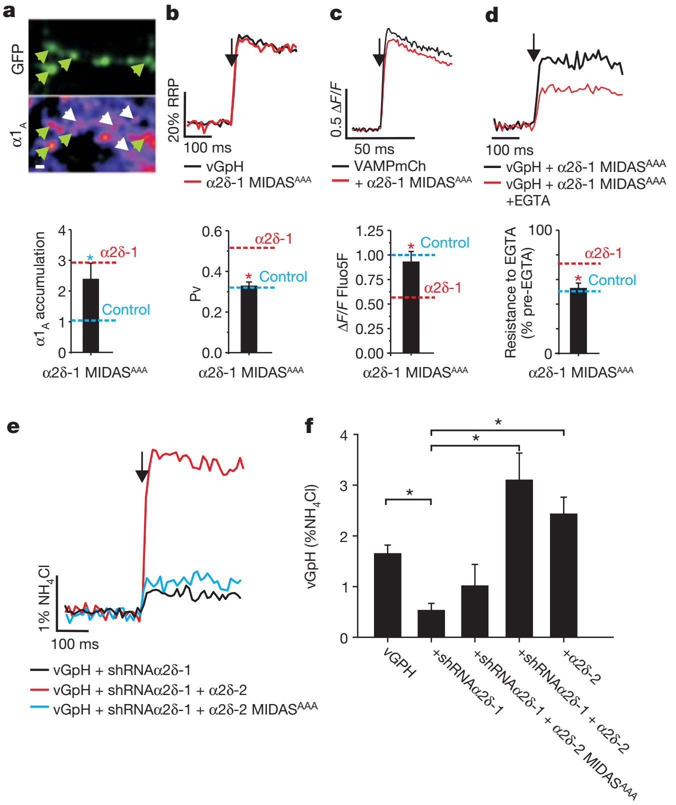


Figure 4 | α2δ MIDAS motif is essential for coupling Ca^{2+} channels to exocytosis. **a**, Top, presynaptic $\alpha 1_A$ abundance. Green arrows indicate transfected boutons, white arrows indicate non-transfected immunopositive $\alpha 1_A$ channel puncta. Scale bar, 2 μm. Bottom, ratio of $\alpha 1_A$ staining in synaptic boutons. Dashed lines represent ratios taken from Fig. 1f as indicated. **b**, Top, representative vGpH responses to one action potential (arrow) as a fraction of the measured RRP Bottom: vGpH and α2δ-1 $MIDAS^{AAA}$ (Pv = 0.33 ± 0.017) compared to data from Fig. 2f as indicated. **c**, Top, representative responses to one action-potential-driven Ca^{2+} influx (Fluo5F $\Delta F/F$). Bottom, peak one action potential Fluo 5F $\Delta F/F$ values in cells cotransfected with VAMPmCh ($n = 11$) and α2δ-1 $MIDAS^{AAA}$ (0.88 ± 0.1; $n = 6$) normalized to VAMPmCh alone ($*P < 0.05$). **d**, Top, representative vGpH response to one action potential in a neuron co-expressing α2δ-1 $MIDAS^{AAA}$ as indicated. Bottom, resistance to EGTA block (% block = 51 ± 5, $P = 0.63$) dashed lines compare data from Fig. 3e as indicated. **e**, Representative vGpH responses to one action potential. **f**, One action potential response (%NH_4Cl): vGpH = 1.65 ± 0.17; vGpH + shRNAα2δ-1 = 0.4 ± 0.08, vGpH + shRNAα2δ-1 + α2δ-2 = 3.10 ± .53; vGpH + shRNAα2δ-1 + α2δ-2 $MIDAS^{AAA}$ = 1.02 ± .41; vGpH + α2δ-2 = 2.44 ± 0.36. Values are mean ± s.e.m., $*P < 0.01$, $n \geq 7$, ($*P < 0.05$).

controlling exocytosis is experiencing higher levels of Ca^{2+} influx, even though overall synaptic Ca^{2+} transients are reduced. Single action-potential-driven Ca^{2+} influx remained equally sensitive to ω-conotoxin GVIA following α2δ overexpression, indicating that this condition did not lead to a significant shift in VGCC type at nerve terminals (Supplementary Fig. 7).

The finding that α2δ subunits form GPI-anchored proteins[3] implies that their ability to change VGCC-exocytosis coupling is probably conveyed through an extracellular interaction. One possible candidate for exerting such influence lies in the highly conserved von Willebrand A (VWA) domain within α2δ (ref. 22). A characteristic feature of this domain is its ability to interact with adhesion proteins via the MIDAS motif by sharing coordination of a divalent cation[23–25]. To examine the role of the MIDAS motif of α2δ we mutated three out of five conserved key metal coordinating residues within the DxSxS MIDAS motif to alanine[22] and expressed the mutant protein (α2δ-1 $MIDAS^{AAA}$) in neurons together with functional reporters. α2δ-1 $MIDAS^{AAA}$ was similar to wild-type α2δ-1 in its ability to drive $\alpha 1_A$ accumulation at synapses (Fig. 4a). However, measurements of exocytosis from α2δ-1 MIDAS-mutants showed no enhancement of Pv (Fig. 4b), normal Ca^{2+} influx (Fig. 4c) and normal sensitivity to EGTA block of exocytosis (Fig. 4d). Furthermore, unlike intact α2δ-2, α2δ-2 $MIDAS^{AAA}$ was unable to rescue the decrease in exocytosis resulting from shRNA-mediated α2δ-1 depletion (Fig. 4e, f). These data are consistent with the ability of this mutation to block enhancement of Ca^{2+} currents when expressed in heterologous systems[22] (Supplementary Fig. 8), but show that they do not prevent endogenous α2δ from functioning. Taken together, these results demonstrate that α2δ exerts its powerful control of synaptic VGCC function through at least two separate molecular mechanisms: a forward trafficking-step from cell body to presynaptic terminal that is independent of MIDAS motif integrity, and a local MIDAS-dependent interaction critical for proper VGCC function and coupling to exocytosis.

α2δ-1 and α2δ-2 are the targets of the analgesic gabapentin whose binding site lies in close proximity to the MIDAS site[26]. We found no significant impact of gabapentin application (30 min and more than 72 h) on Pv in either control or cells overexpressing α2δ-1 or α2δ-2 (results not shown), similar to previous findings in hippocampal neurons[27]. We also examined the effect of gabapentin on VGCC trafficking to nerve terminals by incubating neurons with gabapentin from the time of transfection with EGFP–$\alpha 1_A$. Analysis of the presynaptic abundance of this probe after 7 days showed that even though gabapentin seems to affect α2δ-2 trafficking in non-neuronal cells[28], it seems unable to affect VGCC trafficking or function in cultured hippocampal neurons (Supplementary Fig. 9).

Our results reveal that α2δ subunits are potent modulators of synaptic transmission. They function through at least two distinct molecular mechanisms: a trafficking step from the cell soma, and a local step at the presynaptic terminal allowing synapses to have increased exocytosis with decreased Ca^{2+} influx. We speculate that increased presynaptic abundance of VGCCs results in increased abundance of active zone VGCCs, and hence a higher density in the vicinity of release sites. This active zone accumulation depends on the VWA domain of α2δ, presumably through interactions with extracellular active-zone-specific proteins. The identity of the interaction partner(s) remains unknown at present; however, it is tempting to speculate that α2δs might recognize cues established in the correct juxtaposition of pre- and post-synaptic membranes, consistent with synaptic defects observed in *Drosophila* α2δ-3 mutants[29]. Additionally, this model requires that α2δ interact with a partner resulting in action potential shortening to limit total calcium entry. As the MIDAS motif is well conserved throughout the α2δ family, identification of α2δ interaction partners in specific neuronal circuits could provide novel targets in the development of future therapeutics, given the potency that these subunits show in controlling synapse function.

METHODS SUMMARY

Hippocampal CA3–CA1 regions were dissected from 1- to 3-day-old Sprague Dawley rats, dissociated, plated and transfected as previously described[30]. Live-cell images were acquired with an Andor iXon$^+$ (model DU-897E-BV) camera. A solid-state diode pumped 488 nm (vGpH and MgG imaging) or 532 nm (vGmOr2 imaging) laser was shuttered using acousto-optic modulation. For vGpH and GCaMP3 imaging, data were acquired at 100 Hz by integrating for 9.74 ms in frame transfer mode and restricting imaging to a sub-area of the CCD chip. Fluo5F imaging data was acquired at 1 kHz (0.974 ms integration time). To estimate one action potential ΔF of vGpH, we took the difference between the average 20 frames before and after the stimulus. The rise in vGpH fluorescence in response to a single action potential always took two frames when acquiring at 100 Hz time resolution. For display purposes the images in Fig. 1e were given a 2-pixel Gaussian average filter. Pv and single action potential calcium signals were measured with 4 mM extracellular $CaCl_2$. All stated values are mean ± s.e.m., statistical significance for groups of three or more were determined by one-way analysis of variance (ANOVA) with Tukey's HSD for post-hoc analysis. Otherwise Student's *t*-test was used for determining statistics.

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

Received 23 June 2011; accepted 13 March 2012.
Published online 13 May 2012.

1. Llinás, R., Steinberg, I. Z. & Walton, K. Presynaptic calcium currents and their relation to synaptic transmission: voltage clamp study in squid giant synapse and theoretical model for the calcium gate. *Proc. Natl Acad. Sci. USA* **73,** 2918–2922 (1976).
2. Neher, E. & Sakaba, T. Multiple roles of calcium ions in the regulation of neurotransmitter release. *Neuron* **59,** 861–872 (2008).
3. Davies, A. *et al.* The α2δ subunits of voltage-gated calcium channels form GPI-anchored proteins, a posttranslational modification essential for function. *Proc. Natl Acad. Sci. USA* **107,** 1654–1659 (2010).
4. Field, M. J. *et al.* Identification of the α2-δ-1 subunit of voltage-dependent calcium channels as a molecular target for pain mediating the analgesic actions of pregabalin. *Proc. Natl Acad. Sci. USA* **103,** 17537–17542 (2006).
5. Wang, M., Offord, J., Oxender, D. L. & Su, T. Z. Structural requirement of the calcium-channel subunit α2δ for gabapentin binding. *Biochem. J.* **342,** 313–320 (1999).
6. Neely, G. G. *et al.* A genome-wide *Drosophila* screen for heat nociception identifies α2δ3 as an evolutionarily conserved pain gene. *Cell* **143,** 628–638 (2010).
7. Gao, B. *et al.* Functional properties of a new voltage-dependent calcium channel α2δ auxiliary subunit gene (*CACNA2D2*). *J. Biol. Chem.* **275,** 12237–12242 (2000).
8. Arikkath, J. & Campbell, K. P. Auxiliary subunits: essential components of the voltage-gated calcium channel complex. *Curr. Opin. Neurobiol.* **13,** 298–307 (2003).
9. Catterall, W. A. Structure and regulation of voltage-gated Ca^{2+} channels. *Annu. Rev. Cell Dev. Biol.* **16,** 521–555 (2000).
10. Dunlap, K., Luebke, J. I. & Turner, T. J. Exocytotic Ca^{2+} channels in mammalian central neurons. *Trends Neurosci.* **18,** 89–98 (1995).
11. Cao, Y. Q. *et al.* Presynaptic Ca^{2+} channels compete for channel type-preferring slots in altered neurotransmission arising from Ca^{2+} channelopathy. *Neuron* **43,** 387–400 (2004).
12. Watschinger, K. *et al.* Functional properties and modulation of extracellular epitope-tagged $Ca_V2.1$ voltage-gated calcium channels. *Channels* **2,** 461–473 (2008).
13. Winterfield, J. R. & Swartz, K. J. A hot spot for the interaction of gating modifier toxins with voltage-dependent ion channels. *J. Gen. Physiol.* **116,** 637–644 (2000).
14. Dolphin, A. C. Calcium channel diversity: multiple roles of calcium channel subunits. *Curr. Opin. Neurobiol.* **19,** 237–244 (2009).
15. Pragnell, M. *et al.* Calcium channel β-subunit binds to a conserved motif in the I–II cytoplasmic linker of the α1-subunit. *Nature* **368,** 67–70 (1994).
16. Schneggenburger, R., Sakaba, T. & Neher, E. Vesicle pools and short-term synaptic depression: lessons from a large synapse. *Trends Neurosci.* **25,** 206–212 (2002).
17. Ariel, P. & Ryan, T. A. Optical mapping of release properties in synapses. *Front. Neural Circuits* **4,** 18 (2010).
18. Catterall, W. A. & Few, A. P. Calcium channel regulation and presynaptic plasticity. *Neuron* **59,** 882–901 (2008).
19. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* **6,** 875–881 (2009).
20. Dodge, F. A. Jr & Rahamimoff, R. Co-operative action a calcium ions in transmitter release at the neuromuscular junction. *J. Physiol. (Lond.)* **193,** 419–432 (1967).
21. Parekh, A. B. Ca^{2+} microdomains near plasma membrane Ca^{2+} channels: impact on cell function. *J. Physiol. (Lond.)* **586,** 3043–3054 (2008).
22. Cantí, C. *et al.* The metal-ion-dependent adhesion site in the Von Willebrand factor-A domain of α2δ subunits is key to trafficking voltage-gated Ca^{2+} channels. *Proc. Natl Acad. Sci. USA* **102,** 11230–11235 (2005).
23. Springer, T. A. Complement and the multifaceted functions of VWA and integrin I domains. *Structure* **14,** 1611–1616 (2006).
24. Lacy, D. B., Wigelsworth, D. J., Scobie, H. M., Young, J. A. & Collier, R. J. Crystal structure of the von Willebrand factor A domain of human capillary morphogenesis protein 2: an anthrax toxin receptor. *Proc. Natl Acad. Sci. USA* **101,** 6367–6372 (2004).
25. Whittaker, C. A. & Hynes, R. O. Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Mol. Biol. Cell* **13,** 3369–3387 (2002).
26. Davies, A. *et al.* The calcium channel α2δ-2 subunit partitions with CaV2.1 into lipid rafts in cerebellum: implications for localization and function. *J. Neurosci.* **26,** 8748–8757 (2006).
27. Brown, J. T. & Randall, A. Gabapentin fails to alter P/Q-type Ca^{2+} channel-mediated synaptic transmission in the hippocampus *in vitro*. *Synapse* **55,** 262–269 (2005).
28. Tran-Van-Minh, A. & Dolphin, A. C. The α2δ ligand gabapentin inhibits the Rab11-dependent recycling of the calcium channel subunit α2δ-2. *J. Neurosci.* **30,** 12856–12867 (2010).
29. Kurshan, P. T., Oztan, A. & Schwarz, T. L. Presynaptic α2δ-3 is required for synaptic morphogenesis independent of its Ca^{2+}-channel functions. *Nature Neurosci.* **12,** 1415–1423 (2009).
30. Kim, S. H. & Ryan, T. A. CDK5 serves as a major control point in neurotransmitter release. *Neuron* **67,** 797–809 (2010).

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

Acknowledgements We thank W. Pratt for performing mutagenesis on $\alpha_2\delta$-1 and M. D'Arco for advice on immunocytochemistry. We thank S. Kim for designing the vGmOr2 and VAMPmCh plasmids, P. Ariel for helpful discussion and analysis of exocytosis data, R. Kwan and Y. Gera for technical assistance in cell culture, G. Obermair for providing the wild-type EGFP–$\alpha1_A$ plasmid and L. Looger for providing the GCaMP3 plasmid and helpful communication.

Author Contributions M.B.H. performed opto-physiological experiments, B.L. and W.M. performed electrophysiological experiments, A.C.D., B.L. and W.M. designed electrophysiological experiments, M.B.H., A.C.D. and T.A.R. designed all other experiments, M.B.H., A.C.D. and T.A.R. wrote the manuscript.

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

METHODS

Plasmids. VAMP-mCherry was designed by replacing the pHluorin of VAMP-pHluorin[31] with mCherry. Vesicular Glutamate-mOrange2 (vGmOr2) was designed by replacing the pHluorin of vGlut1-pHluorin[32] with mOrange2[33]. EGFP–$\alpha 1_A$ was kindly provided by G. Obermair[12], E1656K mutagenesis of was outsourced to Mutagenex. GCaMP3[19] cDNA was kindly provided by L. Looger. The auxiliary subunits used were rat $\alpha_2\delta$-1 (GenBank accession number AF_286488), mouse $\alpha_2\delta$-2 (AF_247139), rat $\alpha_2\delta$-3 (NM_175595) and rat β4 (NM_001105733). The α2δ-1 and α2δ-2 MIDAS^{AAA} constructs were made by standard molecular biological techniques and verified by DNA sequencing. MIDAS mutations were generated by mutating three MIDAS motif amino acids to Ala. D^{259}VSGS in α2δ-1 and D^{300}VSGS in α2δ-2 became AVAGA. All auxiliary subunits were cloned into pcDNA3.0 vectors. For knockdown of endogenous $\alpha 1_A$ or α2δ−1 subunits, mRNA target sequences (GCATTCTCCTCTGGACTTTCG) and (ACTCAACTGGACAAGTGCCTTAGATGAAG), respectively, were cloned into an shRNA vector[30].

Immunofluorescence and quantification. To quantify $\alpha 1_A$ in presynaptic boutons, following live cell imaging, neurons were fixed with 4% paraformaldehyde, permeabilized with 0.2% Triton X-100, blocked with 5% BSA and 1% goat serum for 1 h, and subsequently incubated overnight at 4 °C with anti-GFP (Invitrogen) 1:3,000 and anti-$\alpha 1_A$ (Synaptic Systems). Alexa-488- or Alexa-546-conjugated secondary antibodies (1:1,000) were applied post primary antibody incubation. Expression level ratios of $\alpha 1_A$ were measured as follows: 2-μm diameter circular regions of interest (ROIs) were centred on transfected synaptic marker to indentify the axonal boutons of transfected neurons and then compared to an equal number of adjacent ROIs that were centred on any small punctate spots of $\alpha 1_A$ fluorescence within 2–10 μm of measured transfected boutons with local background correction. For measurements of α2δ-1, cells were fixed and permeabilized with 10% MES buffer pH 6.9 and 90% methanol at −20 °C for 5 min then blocked for 2 h at room temperature. Primary antibodies for α2δ-1 (1:100 dilution; C14882 LSBio) were incubated at 37 °C for 4 h before wash and detection with secondary antibodies.

Live cell imaging. Action potentials were evoked by passing 1-ms current pulses, yielding fields of approximately 10 V cm^{-1} via platinum-iridium electrodes. Live-cell imaging experiments were performed at 30.0 ± 0.2 °C. Cells were continuously perfused at 0.2–1.0 ml min^{-1} in a saline solution containing (in mM) 119 NaCl, 2.5 KCl, 4 $CaCl_2$, 25 HEPES, buffered to pH 7.4, 30 glucose, 10 μM 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX), and 50 μM D,L-2-amino-5-phosphonovaleric acid (AP5). NH_4Cl applications were done with 50 mM NH_4Cl in substitution of 50 mM of NaCl (buffered to pH 7.4). All chemicals were obtained from Sigma except for Ca^{2+} channel toxins (Alomone Labs) and Ca^{2+}dyes (Invitrogen). During experiments, cells were allowed to rest for ∼30 s between one action potential trials and at least 5 min between 100-Hz action potential bursts. All RRP measurements were the average of eight trials of 20 action potential at 100 Hz, details of measurements previously described[17]. 1-kHz imaging used 0.972 ms in-frame transfer mode with an imaging field of 5 pixel width (2 μm) and 512 pixel height. Toxins to block P/Q- and N-type channels were applied for ∼2 min before stimulation with washout at the following concentrations: agatoxin IVA (400 nM) and conotoxin GVIA (400 nM). Fluo5F measurements were obtained by diluting a DMSO stock 1 μg μl^{-1} 1:150 and loading or 10 min at 30 °C before washing for 30 min before imaging. For experiments with vGlut-pHluorin involving EGTA-AM, 200 μM was loaded for 90 s, followed by a 10 min wash before beginning experiments.

GCaMP3 measurements. We found that GCaMP3 had a highly nonlinear response to changes in [Ca^{2+}] as previously described *in vitro*[19]. We compared the change in fluorescence relative to the response of single action potential in cells transfected with GCaMP compared to similar cells loaded with magnesium green dye (MgG AM-ester dye)[17]. Cytosolic GCaMP3 used to measure presynaptic bouton intracellular Ca^{2+} in hippocampal neurons stimulated by field-potential-generated action potentials. GCaMP3 peak fluorescence (ΔF) for each stimulation was found by averaging the 5 highest points post stimulation and subtracting the average of 10 points before stimulation. All changes in fluorescence were normalized to GCaMP intensity to Ca^{2+} saturation (F_{MAX}). F_{MAX} was calculated by applying a solution of Tyrode's buffer (pH 6.9) at the end of experiments containing ionomycin (1 mM). Ionomycin exposure caused a 6.1- to 6.6-fold increase in GCaMP fluorescence in good agreement with results found previously under these pH conditions (personal communication, L. Looger). These values were then compared to MgG values under similar conditions. We found that these changes in fluorescence could be well fit with a Hill equation: $V\text{max} = 1$, $k = 9.077$, $n = 2.46$; adjusted $R^2 = 0.997$. The fit to the Hill equation was in good agreement with published expectations[19]. To linearize GCaMP 3 signals we inverted the Hill equation to obtain an expression of the signal relative to that obtained with MgG following one action potential stimulation using the following equation: linearized GCaMP = $(((\Delta F/F)/F_{MAX}) \times k^n)/(1 - ((\Delta F/F)/F_{MAX}))^{(1/n)}$.

Electrophysiology. Calcium channel expression in tsA-201 cells was investigated by whole-cell patch-clamp recording. The patch pipette solution contained in mM: 140 Cs-aspartate, 5 EGTA, 2 $MgCl_2$, 0.1 $CaCl_2$, 2 ATP, 20 HEPES pH 7.2, 310 mOsm with sucrose. The external solution for recording Ba^{2+} currents contained in mM: tetraethylammonium (TEA) 160 Br, 3 KCl, 1.0 $NaHCO_3$, 1.0 $MgCl_2$, 10 HEPES, 4 glucose, 1 $BaCl_2$ pH 7.4, 320 mOsm with sucrose. Measurements and analysis were performed as previously described[28].

Action potential recordings were performed in a bath solution containing (mM): 145 NaCl, 5 KCl, 2 $CaCl_2$, 1 $MgSO_4$, 10 HEPES, 10 glucose, pH 7.4. The internal solution contained (mM): 130 KCl, 10 EGTA, 10 HEPES, 8 NaCl, 4 Mg-ATP, 1 $MgCl_2$, 1 $CaCl_2$, 0.4 Na_2-GTP, pH 7.25 adjusted with 1 M KOH, 318 mOsm. Data were analysed with Clampfit 9 (Molecular Devices), recorded traces were post-processed with 1 kHz 8-pole Bessel digital filter, the action potential initiated at rheobase was used for measurement of the peak of action potential overshoot, the peak of after-hyperpolarization (AHP) and the duration of action potential. Measured parameters between two groups were compared using Student's *t*-test. Dorsal root ganglion neurons (DRGs) were prepared from P10 Sprague Dawley rats and transfected as previously described[34].

31. Miesenböck, G., De Angelis, D. A. & Rothman, J. E. Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins. *Nature* **394,** 192–195 (1998).
32. Voglmaier, S. M. *et al.* Distinct endocytic pathways control the rate and extent of synaptic vesicle protein recycling. *Neuron* **51,** 71–84 (2006).
33. Shaner, N. C. *et al.* Improving the photostability of bright monomeric orange and red fluorescent proteins. *Nature Methods* **5,** 545–551 (2008).
34. Viard, P. *et al.* PI3K promotes voltage-dependent calcium channel trafficking to the plasma membrane. *Nature Neurosci.* **7,** 939–946 (2004).

doi:10.1038/nature11055

An oxygen-regulated switch in the protein synthesis machinery

James Uniacke[1], Chet E. Holterman[1], Gabriel Lachance[1], Aleksandra Franovic[1], Mathieu D. Jacob[1], Marc R. Fabian[2], Josianne Payette[1], Martin Holcik[3], Arnim Pause[2] & Stephen Lee[1]

Protein synthesis involves the translation of ribonucleic acid information into proteins, the building blocks of life. The initial step of protein synthesis is the binding of the eukaryotic translation initiation factor 4E (eIF4E) to the 7-methylguanosine (m^7-GpppG) 5′ cap of messenger RNAs[1,2]. Low oxygen tension (hypoxia) represses cap-mediated translation by sequestering eIF4E through mammalian target of rapamycin (mTOR)-dependent mechanisms[3–6]. Although the internal ribosome entry site is an alternative translation initiation mechanism, this pathway alone cannot account for the translational capacity of hypoxic cells[7,8]. This raises a fundamental question in biology as to how proteins are synthesized in periods of oxygen scarcity and eIF4E inhibition[9]. Here we describe an oxygen-regulated translation initiation complex that mediates selective cap-dependent protein synthesis. We show that hypoxia stimulates the formation of a complex that includes the oxygen-regulated hypoxia-inducible factor 2α (HIF-2α), the RNA-binding protein RBM4 and the cap-binding eIF4E2, an eIF4E homologue. Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP)[10] analysis identified an RNA hypoxia response element (rHRE) that recruits this complex to a wide array of mRNAs, including that encoding the epidermal growth factor receptor. Once assembled at the rHRE, the HIF-2α–RBM4–eIF4E2 complex captures the 5′ cap and targets mRNAs to polysomes for active translation, thereby evading hypoxia-induced repression of protein synthesis. These findings demonstrate that cells have evolved a program by which oxygen tension switches the basic translation initiation machinery.

Here we describe an oxygen-regulated mechanism that mediates selective cap-dependent translation during hypoxia and eIF4E inactivation (Supplementary Fig. 1). We began our investigation into this alternative mechanism of protein synthesis by examining the epidermal growth factor receptor (EGFR), a receptor tyrosine kinase that has a critical function in cell proliferation, tissue development and cancer[11]. Hypoxia activates the translation of *EGFR* mRNA through HIF-2α (ref. 12), a member of the HIF family of transcription factors involved in maintaining cellular oxygen homeostasis[13–17]. We speculated that HIF-2α might orchestrate a gene program that enabled the translation of *EGFR* mRNA during periods of eIF4E inactivation. We treated cells with transcription inhibitors to preclude activation of HIF-2α target genes during hypoxia, and found that hypoxia caused the accumulation of EGFR protein in a HIF-2α-dependent manner even in transcription-incompetent glioma or primary cultures of renal epithelial cells (Fig. 1a and Supplementary Fig. 2). The *EGFR* mRNA was captured by polysomes and translated *de novo* in hypoxic cells treated with actinomycin D (Fig. 1b, and Supplementary Figs 3 and 4). Silencing of HIF-2α abolished the association of *EGFR* mRNA with polysomes and prevented its *de novo* synthesis (Fig. 1b, and Supplementary Figs 3 and 4). In contrast, silencing of HIF-1α, a paralogue of HIF-2α, did not prevent the hypoxic induction of *EGFR* translation and capture by polysomes

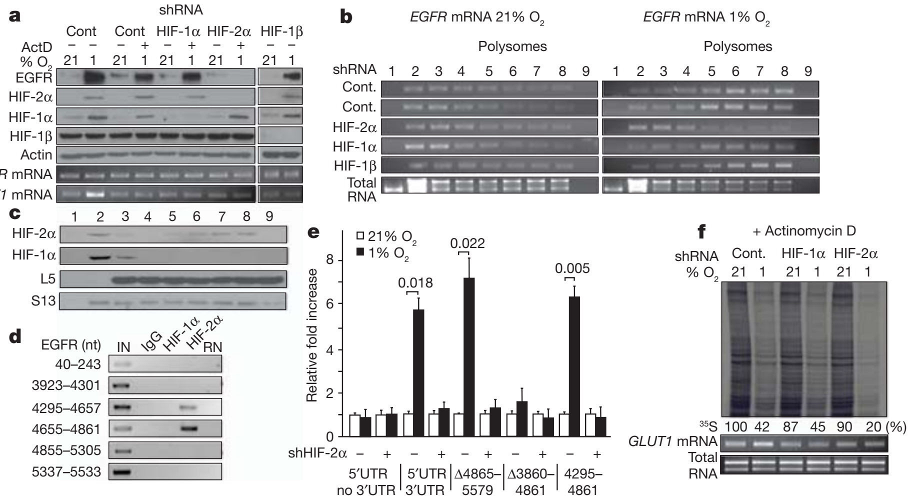


Figure 1 | HIF-2α activates *EGFR* mRNA translation by interacting with its 3′ UTR. a, **b**, Western blot (**a**) and polysomal distribution (**b**) of EGFR protein and mRNA in HIF-2α, HIF-1α or HIF-1β knockdown cells in the presence of actinomycin D (ActD). *GLUT1* (also known as *SLC2A1*) was used as a control. Cont, control; shRNA, short hairpin RNA. See also Supplementary Fig. 3b. **c**, Polysomal distribution of HIF-2α and HIF-1α in hypoxia. **d**, RNA immunoprecipitation of HIF-1α and HIF-2α. IN, input; RN, RNase-treated; nt, nucleotides. **e**, Dual luciferase assays in cells transfected with *EGFR* 3′ UTR reporter constructs. Significances of fold changes (Student's *t*-test) are shown. Results are means and s.e.m. ($n = 3$). **f**, Global translation rates of transcription-incompetent cells expressing shRNA targeting HIF-2α or HIF-1α. Experiments performed in U87MG glioblastoma.

[1]Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5. [2]Department of Biochemistry, Goodman Cancer Research Center, McGill University, Montreal, Quebec, Canada H3G 1Y6. [3]Apoptosis Research Centre, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada K1H 8L1.

(Fig. 1a, b, and Supplementary Figs 3 and 4). In addition, ablation of HIF-1β, a protein required for HIF transcriptional activity, had no effect on hypoxia-inducible translation of the *EGFR* mRNA (Fig. 1a, b, and Supplementary Fig. 3). HIF-2α, but not HIF-1α, was observed in polysome fractions of hypoxic cells, suggesting its direct involvement in the translational machinery (Fig. 1c and Supplementary Fig. 5). RNA immunoprecipitation revealed that HIF-2α associates with the 3′ untranslated region (UTR) of *EGFR* mRNA between nucleotides 4295 and 4861 (Fig. 1d, Supplementary Fig. 6 and Supplementary Table 1). This segment was both necessary and sufficient to enhance the translation of a luciferase reporter during hypoxia in a HIF-2α-dependent manner even in the absence of transcription (Fig. 1e and Supplementary Fig. 7). Silencing HIF-2α, but not HIF-1α, considerably decreased the rate of global hypoxic translation, highlighting its participation in hypoxic protein synthesis beyond *EGFR* translation (Fig. 1f and Supplementary Fig. 8).

HIF-2α does not contain a classical RNA recognition motif; we therefore searched for potential interacting partners that could bind the *EGFR* 3′ UTR. Immunoprecipitation revealed a band of 40 kDa specifically associated with HIF-2α that was identified as RNA-binding motif protein 4 (RBM4; Supplementary Fig. 9), a protein involved in translation control[18,19]. Co-immunoprecipitation revealed that endogenous RBM4 interacted with the amino-terminal region of HIF-2α but not with HIF-1α during hypoxia (Fig. 2a and Supplementary Fig. 10). Furthermore, RBM4 assembled with the *EGFR* 3′ UTR *in vivo* and *in vitro* independently of oxygen tension (Fig. 2b and Supplementary Fig. 11). Silencing experiments revealed that RBM4 is essential for HIF-2α recruitment to the *EGFR* 3′ UTR (Fig. 2b and Supplementary Fig. 11a), the hypoxic induction of EGFR protein (Supplementary Fig. 12) and the ability of the 4295–4861 *EGFR* 3′ UTR segment to induce hypoxia-dependent translation of a reporter construct (Fig. 2c). In addition, depletion of RBM4 caused decreased hypoxic cellular translation to levels similar to those observed in HIF-2α-incompetent cells (Fig. 2d and Supplementary Fig. 13). Consistent with the RNA immunoprecipitation, multiple PAR-CLIP sequenced reads for HIF-2α–RBM4 and RBM4 concentrated at the same site within the *EGFR* 3′ UTR 4295–4861 fragment that confers hypoxic translation on a reporter protein but not elsewhere in the transcript (Supplementary Figs 14 and 15). The crosslink sites were near a CGG trinucleotide, a feature of RBM4-binding motifs[20,21]. Mutation of this CGG motif, but not that of another CGG sequence, was sufficient to

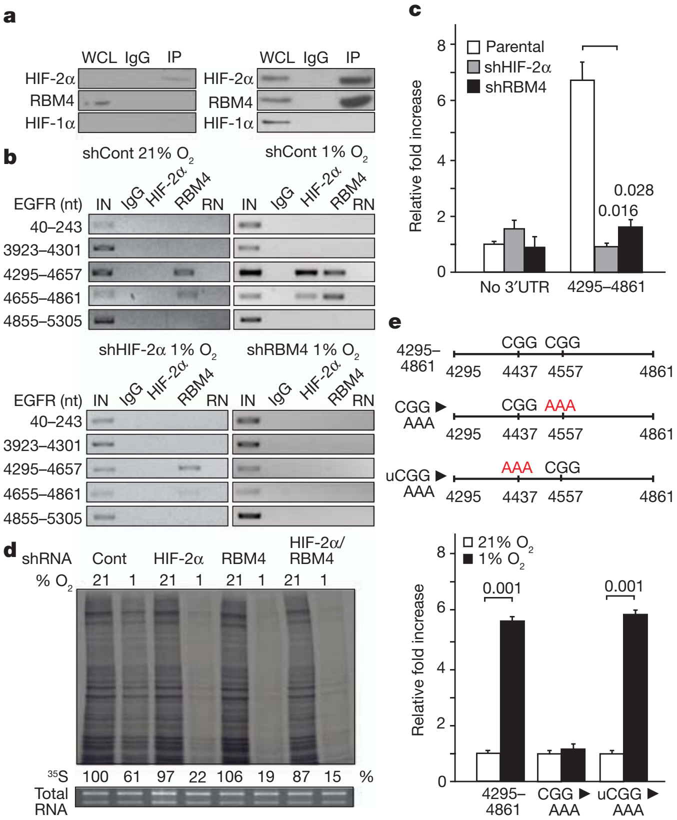


Figure 2 | RBM4 recruits HIF-2α to the 3′ UTR for hypoxic translation. **a**, Co-immunoprecipitation of HIF-2α in normoxia (21% O_2, left) and hypoxia (1% O_2, right). WCL, whole cell lysate. **b**, RNA immunoprecipitation of HIF-2α and RBM4 in HIF-2α or RBM4 knockdown cells. IN, input; nt, nucleotides; RN, RNase-treated. **c**, Effect of silencing HIF-2α or RBM4 on the hypoxic expression of a luciferase reporter fused to the 4295–4861 segment of the *EGFR* 3′ UTR. **d**, Global translation rates in normoxic or hypoxic HIF-2α and/or RBM4 knockdown cells. **e**, Expression of a luciferase reporter containing a CGG→AAA mutation near RBM4 crosslinking sites, or in an unrelated upstream region (uCGG). Results in **c** and **e** are means and s.e.m. ($n = 3$). Significances of fold changes (Student's *t*-test) are shown. Experiments were performed in U87MG glioblastoma.

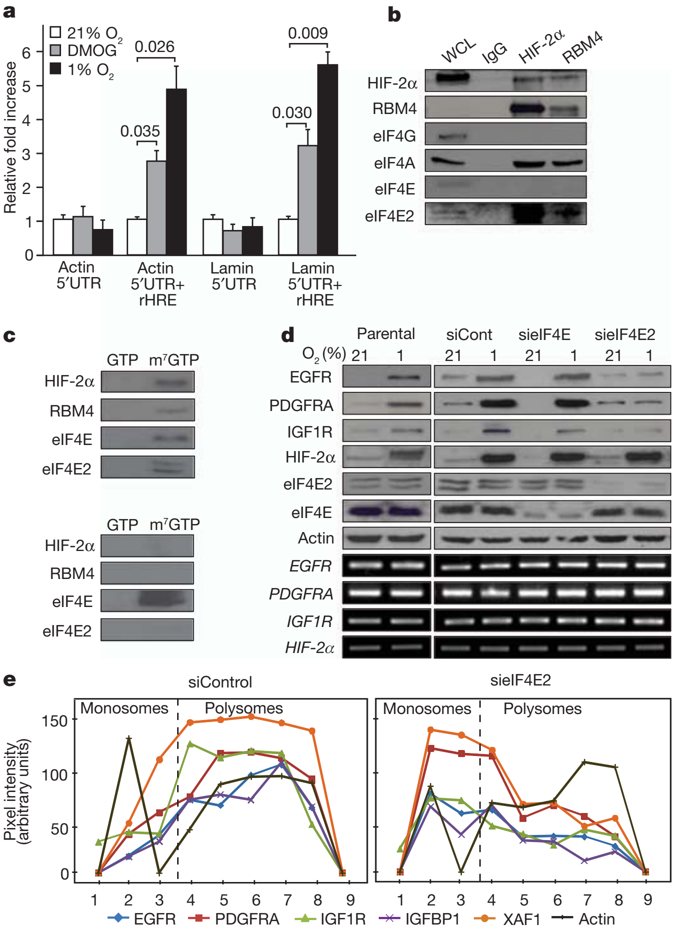


Figure 3 | HIF-2α–RBM4 recruits the m^7-GTP cap by means of an interaction with eIF4E2. **a**, Dual luciferase assays in cells transfected with reporter constructs containing the 5′ UTR of actin or lamin a/c with or without a 3′ rHRE. Significances of fold changes (Student's *t*-test) are shown. Results are means and s.e.m. ($n = 3$). DMOG, dimethyloxalylglycine. **b**, Co-immunoprecipitation of HIF-2α and RBM4 in hypoxia (1% O_2). **c**, Capture assays using m^7-GTP beads in hypoxic cell lysates depleted in eIF4E (top) or eIF4E2 (bottom). GTP, proteins dislodged from the beads by GTP; m^{7}GTP, proteins bound to m^7-GTP beads after GTP wash. **d**, Western blot of total EGFR, PDGFRA, IGF1R, HIF-2α, eIF4E and eIF4E2 levels in eIF4E or eIF4E2 knockdown cells. The lower (darker) panels display mRNA levels. **e**, Polysomal distribution of mRNA coding for HIF-2α–RBM4 targets in hypoxic eIF4E2 knockdown cells. XAF1, XIAP-associated factor 1. Experiments were performed in U87MG glioblastoma.

disrupt the secondary structure and to abolish hypoxia-inducible translation of a reporter construct (Fig. 2e, and Supplementary Figs 16 and 17). In addition, shorter segments of the *EGFR* 3′ UTR that altered the secondary structure were unresponsive to hypoxia in luciferase assays (Supplementary Figs 18 and 19). Consistent with the ^{35}S-labelling experiments, a wide array of mRNAs interacted with the HIF-2α–RBM4 complex (Supplementary Fig. 20a, Supplementary Information and Supplementary Table 2). Similarly to *EGFR* mRNA, multiple sequenced reads for HIF-2α–RBM4 and RBM4 concentrated at the same site near CG(G) nucleotides in most candidates (Supplementary Fig. 20b, c). Several PAR-CLIP candidates were validated for HIF-2α-dependent hypoxic induction (Supplementary Figs 20c and 21). We suggest that RBM4 binds to specific regions in the 3′ UTR of mRNAs to recruit HIF-2α and induce hypoxic translation. These RNA sequences are referred to as rHREs.

A key characteristic of the *EGFR* rHRE is that it confers hypoxia-inducible translation to several unrelated 5′ UTRs that are otherwise unable to initiate translation during hypoxia (Fig. 3a). We therefore suspected that the rHRE might exploit the cap, because it is a common feature used by the 5′ UTRs of mRNAs to initiate protein synthesis. The RBM4–HIF-2α complex of hypoxic cells was captured by m^7-GTP beads (Supplementary Fig. 22). Immunoprecipitation experiments showed that HIF-2α–RBM4 specifically assembles with the cap-binding protein eIF4E2, a homologue of eIF4E (Fig. 3b and Supplementary Fig. 23a, b). We therefore reasoned that eIF4E2 is recruited by HIF-2α–RBM4 to activate selective cap-dependent translation of rHRE-containing mRNAs during hypoxia and inhibition of eIF4E by 4E-binding protein (4EBP)[6,22,23]. Immunoprecipitation revealed that 4EBP has more affinity for eIF4E than for eIF4E2, which is consistent with previously published results (Supplementary Fig. 23c)[24,25]. Silencing of eIF4E2, but not that of eIF4E, prevented the binding of HIF-2α–RBM4 from hypoxic cells to m^7-GTP beads (Fig. 3c). In addition, ablation of eIF4E2 prevented the hypoxic induction of multiple proteins identified by PAR-CLIP, including EGFR, whereas silencing of eIF4E had no discernible effect (Fig. 3d and Supplementary Figs 20b and 24). Ablation of eIF4E2 prevented the capture of rHRE-containing mRNAs by polysomes (Fig. 3e, and Supplementary Figs 25 and 26). The HIF-2α–RBM4–eIF4E2 complex also recruited the RNA helicase eIF4A, a fundamental component of translation initiation[26] (Fig. 3b and Supplementary Fig. 23b). Taken together, these results demonstrate that eIF4E2 is a member of a hypoxic translation initiation complex that mediates selective cap-dependent protein synthesis independently of eIF4E.

Figure 3d shows that eIF4E might be involved in the translation of mRNAs coding for EGFR, platelet-derived growth factor receptor α (PDGFRA) and insulin-like growth factor 1 receptor (IGF1R) during normoxia. This raises the intriguing possibility that the cap-dependent translational machinery switches from eIF4E to eIF4E2 as a function of oxygen tension. This oxygen-dependent switch was clearly observed between eIF4E and eIF4E2 in polysomes: eIF4E participation in the translational machinery was essentially limited to normoxia, and eIF4E2 participation to hypoxia (Fig. 4a, and Supplementary Figs 27 and 28). Ablation of HIF-2α abolished the hypoxic shift of eIF4E2 to polysomes, attesting to the role of HIF-2α as the oxygen-regulated subunit of the eIF4E2–RBM4–HIF-2α complex (Fig. 4a, and Supplementary Fig. 27c). In contrast, treatment with rapamycin, an inhibitor of mTOR and eIF4E, prevented the accumulation of eIF4E in normoxic polysomes but had no effect on eIF4E2 (Supplementary Fig. 27e). Treatment with rapamycin, or silencing of eIF4E, significantly decreased the expression of rHRE-containing reporter mRNAs in normoxia but had no effect in hypoxia (Fig. 4b and Supplementary Fig. 29). Silencing any of the participants of the HIF-2α–RBM4–eIF4E2 cap-binding complex prevented the translation of rHRE-containing mRNAs in hypoxia but not in normoxia. Silencing of eIF4E decreased the rate of cellular protein synthesis in normoxia but had no effect in hypoxia (Fig. 4c and Supplementary Fig. 30). In stark contrast, depletion of eIF4E2 considerably limited the global rate of hypoxic translation without affecting protein synthesis in cells maintained in normoxia (Fig. 4c and Supplementary Fig. 30). These results demonstrate that oxygen tension regulates the cap-dependent protein synthesis machinery by switching from eIF4E- to eIF4E2-dependent translation in a HIF-2α-dependent manner (Supplementary Fig. 1).

Here we have identified a selective cap-dependent translation initiation mechanism that operates independently of eIF4E and that targets mRNAs for protein synthesis during hypoxia. The results suggest that the HIF-2α–RBM4–eIF4E2 complex is extensively involved in coordinating the translation response to low oxygen availability and is therefore essential in cellular oxygen homeostasis. This complex probably recruits functional homologues of the canonical eIF4E-dependent pathway, as well as distinct components, to initiate hypoxic protein synthesis. This process is regulated by the oxygen-sensing machinery first identified as the main regulator of the transcriptional response to hypoxia[13–16]. A human population that recently migrated to the Tibetan highlands contains a point mutation in the gene encoding HIF-2α (*EPAS1*), further emphasizing the evolutionary role of HIF-2α in the adaptation to high altitude and low oxygen tension[27]. The target mRNAs code for proteins such as EGFR, PDGFRA and IGF1R that are implicated in the adaptive response to hypoxia as well as a wide variety of biological processes including development and cancer. The role of these receptor tyrosine kinases in human malignancy

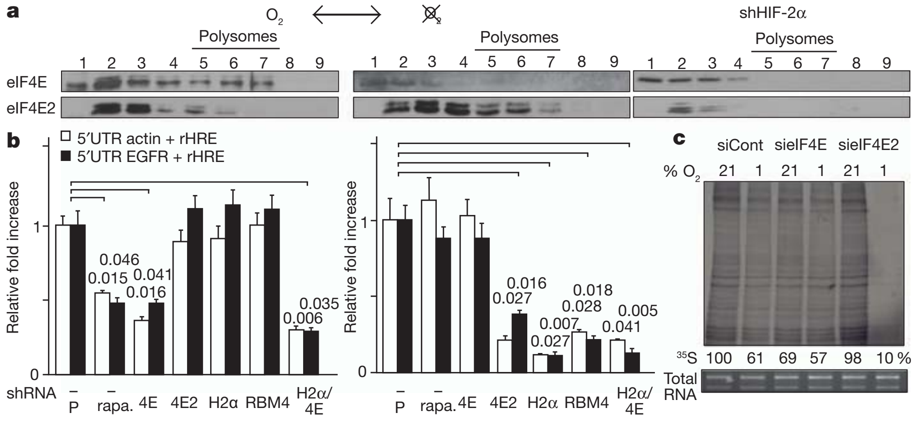


Figure 4 | An oxygen-regulated switch from eIF4E- to eIF4E2-dependent protein synthesis. **a**, eIF4E and eIF4E2 polysome association in normoxia and hypoxia. **b**, Dual luciferase assays in normoxic (left) and hypoxic (right) cells transfected with constructs containing actin or *EGFR* 5′ UTRs and *EGFR* rHREs. Assays were performed on cells treated with rapamycin (rapa.), an inhibitor of mTOR and eIF4E (4E), and in knockdown cells of eIF4E, eIF4E2 (4E2), HIF-2α (H2α) or RBM4 and in a eIF4E–HIF-2α double knockdown. Significances of fold changes (Student's *t*-test) are shown. P, parental. Results are means and s.e.m. ($n = 3$). **c**, Global translation rates in normoxic or hypoxic eIF4E or eIF4E2 knockdown cells. Experiments performed in U87MG glioblastoma.

is particularly well documented and they are at the centre of targeted therapy[11,28]. EGFR is often overproduced by tumours that harbour a wild-type *EGFR* gene, suggesting that cancer cells hijack the eIF4E2 pathway for their proliferative advantage[29,30]. The results shown here provide the foundation for further investigation of the adaptive properties of the basic protein synthesis machinery in response to environmental conditions.

METHODS SUMMARY

With the exception of the human renal proximal tubular epithelial cells (a gift from C. Kennedy), cell lines were obtained from the American Type Culture Collection and propagated as suggested and maintained in epithelial cell medium (ScienCell). Hypoxia was induced by incubation for 24 h at 37 °C in a 1% O_2, 5% CO_2 and N_2-balanced atmosphere unless otherwise indicated. Polysome analysis was performed as described previously[12], with the addition that proteins were isolated by precipitation with trichloroacetic acid and analysed by western blotting. Generation of luciferase constructs, RNA isolation, polymerase chain reaction and polysome analysis are outlined in Supplementary Methods. All short interfering RNA, short hairpin RNA and adenoviral infections were performed with commercially available products and are further described in Supplementary Methods. PAR-CLIP analysis was performed as described previously[10] and is further outlined in Supplementary Methods. Statistical analyses were performed with paired two-tailed Student's *t*-tests.

The protocols for the following are described in detail in Supplementary Methods: western blotting, radioisotope labelling, RNA immunoprecipitation and cap-binding assays.

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

Received 4 July 2011; accepted 15 March 2012.
Published online 6 May 2012.

1. Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nature Rev. Mol. Cell Biol.* **5,** 827–835 (2004).
2. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136,** 731–745 (2009).
3. Braunstein, S. *et al.* A hypoxia-controlled cap-dependent to cap-independent translation switch in breast cancer. *Mol. Cell* **28,** 501–512 (2007).
4. Brugarolas, J. *et al.* Regulation of mTOR function in response to hypoxia by REDD1 and the TSC1/TSC2 tumor suppressor complex. *Genes Dev.* **18,** 2893–2904 (2004).
5. Koritzinsky, M. *et al.* Gene expression during acute and prolonged hypoxia is regulated by distinct mechanisms of translational control. *EMBO J.* **25,** 1114–1125 (2006).
6. Liu, L. *et al.* Hypoxia-induced energy stress regulates mRNA translation and cell growth. *Mol. Cell* **21,** 521–531 (2006).
7. Holcik, M. & Sonenberg, N. Translational control in stress and apoptosis. *Nature Rev. Mol. Cell Biol.* **6,** 318–327 (2005).
8. Young, R. M. *et al.* Hypoxia-mediated selective mRNA translation by an internal ribosome entry site-independent mechanism. *J. Biol. Chem.* **283,** 16309–16319 (2008).
9. Merrick, W. C. Eukaryotic protein synthesis: still a mystery. *J. Biol. Chem.* **285,** 21197–21201 (2010).
10. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–141 (2010).
11. Yarden, Y. & Sliwkowski, M. X. Untangling the ErbB signalling network. *Nature Rev. Mol. Cell Biol.* **2,** 127–137 (2001).
12. Franovic, A. *et al.* Translational up-regulation of the EGFR by tumor hypoxia provides a nonmutational explanation for its overexpression in human cancer. *Proc. Natl Acad. Sci. USA* **104,** 13092–13097 (2007).
13. Ivan, M. *et al.* HIFα targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. *Science* **292,** 464–468 (2001).
14. Jaakkola, P. *et al.* Targeting of HIF-α to the von Hippel–Lindau ubiquitylation complex by O_2-regulated prolyl hydroxylation. *Science* **292,** 468–472 (2001).
15. Kaelin, W. G. Jr & Ratcliffe, P. J. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Mol. Cell* **30,** 393–402 (2008).
16. Semenza, G. L. Regulation of mammalian O_2 homeostasis by hypoxia-inducible factor 1. *Annu. Rev. Cell Dev. Biol.* **15,** 551–578 (1999).
17. Wiesener, M. S. *et al.* Widespread hypoxia-inducible expression of HIF-2α in distinct cell populations of different organs. *FASEB J.* **17,** 271–273 (2003).
18. Lin, J. C., Hsu, M. & Tarn, W. Y. Cell stress modulates the function of splicing regulatory protein RBM4 in translation control. *Proc. Natl Acad. Sci. USA* **104,** 2235–2240 (2007).
19. Lin, J. C. & Tarn, W. Y. RNA-binding motif protein 4 translocates to cytoplasmic granules and suppresses translation via argonaute2 during muscle cell differentiation. *J. Biol. Chem.* **284,** 34658–34665 (2009).
20. Kazan, H., Ray, D., Chan, E. T., Hughes, T. R. & Morris, Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLOS Comput. Biol.* **6,** e1000832 (2010).
21. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnol.* **27,** 667–670 (2009).
22. Pause, A. *et al.* Insulin-dependent stimulation of protein synthesis by phosphorylation of a regulator of 5′-cap function. *Nature* **371,** 762–767 (1994).
23. Richter, J. D. & Sonenberg, N. Regulation of cap-dependent translation by eIF4E inhibitory proteins. *Nature* **433,** 477–480 (2005).
24. Rom, E. *et al.* Cloning and characterization of 4EHP, a novel mammalian eIF4E-related cap-binding protein. *J. Biol. Chem.* **273,** 13104–13109 (1998).
25. Tee, A. R., Tee, J. A. & Blenis, J. Characterizing the interaction of the mammalian eIF4E-related protein 4EHP with 4E-BP1. *FEBS Lett.* **564,** 58–62 (2004).
26. Parsyan, A. *et al.* mRNA helicases: the tacticians of translational control. *Nature Rev. Mol. Cell Biol.* **12,** 235–245 (2011).
27. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
28. Kuehn, B. M. Genomics illuminates a deadly brain cancer. *J. Am. Med. Assoc.* **303,** 925–927 (2010).
29. Franovic, A., Holterman, C. E., Payette, J. & Lee, S. Human cancers converge at the HIF-2α oncogenic axis. *Proc. Natl Acad. Sci. USA* **106,** 21306–21311 (2009).
30. Giatromanolaki, A. *et al.* Expression of hypoxia-inducible carbonic anhydrase-9 relates to angiogenic pathways and independently to poor outcome in non-small cell lung cancer. *Cancer Res.* **61,** 7992–7998 (2001).

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

Acknowledgements We thank J. Côté, W. Kaelin, C. Kennedy and T. Tuschl for reagents and technical advice. This work was funded by the Canadian Institutes of Health Research (S.L. and M.H.). J.U. is a Research Fellow of the Terry Fox Foundation (Canadian Cancer Society Award no. 700014). A.F. was supported by a Terry Fox Foundation Studentship from the Canadian Cancer Society.

Author Contributions J.U. performed most experiments and made most of the plasmid constructs, with assistance from C.E.H. (who identified RBM4 interaction with HIF-2α, participated in PAR-CLIP experiments, and made some luciferase contructs), G.L. (who performed experiments with human renal proximal tubular epithelial cells), A.F. (who performed actinomycin D experiments on total hypoxic EGFR levels, created stable shHIF-2α and shHIF-1α cell lines and some plasmid constructs), M.D.J. (who created luciferase constructs for CGG mutagenesis and rHRE mapping and created HIF-2α truncation mutants), M.R.F. (who performed eIF4E2 co-IP assays) and J.P. (who created some luciferase constructs). J.U., C.E.H., G.L., A.F., M.R.F., M.H., A.P. and S.L. conceived the experiments and analysed the data. J.U. and S.L. wrote the paper.

Author Information Illumina sequencing data are deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE36247. 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.L. (slee@uottawa.ca).

METHODS

Cell culture and reagents. Human renal proximal tubular epithelial cells were maintained in epithelial cell medium (ScienCell). All other cell lines were obtained from the American Type Culture Collection and propagated as suggested. Cells were incubated at 37 °C in a 5% CO_2 environment. Hypoxia was induced by incubation for 24 h at 37 °C in a 1% O_2, 5% CO_2 and N_2-balanced atmosphere unless otherwise indicated. Heat shock was induced by incubating cells at 42 °C for 30 min. Dimethyloxalylglycine (DMOG; Cayman Chemical) was used at a concentration of 60 μg ml^{-1} to stabilize HIF-2α. Transcription inhibitors actinomycin D (EMD Biosciences), 5,6-dichloro-1-β-D-ribobenzimidazole (EMD Biosciences) and α-amanitin (Sigma) were added 30 min before hypoxic exposure and used at concentrations of 10 μg ml^{-1}, 78 μM and 10 μg ml^{-1}, respectively. Cells were treated with 10 nM rapamycin (Sigma) for 1 h before hypoxia.

Western blot analysis. Western blot analysis was performed using standard techniques. Monoclonal antibodies were used to detect EGFR (Ab-12; LabVision), green fluorescent protein (GFP; Roche), HIF-1α (Novus) and HIF-1β (Novus). Polyclonal antibodies were used to detect HIF-2α (Novus), actin (Sigma), L26 (Abcam), L5 (Abcam), S13 (Abcam), PDGFRA (Assay Biotech), IGF1R (Cell Signaling), RBM4 (ProteinTech Group), 4EBP1 (Cell Signaling), 4EBP1-P (Cell Signaling), AKT (Cell Signaling), AKT-P (Cell Signaling), EGFR-P (Cell Signaling), S6-P (Cell Signaling), eIF4E (Santa Cruz), eIF4E2 (Genetex) and eIF4G1 (Novus). Primary antibody against eIF4A1 was created in the laboratory of N. Sonenberg. Secondary antibodies were horseradish peroxidase-conjugated anti-mouse (Amersham Biosciences) or anti-rabbit (Jackson ImmunoResearch). Bands were detected by enhanced chemiluminescence (Pierce). Whole cell lysate (WCL) is defined as 5% of the input used for immunoprecipitations.

Protein synthesis by [^{35}S]Met incorporation. Cells were grown for 48 h in 10-cm plates. Serum-free conditions supplemented with 1% insulin-transferrin-selenium (Invitrogen) were used when cells were incubated under hypoxia to detect stronger *de novo* EGFR accumulation. Actinomycin D was added for 30 min to fresh medium at a concentration of 10 μg ml^{-1} before the addition of DMOG for the indicated times. *GLUT* mRNA was used as a control for actinomycin D activity. At 1 h before the end point, the medium was changed to glutamine-free, methionine-free and cysteine-free DMEM and labelled 30 min later with [^{35}S]Met (33 μCi ml^{-1}) for 30 min. Cells were lysed for 30 min in 1 ml of modified RIPA (50 mM Tris-HCl pH 7.4, 1% Igepal, 0.25% sodium deoxycholate, 150 mM NaCl, 1 mM EDTA, 1 mM phenylmethylsulphonyl fluoride, 1 mM Na_3VO_4, 1 mM NaF, and 1 μg ml^{-1} aprotinin, leupeptin and pepstatin) at 4 °C. Immunoprecipitation of EGFR was performed with 50 μl of agarose-conjugated anti-EGFR antibody (Santa Cruz). Samples were run on a 6% SDS–PAGE gel, dried for 90 min at 80 °C and exposed overnight to X-ray film at −80 °C. For total cellular protein synthesis rates, samples were loaded on a per-cell basis (500,000 cells). Cell viability assays were performed by incubating cells for 5 min in 3 μg ml^{-1} propidium iodide (Sigma; to stain dead cells), 3 μg ml^{-1} Hoechst (Invitrogen; to stain nuclei) and 3 μM fluorescein diacetate (Sigma; to stain live cells). The percentages of live cells over the whole population of cells were plotted. Pixel intensity was measured by densitometry with Adobe Photoshop CS5.1.

RNA immunoprecipitation. Formaldehyde (1%) was added to cells for 30 min at 21 °C. Glycine (200 mM) was added for 5 min to stop the reaction, followed by two washes with cold PBS. Cells were lysed in 1 ml of modified RIPA. RNase inhibitor (40 U ml^{-1}; Ambion) was added to modified RIPA just before use. Samples were sonicated at 50% amplitude for two cycles of 30 s (2 s on, 2 s off) with a 1-min pause between cycles. DNase treatment (12 μl of 20 mg ml^{-1} DNase I, 25 mM $MgCl_2$, 5 mM $CaCl_2$) was performed for 30 min at 37 °C. The reaction was stopped by adding 20 mM EDTA. For RNase-treated control samples, 5 μl of a 10 mg ml^{-1} RNase A solution was added for 30 min at 37 °C (RNase A; Fermentas). Samples were pre-cleared by using 10 μl of Dynabeads (Invitrogen) for 15 min at 4 °C. Beads were removed by using a magnetic stand (Promega). Immunoprecipitation was performed at 2 μg ml^{-1} primary antibody overnight at 4 °C. Samples were centrifuged for 15 min at 12,000*g* and 4 °C. The supernatant was incubated with 20 μl of Dynabeads equilibrated for 1 h in 2% BSA at 4 °C. Beads were recovered and washed five times with modified RIPA and eluted with 20 μl of 0.1 M glycine (pH 3.0). Bound proteins were removed by adding 200 mM NaCl and 20 μg of Proteinase K to the supernatant and incubating for 1 h at 42 °C. Crosslinking was reversed by incubation overnight at 65 °C. RNA extraction and RT–PCR analysis were performed to identify interacting RNA segments. Inputs were 2% of the sample. Primers are listed in Supplementary Table 1.

Adenoviral infections. Adenoviruses encoding GFP, HIF-1α and HIF-2α were generated and used as described previously[31,32].

Analysis of cap-binding proteins. Cells on two 150-mm plates were washed with PBS and lysed in 1 ml of lysis buffer (20 mM Tris-HCl pH 7.4, 100 mM NaCl, 25 mM $MgCl_2$, 0.5% Nonidet P40, plus standard protease and phosphatase inhibitors). Extracts were clarified by centrifugation for 10 min at 10,000*g* and 4 °C. Supernatants were pre-cleared for 10 min with 30 μl of Sepharose 4B beads (Sigma) at 4 °C. Beads were removed by centrifugation for 30 s at 500*g*, and supernatants were incubated for 1 h with 50 μl of 7-methyl GTP-Sepharose 4B beads (GE Healthcare) at 4 °C. Pelleted beads were washed four times with 0.5 ml of lysis buffer and resuspended for 1 h in 0.6 ml of lysis buffer containing 1 mM GTP at 4 °C. After four final washes with lysis buffer, the beads were resuspended in sample buffer and boiled for 1 min. Concentrated GTP wash, m^7-GTP-bound proteins and 5% input taken just before m^7-GTP beads were added were subjected to SDS–PAGE.

Statistical analysis. *P* values associated with all comparisons were based on paired two-tailed Student's *t*-tests. Results are shown as means and s.e.m. ($n = 3$).

31. Gunaratnam, L. *et al.* Hypoxia inducible factor activates the transforming growth factor-α/epidermal growth factor receptor growth stimulatory pathway in VHL$^{-/-}$ renal cell carcinoma cells. *J. Biol. Chem.* **278,** 44966–44974 (2003).
32. Smith, K. *et al.* Silencing of epidermal growth factor receptor suppresses hypoxia-inducible factor-2-driven VHL$^{-/-}$ renal cancer. *Cancer Res.* **65,** 5221–5230 (2005).

LETTER

doi:10.1038/nature11054

Crystal structure of an orthologue of the NaChBac voltage-gated sodium channel

Xu Zhang[1,2*], Wenlin Ren[1,2*], Paul DeCaen[3,4*], Chuangye Yan[1,2], Xiao Tao[5], Lin Tang[6], Jingjing Wang[7], Kazuya Hasegawa[8], Takashi Kumasaka[8], Jianhua He[6], Jiawei Wang[1], David E. Clapham[3,4] & Nieng Yan[1,2]

Voltage-gated sodium ($\mathrm{Na_v}$) channels are essential for the rapid depolarization of nerve and muscle[1], and are important drug targets[2]. Determination of the structures of $\mathrm{Na_v}$ channels will shed light on ion channel mechanisms and facilitate potential clinical applications. A family of bacterial $\mathrm{Na_v}$ channels, exemplified by the $\mathrm{Na^+}$-selective channel of bacteria (NaChBac)[3], provides a useful model system for structure–function analysis. Here we report the crystal structure of $\mathrm{Na_vRh}$, a NaChBac orthologue from the marine *alphaproteobacterium HIMB114* (*Rickettsiales sp. HIMB114*; denoted Rh), at 3.05 Å resolution. The channel comprises an asymmetric tetramer. The carbonyl oxygen atoms of Thr 178 and Leu 179 constitute an inner site within the selectivity filter where a hydrated $\mathrm{Ca^{2+}}$ resides in the crystal structure. The outer mouth of the $\mathrm{Na^+}$ selectivity filter, defined by Ser 181 and Glu 183, is closed, as is the activation gate at the intracellular side of the pore. The voltage sensors adopt a depolarized conformation in which all the gating charges are exposed to the extracellular environment. We propose that $\mathrm{Na_vRh}$ is in an 'inactivated' conformation. Comparison of $\mathrm{Na_vRh}$ with $\mathrm{Na_vAb}$[4] reveals considerable conformational rearrangements that may underlie the electromechanical coupling mechanism of voltage-gated channels.

$\mathrm{Na_v}$ channels initiate and propagate action potentials in excitable cells[1]. Because they underlie several clinical disorders such as epileptic seizures and cardiac arrhythmias, they are important drug targets[2]. Eukaryotic $\mathrm{Na_v}$ channel pore-forming α- subunits[5] are single polypeptide chains that are organized into four repeated domains of six transmembrane-spanning (S1–S6) segments. The S5 and S6 segments from each domain form the channel pore that is flanked by the four S1–S4 voltage-sensing domains (VSDs). The VSD provides the molecular basis for voltage sensing in voltage-dependent channels and enzymes[6–9].

VSDs contain the gating charges[10] embodied in a set of highly conserved positively charged side chains occurring every three residues along the S4 segment. In voltage-gated potassium ($\mathrm{K_v}$) channels, approximately 12 gating charges per channel are transferred across the membrane from the cytosolic side to the extracellular side[11,12]. Although multiple models of voltage sensor activation have been proposed, it is generally accepted that the outward translation of S4 segments is coupled to pore opening by the interactions between the S4–S5 connecting helices and S6 segments[7,8].

Less well understood than the voltage-gated channel activation mechanism is its intricate inactivation mechanism. Fast or N-type inactivation, taking place on a millisecond scale, is performed by a cytoplasmic moiety between repeats III and IV of $\mathrm{Na_v}$ channels[13], or by the amino terminus of the *Shaker* $\mathrm{K^+}$ channel[14–16], respectively. Also during prolonged depolarization, slow or 'C-type' inactivation[17] is thought to result from a conformational change of the selectivity filter[18,19].

The prokaryotic homologues of $\mathrm{Na_v}$ channels are homo-tetramers of 6-TM subunits. Interestingly, the sequence of NaChBac is closer to that of $\mathrm{Ca_v}$ channels[20]. Thus, structural determination of NaChBac homologues is expected to provide insights into both $\mathrm{Na_v}$ and $\mathrm{Ca_v}$ channels. To promote a deeper mechanistic understanding of $\mathrm{Na_v}$ channels, we determined the crystal structure of a NaChBac homologue, $\mathrm{Na_vRh}$ (Supplementary Figs 1–5 and Supplementary Tables 1 and 2). During our structural refinement, the $\mathrm{Na_vAb}$ structure was published[4]. Compared with $\mathrm{Na_vAb}$, $\mathrm{Na_vRh}$ reveals several distinct and mechanistically informative structural features.

As in all the known structures of voltage-gated channels, the VSD of one protomer attaches to the pore-forming unit of the adjacent protomer (Fig. 1a). The activation gate formed by S6 of $\mathrm{Na_vRh}$ is closed, although Leu 219, the residue that occludes the gate, is one helical turn above the functionally equivalent Met 221 in $\mathrm{Na_vAb}$ (Fig. 1b)[4]. Notably, the narrowest point along the pore is at Ser 181, which with Glu 183 encloses the entrance to the selectivity filter vestibule. Although the selectivity filter of $\mathrm{Na_vAb}$ is open and may allow the conductance of hydrated $\mathrm{Na^+}$, that of $\mathrm{Na_vRh}$ is closed (Fig. 1b, c).

The pore domain and the VSDs of $\mathrm{Na_vRh}$ exhibit structural variations among the four protomers, resulting in an asymmetric tetramer. The selectivity filter of $\mathrm{Na_vRh}$ (178-TLSSWE-183) connects P1 (corresponding to the P-helix in $\mathrm{K^+}$ channel) and P2 helices (Fig. 2 and Supplementary Fig. 6). The side groups of Ser 180, Ser 181 and Glu 183, as well as the carbonyl oxygen atoms (C=O) of Thr 178 and Leu 179, constitute the electronegative vestibule of the selectivity filter (Fig. 2a). The entrance to the selectivity filter is negatively charged owing to Glu 183. The side groups of Ser 181 adopt distinctive conformations among the four protomers, leading to the observed asymmetry (Supplementary Fig. 6).

Two residues in the selectivity filter of $\mathrm{Na_vRh}$ and $\mathrm{Na_vAb}$ ($\mathrm{Na_vRh}$-Ser 180/Glu 183 versus Glu 178/Ser 181 of $\mathrm{Na_vAb}$) are swapped in the primary sequences. However, structural superimposition shows that the carboxylate groups of $\mathrm{Na_vRh}$-Glu 183 are positioned similarly to those of Glu 178 in the adjacent protomer of $\mathrm{Na_vAb}$ despite their distinct backbone locations (Fig. 2b). Therefore, $\mathrm{Na_vRh}$-Glu 183 and $\mathrm{Na_vAb}$-Glu 178 seem to be functional equivalents. This provides a basis to begin to understand the function of the negatively charged residues in the eukaryotic $\mathrm{Na_v}$ channels that are located at different positions in the selectivity filter (Fig. 2b, left panel).

Like many other bacterial channels, $\mathrm{Na_vRh}$ did not yield measurable ion currents when expressed in insect or mammalian cell lines or when expressed in *Escherichia coli*, purified, reconstituted into lipids and fused into bilayers with lipid composition of either POPE:POPG (3:1 mass ratio) or DPhPC. To test the selectivity of the $\mathrm{Na_vRh}$ channel, we generated a chimaera by replacing the selectivity filter of NaChBac

[1] State Key Laboratory of Bio-membrane and Membrane Biotechnology, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [2] Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China. [3] Howard Hughes Medical Institute, Department of Cardiology, Children's Hospital Boston, Boston, Massachusetts 02115, USA. [4] Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA. [5] Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, Howard Hughes Medical Institute, 1230 York Avenue, New York, New York 10065, USA. [6] Shanghai Institute of Applied Physics, Chinese Academy of Sciences, 239 Zhangheng Road, Shanghai 201204, China. [7] State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. [8] Japan Synchrotron Radiation Research Institute (SPring-8), 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5198, Japan.
*These authors contributed equally to this work.

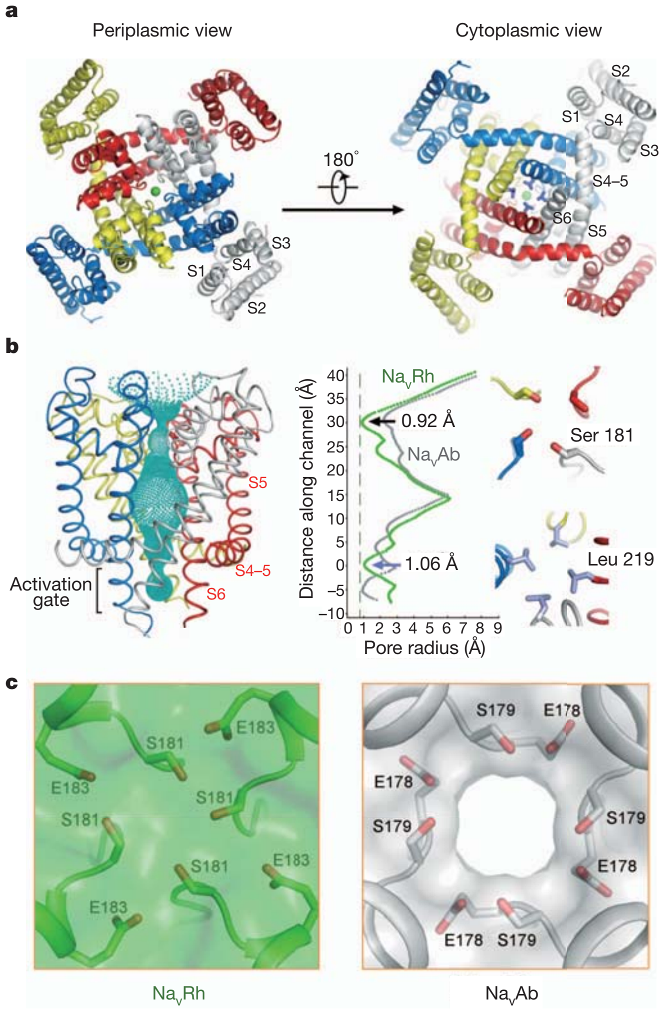


Figure 1 | The structure of Na$_\text{v}$Rh exhibits a closed conformation. a, Na$_\text{v}$Rh crystallized as an asymmetric tetramer. The green sphere indicates the bound ion within the selectivity filter. Leu 219, which occludes the activation gate, is shown in light purple sticks in the cytoplasmic view. **b**, Na$_\text{v}$Rh is closed at the activation gate and the entrance to the selectivity filter. The channel passage (left panel) is indicated by cyan dots. The pore radii (right panel) of Na$_\text{v}$Rh (green) are compared with those of Na$_\text{v}$Ab (grey). The residues that constitute the constriction sites, Ser 181 at the entrance to the selectivity filter and Leu 219 at the activation gate, are shown in sticks in periplasmic and cytoplasmic views, respectively. **c**, A semi-transparent surface illustration of the periplasmic entrance to the selectivity filter in Na$_\text{v}$Rh and Na$_\text{v}$Ab.

with that of Na$_\text{v}$Rh (Supplementary Fig. 7). The chimaeric channel was Na$^+$-selective when expressed in HEK-293 cells and measured under voltage clamp (Fig. 2c). Similar to observations with other NaChBac pore mutations[21,22], the voltage dependence of activation was shifted (+49 mV; Fig. 2d) with an altered rate of inactivation (1.6× increase; Fig. 2d). Similar to NaChBac, the chimaera was blocked by the Na$_\text{v}$ channel antagonist lidocaine, and the Ca$_\text{v}$ antagonist nifedipine (Supplementary Fig. 7d), but was insensitive to tetrodotoxin.

During structure refinement, a spherical electron density was observed in the selectivity filter (Supplementary Fig. 8a). We believe this represents a calcium ion because obtaining well-diffracting crystals depended on the addition 100 mM CaCl$_2$. Furthermore, crystals were obtained from proteins purified in solutions with RbCl instead of NaCl, and diffracted X-ray at the high remote wavelength for Rb$^+$. After structural refinement no anomalous signal for Rb$^+$ was observed whereas the omit electron density persisted, suggesting that the electron density was from Ca^{2+}.

When Ca^{2+} was built into the 3.05 Å structure and further refined, the $2F_\text{o}-F_\text{c}$ electron density at 1.5σ had an elongated tail on the side of the ion facing the central cavity (Supplementary Fig. 8b). A water molecule was then built into the appendage ~2.4 Å away from the ion, fulfilling the geometric restraint of the interaction between water and Ca^{2+} (Fig. 2e). The Ca^{2+} ion is caged by the eight C=O groups from Thr 178 and Leu 179. The distances between Ca^{2+} and the C=O are in the range 3.5–4.6 Å (Fig. 2e). For direct coordination of a Ca^{2+} by C=O, the distance is usually between 2.3 and 2.5 Å (ref. 23). The well-defined electron density suggests that the ion is stabilized in a fully or mostly hydrated state. However, there was no distinguishable electron density for the surrounding water molecules, perhaps because of the moderate resolution of the structure and/or the intrinsic motility of those water molecules. Ca^{2+} and Na$^+$ effective ionic radii are practically identical (1.00 Å compared with 1.02 Å), but including the primary hydration shell the radii are 2.7 Å for Ca^{2+}and 2.2 Å for Na$^+$ (ref. 24). The observation that the inner binding site of the selectivity filter is spacious enough to accommodate a hydrated Ca^{2+} or Na$^+$ thus provides structural evidence for the hypothesis that Na$_\text{v}$ channels allow the passage of mostly hydrated Na$^+$ (ref. 25).

Because the chimaera was impermeant to Ca^{2+} (Fig. 2c), we proposed that Ca^{2+} might block Na$^+$ permeation[26]. Indeed, the Na$^+$ currents from NaChBac and the chimaera were substantially blocked by millimolar concentrations of Ca^{2+} and micromolar concentrations of Cd^{2+} (Fig. 2f). We speculate that divalent ions are able to enter the channel and occlude the pore at the Leu 179/Thr 178 site.

Compared with the subtle conformational variations of the filter residues among the four protomers, the divergences of VSDs are more prominent, particularly for S3–S4 linkers (Fig. 3a). Unlike the VSDs of K$_\text{v}$ channels, in which the carboxy-terminal end of S3 and the N-terminal half of S4 form a paddle-like structure[6–8], the C-terminal fragments of S3 in Na$_\text{v}$Rh and Na$_\text{v}$Ab are unwound. We name the four Na$_\text{v}$Rh protomers Mol A to Mol D. S3–S4 linkers in Mol A and Mol C are not resolved, whereas those in Mol B and Mol D show distinct conformations, dissimilar to that of Na$_\text{v}$Ab (Supplementary Fig. 9). The flexibility of the S3–S4 linker may allow the movement of the S4 segment during voltage sensing.

Transmembrane segments S1–S4 of the four VSDs can be superimposed with root mean squared deviation (r.m.s.d.) values within 0.9 Å over 71–81 Cα atoms. Consistent with a 0 mV field during crystallization, all four conserved Arg residues on the S4 segment point extracellularly, representing a depolarized ('up') conformation (Fig. 3b). The external negative clusters stabilize the gating charges through two invariant interactions: R4 interacts with Asp 48 on S2 and R3 is hydrogen-bonded to the C=O of Ile 90 on S3 (Fig. 3c). In addition to these invariant interactions, there are additional stabilizing contacts specific to individual VSDs. In Mol A, an extra hydrogen bond is between R2 and the C=O of Asn 25 on S1. In Mol B, R3 binds to the anion position, An1. In Mol C, R3 is further hydrogen-bonded to the C=O of Ser 88.

The structure of Na$_\text{v}$Rh has a closed inner gate and the VSDs are in a depolarized ('open') conformation. Similar features were described for Na$_\text{v}$Ab, which was proposed to be in the 'pre-open' conformation[4]. Although the structure of Na$_\text{v}$Rh may also represent a pre-open state, an alternative interpretation is that Na$_\text{v}$Rh is in an inactivated state. The following lines of evidence support this speculation: NaChBac homologues and the NaChBac/Na$_\text{v}$Rh-filter chimaera undergo inactivation on a millisecond to second timescale (Fig. 2d)[17], and because purification and crystallization of the proteins occurs over days at 0 mV, we assume this should favour complete inactivation of the channel. There is little interaction between S4–S5 connecting helices and S6 segments (Fig. 3a), indicating a loss of coupling between the voltage sensor and the inner gate. Furthermore, the structure of Na$_\text{v}$Rh is consistent with a possible form of inactivation discussed by Schmidt *et al.* to account for the gating properties of K$_\text{v}$AP[27]; when S4–S5 linkers release their constriction of the S6 helices, the inner gate may close even if the VSDs are

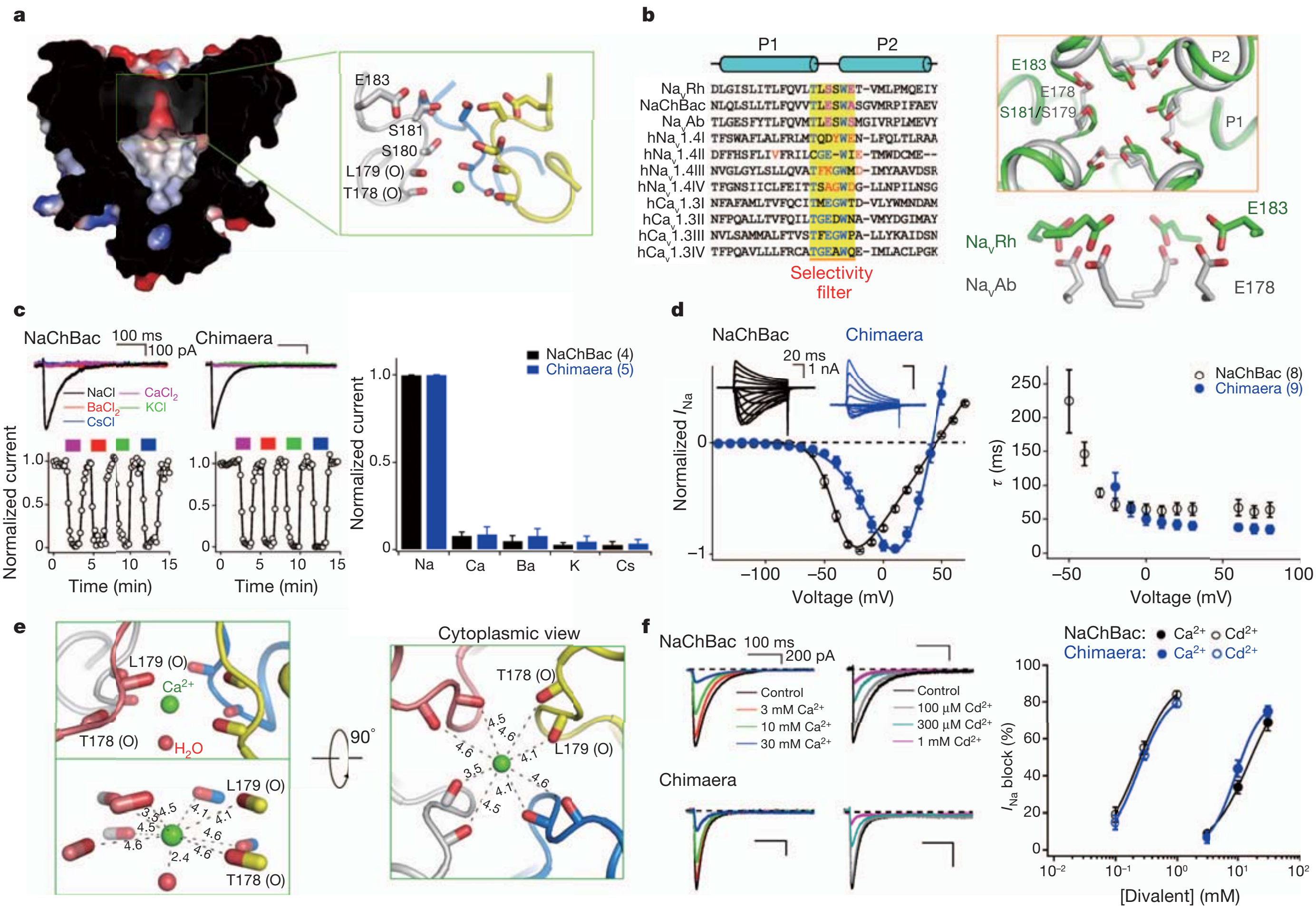


Figure 2 | A Ca^{2+} ion is bound in the asymmetric selectivity filter of Na$_v$Rh. **a**, Side view of the selectivity filter of Na$_v$Rh. **b**, The carboxylate groups of Na$_v$Rh-E183 and Na$_v$Ab-E178 are positioned similarly despite their distinct Cα locations within the selectivity filter. The residues that are important for slow inactivation of Na$_v$1.4 are coloured red in the sequence alignment. **c**, The Na$_v$Rh selectivity filter is sodium selective. Error bars, s.e.m. **d**, Current–voltage relationships for NaChBac and the NaChBac/Na$_v$Rh-filter chimaera, and rates of inactivation (τ) of I_{Na}. **e**, A Ca^{2+} ion is bound at an inner site within the selectivity filter. The distances between the ion and the surrounding groups are in Å. '(O)' is the carbonyl oxygen. **f**, Progressive reduction in I_{Na} by the addition of Ca^{2+} or Cd^{2+}. I_{Na} block potency is estimated in the right panel.

still in the up conformation. In both Na$_v$ and Shaker K channels, mutagenesis analyses suggested that the selectivity filter residues are involved in C-type inactivation[17,19]. In rat Na$_v$1.4, residues Glu 403, Glu 758, Asp 1241 and Asp 1532, which correspond to Glu 183 in Na$_v$Rh (Fig. 2b and Supplementary Fig. 1) are important for inactivation[28]. In our structure of Na$_v$Rh, Glu 183 and Ser 181 collectively close the outer mouth to the selectivity filter (Fig. 1c), supporting the reported functional significance of the outer negative charges in the inactivation process. We believe that the structure of Na$_v$Rh shown here represents an inactivated conformation.

Superposition of the pore domains of Na$_v$Rh and Na$_v$Ab reveal prominent conformational changes of the VSDs. Viewed from the cytoplasm, the VSDs of Na$_v$Rh are rotated anticlockwise around the pore axis by ~30° and the relative positions of the VSDs in Na$_v$Rh are more like those in the depolarized and open conformation of K$_v$1.2 (ref. 7) (Fig. 4a). When the individual VSDs of Na$_v$Rh and Na$_v$Ab were compared by superimposing the S4–S5 linkers (Supplementary Fig. 10a), the VSDs diverge from each other, suggesting that the VSD and S4–S5 linker do not move as a single unit. Asp 48 (An1), Phe 55 and Glu 58 (An2) on the S2 segment constitute the charge transfer centre (CTC)[29] in Na$_v$Rh. Superimposing the VSDs of Na$_v$Rh and Na$_v$Ab relative to the CTC, it is clear that the other transmembrane segments now are discordant, indicating a considerable intra-domain rearrangement within the VSD (Supplementary Fig. 10b).

Gating charges are transferred in response to a change in transmembrane voltage. Superimposition of Na$_v$Rh and Na$_v$Ab VSDs relative to the CTC unambiguously shows that there is a one helical turn shift of Na$_v$Rh-S4 towards the extracellular side (Fig. 4b): for each Na$_v$Rh VSD, one more charge is transferred than for each Na$_v$Ab VSD. In Na$_v$Ab, S4 exists as a 3_{10}-helix from R1 to R4 (ref. 4). In Na$_v$Rh, however, whereas the segment from R3 to R4 forms a 3_{10}-helix, the preceding segment is an α-helix (Fig. 4b). We also compared the voltage sensors of Na$_v$Rh with those of K$_v$1.2 and the paddle chimaera, in which only the C-terminal halves of S4 segments containing R3–R5 adopt a 3_{10}-helix whereas the segments containing R0–R2 are relaxed into α-helices. However, when the CTCs are superimposed, the relative positions of the gating charges, exemplified by R4 of Na$_v$Rh, are located between those of K$_v$1.2 and the paddle chimaera (Fig. 4c).

The availability of voltage sensor structures with unique positions of the gating charges relative to the CTC provides evidence that supports our structure-based animation of gating charge transfer (Fig. 4d and Supplementary Movie 1), which illustrates how R3 and R4 are stabilized sequentially by An1 and An2 to lower the energy barrier during the transfer of R4 across the occluding Phe residue. It also shows the secondary structure transition between 3_{10}- and α-helices concurrent with the translational motion of S4 segment relative to CTC, which exemplifies the 'concertina effect' discussed for K$_v$ channels[8], and is consistent with the disulphide cross-linking experiments of NaChBac[30].

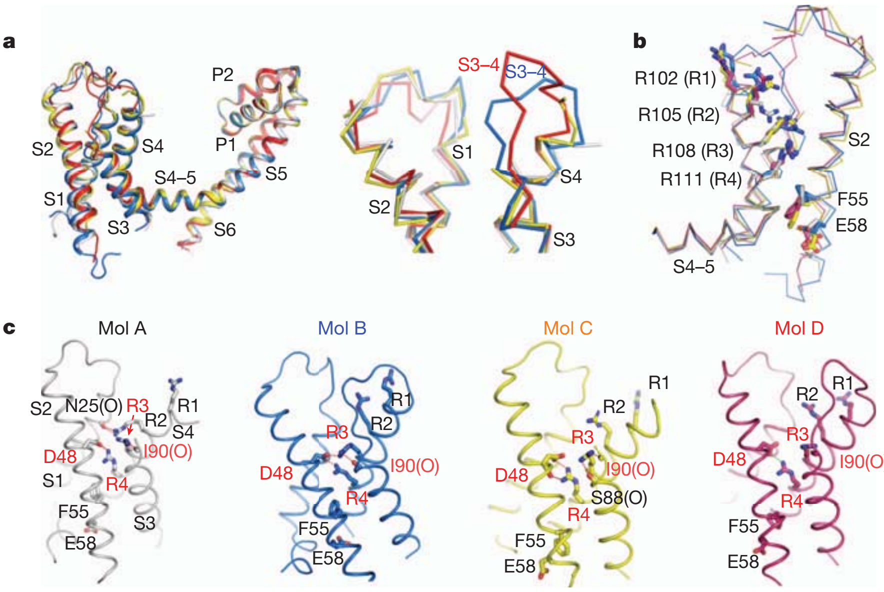


Figure 3 | The VSDs of Na$_v$Rh exhibit a depolarized conformation. **a**, Superimposition of the four protomers in the Na$_v$Rh structure. An enlarged view of S3–S4 linkers is shown (right) to highlight their conformational distinctions. Details can be found in Supplementary Fig. 9. **b**, The S1–S4 segments of the four VSDs exhibit similar conformations with all the gating charges pointing to the extracellular surface. The gating charges (R1–R4) on the S4 segment as well as Phe 55 and Glu 58 on the S2 segment are shown as sticks. **c**, The coordination of the gating charges in the four VSDs of Na$_v$Rh. Hydrogen bonds are represented by red (Mols A–C) or green (Mol D) dashed lines. Residues that mediate invariant interactions between gating charges and the external-negative cluster are labelled in red.

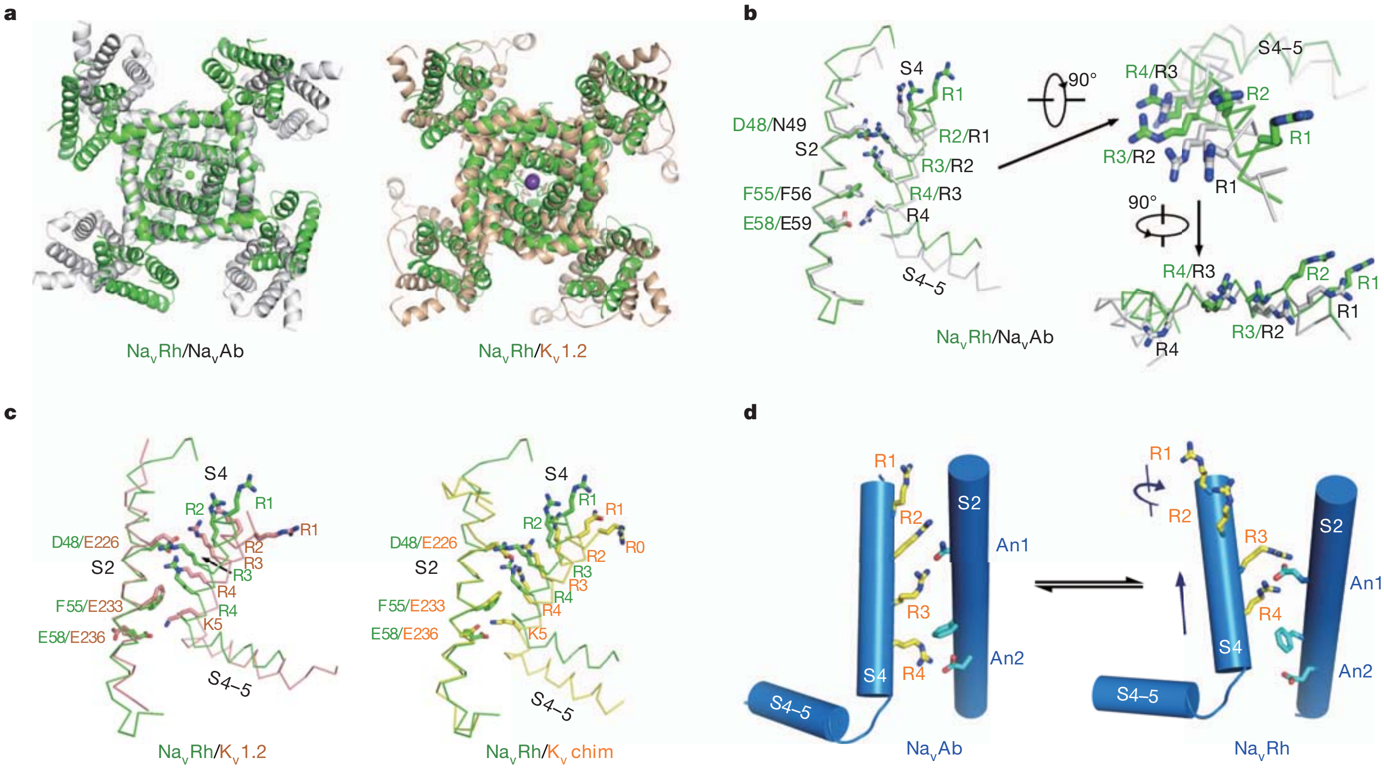


Figure 4 | Molecular basis of charge transfer of VSDs. **a**, Superimposition of the structures of Na$_v$Rh (green), Na$_v$Ab (grey) and K$_v$1.2 (brown), relative to the pore domains. Cytoplasmic views are shown. Ca^{2+} and K$^+$ are shown in green and purple spheres. **b**, There is a one helical turn shift towards the extracellular side of Na$_v$Rh-S4 compared with Na$_v$Ab-S4 when the CTCs are superimposed. The segment of Na$_v$Rh-S4 containing R1 and R2 is an α-helix whereas the corresponding segment of Na$_v$Ab-S4 is a 3_{10}-helix. **c**, Structural comparison of the S4 segments from Na$_v$Rh, K$_v$1.2 and the paddle chimaera. Structures are superimposed against the CTC. **d**, The process of one charge (R4) transfer across the occluding residue, Phe, within the CTC. An animation is shown in Supplementary Movie 1.

The flexibility of the S3–S4 linker may lower the energy barrier during the motion and the secondary structural transition of S4 segment (Fig. 3a). The complementary studies of Na_vRh and Na_vAb thus provide an important framework for future functional and mechanistic investigations of voltage-gated ion channels.

METHODS SUMMARY

Details of the subcloning, purification, crystallization and structural determination of Na_vRh are described in Methods. In brief, full-length Na_vRh was overexpressed in *E. coli* BL21(DE3) and purified to homogeneity in the presence of 0.4% (w/v) *n*-nonyl-β-D-glucopyranoside and 0.1 mg ml^{-1} POPC:POPE:POPG at 3:1:1 mass ratios. The wild-type protein crystallized in two space groups: $P4_12_12$ and $P4_2$, diffracting to ~4.5 and 3.7 Å, respectively. Introduction of a single-point mutation G208S or G208A and post-crystallization manipulation improved the resolution to beyond 3 Å. Addition of 100 mM $CaCl_2$ is a prerequisite for obtaining well-diffracting crystals. The initial phases were derived from mercury-based single-wavelength anomalous dispersion (Supplementary Fig. 2 and Supplementary Table 1). The electron density was of excellent quality for the pore domain and reasonably good for VSDs (Supplementary Fig. 3). The final atomic model was refined to 3.05 Å resolution (Supplementary Table 1). The structure of wild-type protein was refined to 3.7 Å (Supplementary Table 2). There were no detectible differences between the structures of wild-type and the G208S variant at the available resolutions (Supplementary Fig. 4). For simplicity, we use the name Na_vRh to refer to the variant containing G208S for structural description. Structure of Na_vRh in the $P4_2$ space group allowed the structural determination of Na_vRh in the $P4_12_12$ space group to 4.4 Å (Supplementary Table 2). One Na_vRh tetramer is in each asymmetric unit in both crystal forms, and the protein exhibits a nearly identical conformation despite the distinct crystal lattice packing in the two space groups (Supplementary Fig. 5). Structural analysis is based on the 3.05 Å structure in the $P4_2$ space group. For electrophysiology, whole-cell voltage-clamp experiments were performed at 22 °C on transiently transfected HEK-293 cells.

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

Received 11 November 2011; accepted 16 March 2012.
Published online 20 May 2012.

1. Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, 2001).
2. Mantegazza, M., Curia, G., Biagini, G., Ragsdale, D. S. & Avoli, M. Voltage-gated sodium channels as therapeutic targets in epilepsy and other neurological disorders. *Lancet Neurol.* **9,** 413–424 (2010).
3. Ren, D. *et al.* A prokaryotic voltage-gated sodium channel. *Science* **294,** 2372–2375 (2001).
4. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475,** 353–358 (2011).
5. Catterall, W. A. The molecular basis of neuronal excitability. *Science* **223,** 653–661 (1984).
6. Jiang, Y. *et al.* X-ray structure of a voltage-dependent K^+ channel. *Nature* **423,** 33–41 (2003).
7. Long, S. B., Campbell, E. B. & Mackinnon, R. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* **309,** 903–908 (2005).
8. 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).
9. Butterwick, J. A. & MacKinnon, R. Solution structure and phospholipid interactions of the isolated voltage-sensor domain from KvAP. *J. Mol. Biol.* **403,** 591–606 (2010).
10. Armstrong, C. M. & Bezanilla, F. Charge movement associated with the opening and closing of the activation gates of the Na channels. *J. Gen. Physiol.* **63,** 533–552 (1974).
11. Aggarwal, S. K. & MacKinnon, R. Contribution of the S4 segment to gating charge in the Shaker K^+ channel. *Neuron* **16,** 1169–1177 (1996).
12. Seoh, S. A., Sigg, D., Papazian, D. M. & Bezanilla, F. Voltage-sensing residues in the S2 and S4 segments of the *Shaker* K^+ channel. *Neuron* **16,** 1159–1167 (1996).
13. Vassilev, P. M., Scheuer, T. & Catterall, W. A. Identification of an intracellular peptide segment involved in sodium channel inactivation. *Science* **241,** 1658–1661 (1988).
14. Armstrong, C. M. & Bezanilla, F. Currents related to movement of the gating particles of the sodium channels. *Nature* **242,** 459–461 (1973).
15. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. Biophysical and molecular mechanisms of *Shaker* potassium channel inactivation. *Science* **250,** 533–538 (1990).
16. Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Restoration of inactivation in mutants of *Shaker* potassium channels by a peptide derived from ShB. *Science* **250,** 568–571 (1990).
17. Ulbricht, W. Sodium channel inactivation: molecular determinants and modulation. *Physiol. Rev.* **85,** 1271–1301 (2005).
18. Todt, H., Dudley, S. C. Jr, Kyle, J. W., French, R. J. & Fozzard, H. A. Ultra-slow inactivation in mu1 Na^+ channels is produced by a structural rearrangement of the outer vestibule. *Biophys. J.* **76,** 1335–1345 (1999).
19. Yellen, G., Sodickson, D., Chen, T. Y. & Jurman, M. E. An engineered cysteine in the external mouth of a K^+ channel allows inactivation to be modulated by metal binding. *Biophys. J.* **66,** 1068–1075 (1994).
20. Durell, S. R. & Guy, H. R. A putative prokaryote voltage-gated Ca^{2+} channel with only one 6TM motif per subunit. *Biochem. Biophys. Res. Commun.* **281,** 741–746 (2001).
21. Yue, L., Navarro, B., Ren, D., Ramos, A. & Clapham, D. E. The cation selectivity filter of the bacterial sodium channel, NaChBac. *J. Gen. Physiol.* **120,** 845–853 (2002).
22. Pavlov, E. *et al.* The pore, not cytoplasmic domains, underlies inactivation in a prokaryotic sodium channel. *Biophys. J.* **89,** 232–242 (2005).
23. Harding, M. M. The geometry of metal-ligand interactions relevant to proteins. *Acta Crystallogr. D* **55,** 1432–1443 (1999).
24. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32,** 751–767 (1976).
25. Hille, B. The hydration of sodium ions crossing the nerve membrane. *Proc. Natl Acad. Sci. USA* **68,** 280–282 (1971).
26. Armstrong, C. M. & Cota, G. Calcium ion as a cofactor in Na channel gating. *Proc. Natl Acad. Sci. USA* **88,** 6528–6531 (1991).
27. Schmidt, D., Cross, S. R. & MacKinnon, R. A gating model for the archeal voltage-dependent K^+ channel KvAP in DPhPC and POPE:POPG decane lipid bilayers. *J. Mol. Biol.* **390,** 902–912 (2009).
28. Xiong, W., Li, R. A., Tian, Y. & Tomaselli, G. F. Molecular motions of the outer ring of charge of the sodium channel: do they couple to slow inactivation? *J. Gen. Physiol.* **122,** 323–332 (2003).
29. Tao, X., Lee, A., Limapichat, W., Dougherty, D. A. & MacKinnon, R. A gating charge transfer center in voltage sensors. *Science* **328,** 67–73 (2010).
30. DeCaen, P. G., Yarov-Yarovoy, V., Sharp, E. M., Scheuer, T. & Catterall, W. A. Sequential formation of ion pairs during activation of a sodium channel voltage sensor. *Proc. Natl Acad. Sci. USA* **106,** 22498–22503 (2009).

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

Acknowledgements We thank R. MacKinnon at Rockefeller University for discussions and reading the manuscript. We thank L. Feng at Rockefeller University for help. We thank S. Huang and F. Yu at Shanghai Synchrotron Radiation Facility beamline BL17U. K.H. acknowledges SPring-8 beamline BL41XU for proposal 2011A2039. This work was supported by funds from the Ministry of Science and Technology (grant numbers 2009CB918802, 2011CB910501 and 2011CB911102), projects 31125009 and 91017011 of the National Natural Science Foundation of China, and funds from Tsinghua University.

Author Contributions X.Z., W.R., P.D., X.T., D.E.C. and N.Y. designed experiments. X.Z., W.R., P.D., C.Y., X.T., L.T., J.W., K.H., T.K., J.H., J.W. and N.Y. performed the experiments. X.Z., W.R., P.D., C.Y., X.T., J.W., D.E.C. and N.Y. analysed the data. X.Z., P.D., X.T., C.Y., J.W. and D.E.C. contributed to manuscript preparation. N.Y. wrote the manuscript.

Author Information The atomic coordinates of Na_vRh are deposited in Protein Data Bank under accession number 4DXW. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.Y. (nyan@tsinghua.edu.cn).

METHODS

Protein preparation. The complementary DNAs of NaChBac homologues, the sequences of which were codon optimized for *E. coli* expression, were cloned into bacteria expression vectors and the recombinant proteins were overexpressed in *E. coli* BL21(DE3). After screening several dozen homologues, only *alphaproteobacterium HIMB114* (*Rickettsiales sp. HIMB114*; denoted Rh) yielded crystals. The proteins from all homologues were purified without protease inhibitors, which we surmise helps in the selection of the most compact and stable targets[31]. The full-length Na$_v$Rh was cloned into the pET21b vector (Novagen). The Na$_v$Rh mutants were generated using two-step PCR and were subcloned, overexpressed and purified in the same way as wild-type protein. Overexpression of Na$_v$Rh was induced in *E. coli* BL21 (DE3) by 0.2 mM isopropyl-β-D-thiogalactoside when the cell density reached an attenuance, $D_{600\,\mathrm{nm}}$, of 1.5. After growth at 30 °C for 12 h, the cells were collected, re-suspended in a buffer containing 25 mM Tris-HCl, pH 8.0, and 150 mM NaCl, and disrupted by sonication. Cell debris was removed by centrifugation at 27,000*g* for 10 min. The supernatant containing the membrane was collected and applied to ultracentrifugation at 150,000*g* for 1 h. The membrane fraction was collected and incubated with 1.6% (w/v) *n*-dodecyl-β-D-maltopyranoside (Anatrace) for 2 h at 4 °C. After additional ultracentrifugation at 150,000*g* for 30 min, the supernatant was loaded onto a Ni^{2+}-nitrilotriacetate affinity resin (Qiagen). Subsequently, the resin was rinsed three times with 10 ml buffer containing 25 mM Tris-HCl, pH 8.0, 500 mM NaCl, 50 mM imidazole-HCl, pH 8.0, and 0.02% *n*-dodecyl-β-D-maltopyranoside. The protein was eluted from the affinity resin with wash buffer supplemented with 400 mM imidazole-HCl, pH 8.0. The proteins were concentrated to about 15 mg ml^{-1} before applying to gel-filtration chromatography (Superdex-200 10/30, GE Healthcare), which was equilibrated in the buffer containing 25 mM Tris-HCl, pH 8.0, 150 mM NaCl and 0.4% *n*-nonyl-β-D-glucopyranoside (Anatrace). The peak fractions of the protein (~8 mg ml^{-1}) were collected and incubated with 0.1 mg ml^{-1} lipids POPC:POPE:POPG (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine:1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine:1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-1-glycerol, Avanti) at mass ratio 3:1:1 for crystallization trials.

Crystallization. Crystals were grown at 18 °C by the hanging-drop vapour diffusion method. To improve resolution, multiple steps of construct modification, crystal growth optimization and post-crystallization manipulation were explored. In the beginning, the C-terminal His$_6$ tagged, wild-type proteins yielded cubic-shaped crystals in the buffer containing 0.7 M MgSO$_4$, 0.1 M MES-NaOH, pH 6.0. The crystals diffracted to ~8 Å at BL41XU, SPring-8, Japan. Removal of His$_6$-tag improved the crystal quality markedly. Crystals of the wild-type, non-tagged protein appeared overnight in the buffer containing 16% PEG 400 (v/v), 100 mM MES-NaOH, pH 6.5, 100 mM CaCl$_2$, and diffracted to ~4.5 Å at the synchrotron radiation resource. However, it was difficult to scale the data sets to a specific space group. A single-point mutation G208S or G208A improved the data quality. The space group was ultimately assigned as $P4_12_12$ using the data sets obtained for Na$_v$Rh-G208S. Further improvement was achieved with a new crystallization condition. Crystals appeared in the buffer containing 5% PEG 8,000 (w/v), 100 mM HEPES-NaOH, pH 7.0, 100 mM CaCl$_2$, 10% glycerol and 20% 1,4-butandiol in 2 days, and grew to 50 μm × 50 μm × 100 μm tetragonal rods in 5 days. The crystals, in the space group of $P4_2$, were able to break the 4.0 Å diffraction limit but with poor reproducibility and a high mosaicity value (>5). Finally, the best crystals were obtained through dehydration manipulation by equilibrating the crystals at 4 °C with increasing precipitant concentrations in crystallization buffer to 15% PEG 400 and 20% PEG 8,000. The crystals were flash frozen in liquid nitrogen, and diffracted beyond 3.05 Å at Shanghai Synchrotron Radiation Facility beamline BL17U. Mercury derivatives were obtained by soaking the crystals for 3 h in the dehydration solution plus 10 mg ml^{-1} methylmercury chloride (CH$_3$HgCl) as the final concentration.

Data collection and processing. All data sets were collected at the Shanghai Synchrotron Radiation Facility beamline BL17U, except for the native data in the space group of $P4_12_12$, which were collected at the SPring-8 beamline BL41XU. All were integrated and scaled with HKL2000 (ref. 32). Further processing was performed using programs from the *CCP4* suite[33]. Data collection statistics are summarized in Supplementary Table 1.

Experimental phasing and structure refinement. The mercury positions in the Hg-derived crystal of the $P4_2$ space group were determined using the program SHELXD[34]. The identified heavy-atom sites were refined and the initial phases were generated in the program PHASER[35] with the single-wavelength anomalous dispersion experimental phasing module. Cross-crystal averaging combined with solvent flattening, histogram matching and NCS averaging in DMMulti[36] gave rise to electron density maps of sufficient quality for model building, using the data sets in Supplementary Table 1. An initial model was built into the high-resolution $P4_2$ native data using COOT[37]. The structure was refined with PHENIX[38]. All structure figures in the manuscript were prepared with PyMol[39]. The surface electrostatic potential presented in the manuscript was calculated with PyMol. The pore radii were calculated with the program 'HOLE'[40].

Electrophysiology. Whole-cell voltage-clamp experiments were performed at 22 °C in transiently transfected HEK-293 cells. Transfected cells were seeded onto glass coverslips and placed in a perfusion chamber for experiments in which extracellular conditions could be exchanged. Unless otherwise stated, the extracellular solution contained (in mM): NaCl 150, CaCl$_2$ 1.5, MgCl$_2$ 1, glucose 10 and HEPES 10; pH 7.4, and the intracellular (pipette) solution contained (in mM): CsF 105, EGTA 10, NaCl 35, MgCl$_2$ 4 and HEPES 10; pH 7.5. For experiments shown in Fig. 2c, representative current traces were elicited by 0.5 s depolarizations from −140 mV (holding potential) to 0 mV. Na$^+$ was substituted by the ions indicated (150 Cs$^+$, K$^+$; 110 Ca^{2+}, Ba^{2+}). Normalized current magnitudes were plotted as a function of time as Na$^+$-containing solution was exchanged for solutions with the indicated ions (coloured boxes). Error bars, s.e.m.; $n = 4$–5 each. For experiments shown in Fig. 2d, current–voltage relationships were fitted to $(V - V_{\mathrm{rev}})/\{1 + \exp[(V - V_{1/2})/k]\}$, in which V_{rev} is the extrapolated reversal potential, $V_{1/2}$ is the half-activation voltage and k is a slope factor equal to RT/zF (z is the apparent gating charge, R the ideal gas constant and F is Faraday's constant). Half-inactivation voltages were derived from fits to $1/\{1+\exp(V - V_{1/2})/k\}$ to derive steady-state inactivation curves. Inactivating currents during 500 ms pulses were fitted to $C + A(\mathrm{e}^{-t/\tau})$, in which τ is the time constant, A the amplitude and C the baseline. Decay in response to a 500 ms pulse to the indicated potentials was fitted to a single exponential. For experiments shown in Fig. 2f, the half maximal inhibitory concentration (IC$_{50}$) was estimated by fitting the average percentage of inward I_{Na} block at each divalent concentration, where the current amplitude = $1/\{([D]/\mathrm{IC}_{50})^n + 1\}$, in which n is the Hill coefficient and $[D]$ is the respective drug or divalent ion concentration.

Animation. To generate the morph to visualize the conformational change of the S4 segments between Na$_v$Rh and Na$_v$Ab, the homology-based model of Na$_v$Rh was generated using the online SWISS-MODEL workspace[41–43] with the structure of Na$_v$Ab (PDB accession number 3RVY, chain A) as the model. The resulting structure was then superimposed on that of Na$_v$Rh relative to CTC. The shifted coordinates of the modelled structure and the original coordinates of Na$_v$Rh were used as the initial and end states, respectively, for morph generation. The intermediate morphs were obtained with the multiple-chain morphing script[44,45] for Crystallography and NMR System[46,47]. The animations were finally produced using PyMol.

31. Sawaya, M. R., Pelletier, H., Kumar, A., Wilson, S. H. & Kraut, J. Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. *Science* **264,** 1930–1935 (1994).
32. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
33. Collaborative Computational Project, Number 4. The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
34. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
35. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
36. Cowtan, K. dm: an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31,** 34–38 (1994).
37. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
38. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
39. DeLano, W. L. The PyMOL Molecular Graphics System. *Pymol* http://www.pymol.org (2002).
40. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. The pore dimensions of gramicidin A. *Biophys. J.* **65,** 2455–2460 (1993).
41. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22,** 195–201 (2006).
42. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31,** 3381–3385 (2003).
43. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18,** 2714–2723 (1997).
44. Echols, N., Milburn, D. & Gerstein, M. MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res.* **31,** 478–482 (2003).
45. Krebs, W. G. & Gerstein, M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.* **28,** 1665–1675 (2000).
46. Brunger, A. T. *et al. Crystallography & NMR System*: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
47. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).

doi:10.1038/nature11077

Crystal structure of a voltage-gated sodium channel in two potentially inactivated states

Jian Payandeh[1]†, Tamer M. Gamal El-Din[1], Todd Scheuer[1], Ning Zheng[1,2] & William A. Catterall[1]

In excitable cells, voltage-gated sodium (Na_V) channels activate to initiate action potentials and then undergo fast and slow inactivation processes that terminate their ionic conductance[1,2]. Inactivation is a hallmark of Na_V channel function and is critical for control of membrane excitability[3], but the structural basis for this process has remained elusive. Here we report crystallographic snapshots of the wild-type Na_VAb channel from *Arcobacter butzleri* captured in two potentially inactivated states at 3.2 Å resolution. Compared to previous structures of Na_VAb channels with cysteine mutations in the pore-lining S6 helices (ref. 4), the S6 helices and the intracellular activation gate have undergone significant rearrangements: one pair of S6 helices has collapsed towards the central pore axis and the other S6 pair has moved outward to produce a striking dimer-of-dimers configuration. An increase in global structural asymmetry is observed throughout our wild-type Na_VAb models, reshaping the ion selectivity filter at the extracellular end of the pore, the central cavity and its residues that are analogous to the mammalian drug receptor site, and the lateral pore fenestrations. The voltage-sensing domains have also shifted around the perimeter of the pore module in wild-type Na_VAb, compared to the mutant channel, and local structural changes identify a conserved interaction network that connects distant molecular determinants involved in Na_V channel gating and inactivation. These potential inactivated-state structures provide new insights into Na_V channel gating and novel avenues to drug development and therapy for a range of debilitating Na_V channelopathies.

Voltage-gated ion channels share a common architecture, consisting of a central ion-conducting pore module (PM) and four peripheral voltage-sensing domains (VSDs)[4,5]. Voltage-gated potassium (K_v) and bacterial Na_V channels are homotetramers of subunits containing six transmembrane segments (S1–S6)[5,6], whereas vertebrate Na_V and calcium (Ca_v) channels contain four linked homologous domains in a single polypeptide[7]. The S5 and S6 segments of four subunits (or domains) form the PM[7]. The VSDs (S1–S4) place highly conserved S4 gating charges in the membrane electric field, where depolarization causes their outward movement during channel activation[7]. S4 movement is coupled through the S4–S5 linker to the intracellular activation gate to open the pore[4,5]. In mammalian Na_V channels, two physically distinct inactivation processes control the channel activity. Fast inactivation operates on the millisecond timescale and is quickly reversed on repolarization, permitting Na_V channels to be rapidly available for reactivation[7]. A tethered cytoplasmic inactivation gate connecting the third and fourth homologous domains confers fast inactivation through a hinged-lid mechanism[7]. Fast inactivation can be removed by intracellular protease treatment[8] or mutations of the inactivation gate[9] and can be restored by addition of inactivation-gate peptides[10]. By contrast, slow inactivation develops much more slowly during repetitive firing of action potentials and opposes high-frequency spike generation[2,3]. Its essential physiological role is highlighted by disease mutations that affect slow-inactivation[3] and clinically relevant channel blocking drugs bind to and stabilize slow-inactivated states[11,12]. In contrast to fast inactivation, the structures and mechanisms involved in slow inactivation of Na_V channels remain poorly defined.

Bacterial Na_V channels share key physiological properties with their more complex vertebrate descendants, including voltage-dependent activation, inactivation, and pharmacological sensitivity[6]. However, their simple homotetrameric structure leaves bacterial Na_V channels without the fast inactivation gate found in mammalian Na_V channels. Therefore, bacterial Na_V channels are thought to undergo an inactivation process similar to slow inactivation[13]. Mutations near the selectivity filter[13,14], along the pore-lining S6 helices[15–17] and within VSDs of bacterial Na_V channels have dramatic effects on inactivation, similar to slow inactivation in mammalian Na_V channels (Supplementary Discussion). We previously reported the structure of Na_VAb from *A. butzleri* captured in a potentially pre-open state with four activated VSDs and a closed PM[4]. In space group *I*222, the structural model of Na_VAb-I217C was nearly four-fold symmetric with two very similar molecules in the asymmetric unit. The ion conductance pathway displayed a selectivity filter with a central orifice of ~4.6 Å lined by four Glu 177 side chains, followed by two sequential carbonyl sites fit to coordinate a Na^+ ion in complex with a square planar array of hydrating waters[4]. The nearly four-fold symmetrical S6 segments formed a large central cavity and a tightly closed intracellular activation gate[4]. Four lateral pore fenestrations of similar size and shape were seen connecting the hydrophobic membrane phase to the central cavity, and the activated VSDs were arranged in a square array around the PM[4]. Here, through crystallographic and electrophysiological studies, we now describe the structure of Na_VAb in two potentially inactivated states.

When expressed in *Trichoplusia ni* cells, the wild-type Na_VAb channel (Na_VAb-WT) activates and inactivates during 7-ms depolarizing pulses from −180 mV to −40 mV (Fig. 1a). Repetitive 7-ms pulses cause a late phase of slow inactivation that is dependent on the frequency of depolarization and nears completion in 600 s at 0.2 Hz (Fig. 1a). By comparison, Na_VAb-I217C enters this deep slow-inactivated state more slowly and less completely, and its voltage-dependence of activation is shifted towards more negative potentials in a manner that is consistent with stabilization of the pre-open state (Supplementary Fig. 1, Supplementary Tables 1 and 2; Supplementary Discussion). The unusually strong, negatively shifted, and slowly reversible inactivation of Na_VAb-WT in *T. ni* cells suggests that it might enter the late slow-inactivated state during our purification and crystallization procedures and never recover from it.

To gain insights into the native structure of Na_VAb-WT, we solubilized and purified it in a mild detergent and reconstituted it into a phosphatidylcholine-based crystallization system as described for Na_VAb-I217C[4]. Crystals formed in space group $P4_2$ and the Na_VAb-WT structure was phased and refined to 3.2 Å resolution (Supplementary Methods and Supplementary Table 3). Na_VAb-WT channels are arranged in the crystals as though each is embedded in a phospholipid membrane bilayer (Supplementary Fig. 2). Remarkably,

[1]Department of Pharmacology, University of Washington, Seattle, Washington 98195, USA. [2]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. †Present address: Department of Structural Biology, Genentech, Inc., South San Francisco, California 94080, USA.

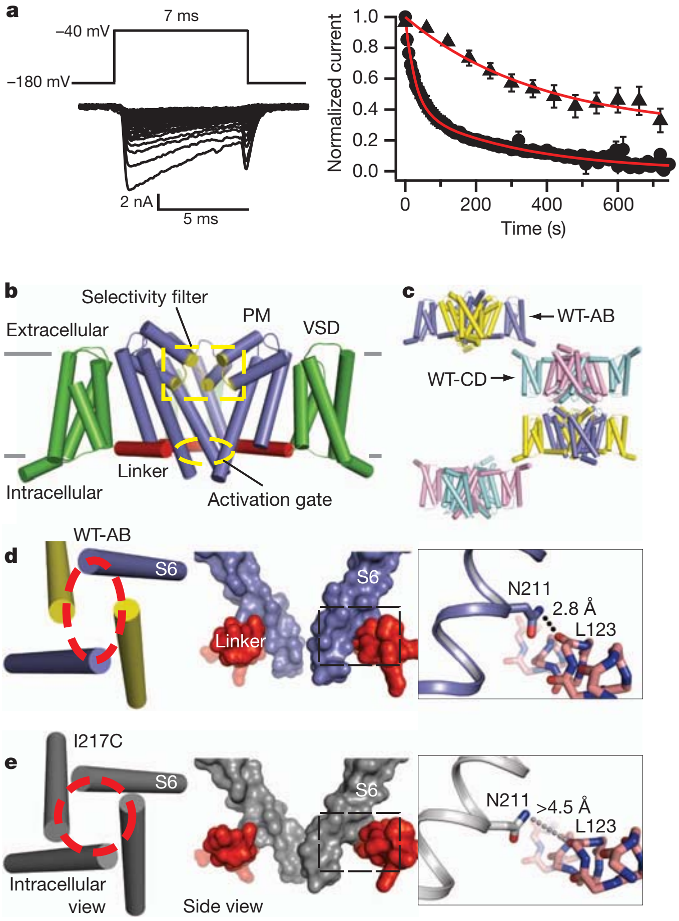


Figure 1 | Structure and function of Na$_V$Ab. **a**, Use-dependent development of slow inactivation in Na$_V$Ab-WT. Depolarizations from a holding potential of −180 mV to −40 mV, 7-ms in duration, were applied at 0.2 Hz (circles) or once per minute (triangles), and the peak current elicited by each pulse was measured. Left, pulse protocol and example current traces in response to stimuli at 0.2 Hz. Right, mean results ± s.e.m. from experiments as shown in left panel. Red lines are fits of a two-exponential function to the data. Currents were normalized to the peak inward current during the first pulse. **b**, Main structural elements of Na$_V$Ab. The nearest VSD and pore domain are removed for clarity. **c**, Packing arrangement in the WT Na$_V$Ab crystals. Four crystallographically independent channel subunits are coloured: blue (chain A), yellow (chain B), cyan (chain C) and pink (chain D). **d**, **e**, Left, red dashed lines indicate the C$_\alpha$ location of D219 (the last S6 residue modelled in WT-AB), where the S6 helices are shown as cylinders. WT-chain A, purple; WT-chain B, yellow; Na$_V$Ab-WT, grey. Middle and right, space-filling models in expanded boxes highlight Asn 211 and the S6 interaction site of WT-chain A with Leu 123 on the S4–S5 linker.

the structure of Na$_V$Ab-WT differs substantially from our previously reported S6-cysteine mutant channels[4] (Fig. 1b–e). Four molecules in the $P4_2$ asymmetric unit give rise to two independent Na$_V$Ab-WT channels composed of dimers of AB and CD subunit conformations, respectively (Na$_V$Ab-AB and Na$_V$Ab-CD; Fig. 1b, c). The Na$_V$Ab-AB and Na$_V$Ab-CD models are each unique and strikingly asymmetric in structure (Fig. 1d and Supplementary Figs 3–5). All VSDs are in an activated conformation, and the PM of both channels appears occluded by collapse of the S6 helices of subunits B or C towards the central axis (Supplementary Figs 3–5).

We aligned the WT Na$_V$Ab-AB and Na$_V$Ab-CD models onto the selectivity filter of Na$_V$Ab-I217C, revealing conformational adjustments that include asymmetric collapse of the S6 activation gate (Fig. 1d), narrowing of the selectivity filter, reshaping of the central cavity and lateral pore fenestrations, and displacement of the VSDs around the PM (see below). These structural features fit well with expectations of a Na$_V$ channel captured in a slow-inactivated state (Supplementary Discussion). Multiple slow-inactivated conformations are predicted from kinetic analyses of Na$_V$ channels[3], and Na$_V$Ab exhibits at least two inactivated states with different kinetics of recovery from inactivation, and differential effects of the S6 mutation I217C (Supplementary Fig. 1). The observation of two discrete conformations of Na$_V$Ab-WT provides a potential structural basis for these functional properties.

The inactivated state of a Na$_V$ channel is expected to be non-conductive. The closed inner ends of the S6 segments in Na$_V$Ab-I217C form a nearly square array (red circle, Fig. 1e) and superimpose well upon other closed-pore tetrameric ion channel structures[4,18]. By contrast, the intracellular activation gate in Na$_V$Ab-AB and Na$_V$Ab-CD has closed in an unprecedented way. Two S6 segments from diagonally opposed subunits have moved closer to the central pore axis, while the adjacent S6 segments have shifted farther away (red oval, Fig. 1d), asymmetrically collapsing the S6 activation gate in these potentially inactivated states. This finding is consistent with biophysical studies[3,15–17] of bacterial and mammalian Na$_V$ channels as well as pathological mutations that have implicated this pore region in slow-inactivation gating[3]. This novel activation gate structure may represent a hallmark of the slow-inactivated state in Na$_V$ channels, and is in sharp contrast to the dilated activation gate observed in inactivated KcsA potassium channels[19].

Conformational shifts of the S5 and S6 segments in Na$_V$Ab are hinged at the extracellular side of the PM near where these segments connect to the pore (P)-helix and P2-helix, respectively (Supplementary Fig. 4c, d). Two S6 segments in Na$_V$Ab-AB interact with the S4–S5 linker from a neighbouring subunit near the intracellular activation gate (Fig. 1d). Electron density suggests that the side chain of Asn 211 forms a stabilizing inter-subunit hydrogen bond with a backbone carbonyl in the S4–S5 linker (Fig. 1d), which is not formed in Na$_V$Ab-I217C (Fig. 1e). Notably, Asn 211 is the only universally conserved S6 residue in all Na$_V$ channels (Supplementary Fig. 6), yet its close interaction with the S4–S5 linker is seen only in two of the four subunits in Na$_V$Ab-AB due to structural asymmetry. Mutations of the equivalent S6 Asn residue in domains I and III of vertebrate Na$_V$ channels, but not in domains II and IV, have dramatic effects on slow-inactivation[20,21]. Thus, Na$_V$Ab may offer the first structural views of a conserved interaction occurring during slow inactivation in the Na$_V$ channel family.

Structural adjustments throughout the PM in Na$_V$Ab-AB and Na$_V$Ab-CD culminate in dramatic effects at the selectivity filter, where accumulating evidence implicates molecular rearrangements in slow inactivation gating in bacterial and mammalian Na$_V$ channels (Supplementary Discussion). The selectivity filter in Na$_V$Ab-I217C is rigidly anchored by a hydrogen bond (~3.0 Å) between Thr 175 and Trp 179 of neighbouring subunits[4]. This conserved landmark interaction forces the Thr 175 and Leu 176 carbonyls to point towards the ion conduction pathway and positions the Glu 177 side chains squarely against the P2-helix (Fig. 2a, b). In Na$_V$Ab-AB, two of the key Thr 175–Trp 179 interactions have become unlatched, as only a very weak hydrogen bond (~3.8 Å) could exist between these partners (Fig. 2a, b). In the unlatched subunits, Ser 180 has also flipped its conformation to engage the Glu 177 carboxylate of a neighbouring subunit (Fig. 2a, b), and formation of this new hydrogen bond may correlate with entry into the inactivated state[13]. Concerted structural changes distort the geometry along the Thr 175–Leu 176 carbonyl funnel, which was perfectly sized to coordinate and conduct a fully hydrated square-planar Na^+ complex in Na$_V$Ab-I217C (Supplementary Figs 7 and 8). Analysis of the Na$_V$Ab-WT pore diameters indicates a 1–2 Å narrowing and distortion of the backbone carbonyl geometry in the central and inner Na^+ coordination sites of the selectivity filter (Fig. 2a, Supplementary Figs 7 and 8), suggesting that Na$_V$Ab would be less permissive to the conduction of optimally hydrated Na^+ ions in these inactivated states.

Consistent with the existence of multiple inactivated states in Na$_V$Ab (Fig. 1, Supplementary Fig. 1), the Thr 175–Trp 179 interaction network remains intact in Na$_V$Ab-CD, as does the predominant

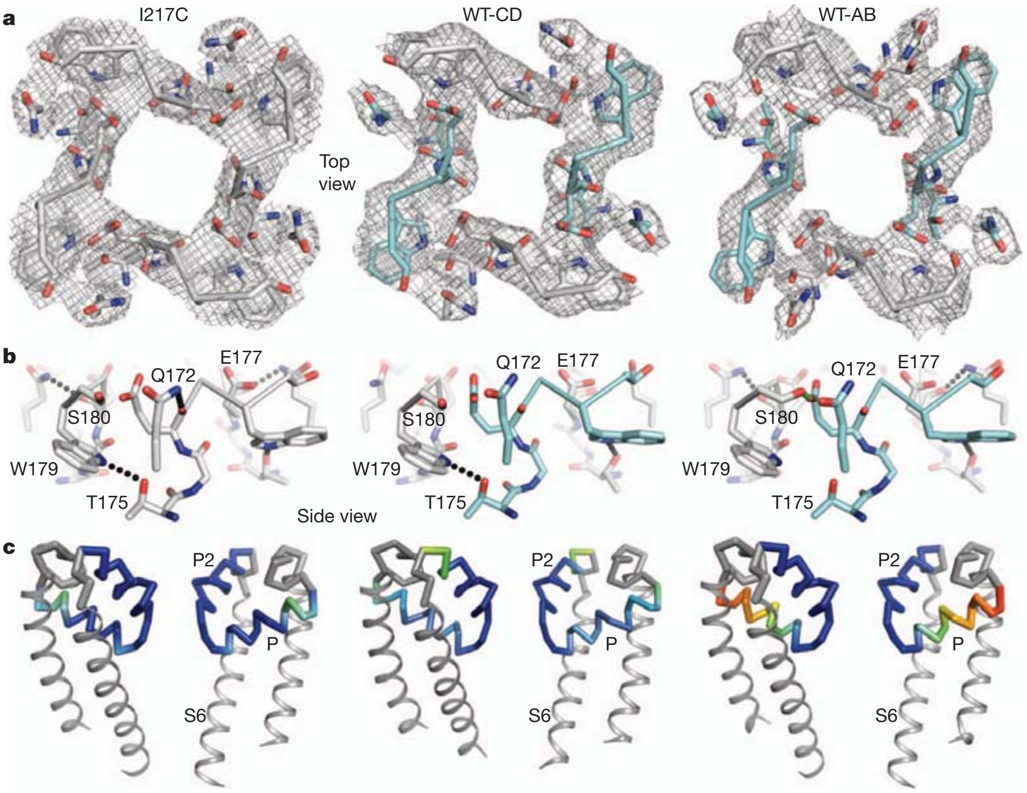


Figure 2 | Structural changes in the selectivity filter of Na_VAb. **a**, Stick representation of the selectivity filter with a $2F_o - F_c$ map calculated at 3.2 Å resolution (grey mesh) contoured at 1.5σ for Na_VAb-I217C and 1.0σ for Na_VAb-AB and Na_VAb-CD. Symmetry-related subunits in WT-AB and WT-CD are coloured white (chains A and D) and cyan (chains B and C), respectively. **b**, Hydrogen bonds (<3.5 Å) supporting the selectivity filter, as discussed in the main text, are shown as dotted lines. **c**, Crystallographic *b*-factors are coloured from low to high (blue to red) along the P-helix, selectivity filter, and P2-helix regions.

side-chain conformation of Ser 180 (Fig. 2a,b). However, the Gln 172 side chain, which makes a strong interaction with the Glu 177 backbone carbonyl in Na_VAb-I217C, is disengaged from this interaction in the Na_VAb-CD channel (Fig. 2a, b). Loss of this supporting interaction would destabilize the selectivity filter, similar to unlatching the Thr 175–Trp 179 network described above in Na_VAb-AB. In fact, comparison of crystallographic temperature factors suggests that the entire P-helix displays increased mobility in these inactivated states (Fig. 2c). Therefore, destabilization of the selectivity filter and concomitant remodelling of the outer pore vestibule in Na_VAb-WT (Fig. 2; Supplementary Figs 5 and 8) may correlate with entry into the inactivated state. This conclusion would be consistent with the effects of toxin binding, permeant ions, and mutations in the selectivity filter of mammalian Na_V channels on slow inactivation (Supplementary Discussion)[3].

Overlaying the structural models of Na_VAb-AB and Na_VAb-CD onto Na_VAb-I217C provides insight into a network of structurally coupled residues unique to Na_V channels that scaffold the selectivity filter and line the surrounding S5 and S6 segments (Fig. 3a, b). Structure-based sequence alignment of the four linked domains from mammalian Na_V channels pinpoints analogous residues in the PM of Na_VAb: Phe 144 and Phe 152 in the S5 segment, Leu 170 and Phe 171 in the P-helix, Trp 179 in the P2-helix, and Phe 198 and Ile 202 in the S6 segment (Fig. 3a, b and Supplementary Fig. 6). Notably, substitution of the Leu 170 equivalent or the Ile 202 equivalent in Na_V1.4 dramatically alters its slow-inactivation[22,23]. These comparisons highlight an evolutionarily conserved network of residues coupling the conformation of the intracellular activation gate to the selectivity filter through a mechanism that results in collapse of the pore into a prominent dimer-of-dimers arrangement for all of the functional elements in Na_VAb-WT.

The structural changes observed in Na_VAb-WT also alter the central cavity, where amino acid side chains analogous to those involved in drug binding in mammalian Na_V channels[24–27] have a different spatial arrangement due to the asymmetric collapse of two S6 segments (Fig. 3c). Local anaesthetics and related pore blockers of mammalian Na_V channels block the Na_VAb homologue NaChBac in a state-dependent manner[6,16]. If similar asymmetric conformational changes occur during inactivation of mammalian Na_V channels (Supplementary Discussion), they could rationalize why pore-blocking drugs bind to and stabilize inactivated states of Na_V channels through interactions with only three of their four S6 segments[25,26].

Four striking lateral pore fenestrations in the PM of Na_VAb-I217C revealed a hydrophobic access pathway to the central cavity[4]. Compared to the nearly identical pore fenestrations in Na_VAb-I217C, two diagonal fenestrations have narrowed in Na_VAb-AB, while the adjacent two fenestrations have opened wider (Fig. 3d, e and Supplementary Figs 9, 10). In Na_VAb-CD, both pairs of pore fenestrations are smaller than in Na_VAb-I217C (Fig. 3d, e and Supplementary Figs 9, 10). Therefore, different-sized drugs and other hydrophobic molecules could potentially gain access to the Na_VAb central cavity through these differentially-sized pore fenestrations (Fig. 3d, e and Supplementary Figs 9, 10). If mammalian Na_V channels have similar pore fenestrations, our Na_VAb structures predict they would provide dynamic drug access to the central cavity during different stages of channel gating, as postulated by the modulated receptor hypothesis of drug action[28].

We can now compare the VSD structures from a single voltage-gated ion channel captured in different pore conformations for the first time. The S4 segments in all Na_VAb structures have a similar 3_{10} helical conformation from R1 to R4 (Fig. 4a), suggesting that S4 does not undergo a major conformational change during inactivation. Surprisingly, crystallographic temperature factors indicate that the S4 segment is the most well-ordered region of the VSD (Fig. 4b). Upon superimposing the pore domains of Na_VAb, a hinge point is seen at the foot of the S5 segment (Fig. 4c), as reported previously when comparing closed-pore Na_VAb-I217C and open-pore K_V1.2 structures[4]. Hence, movements at

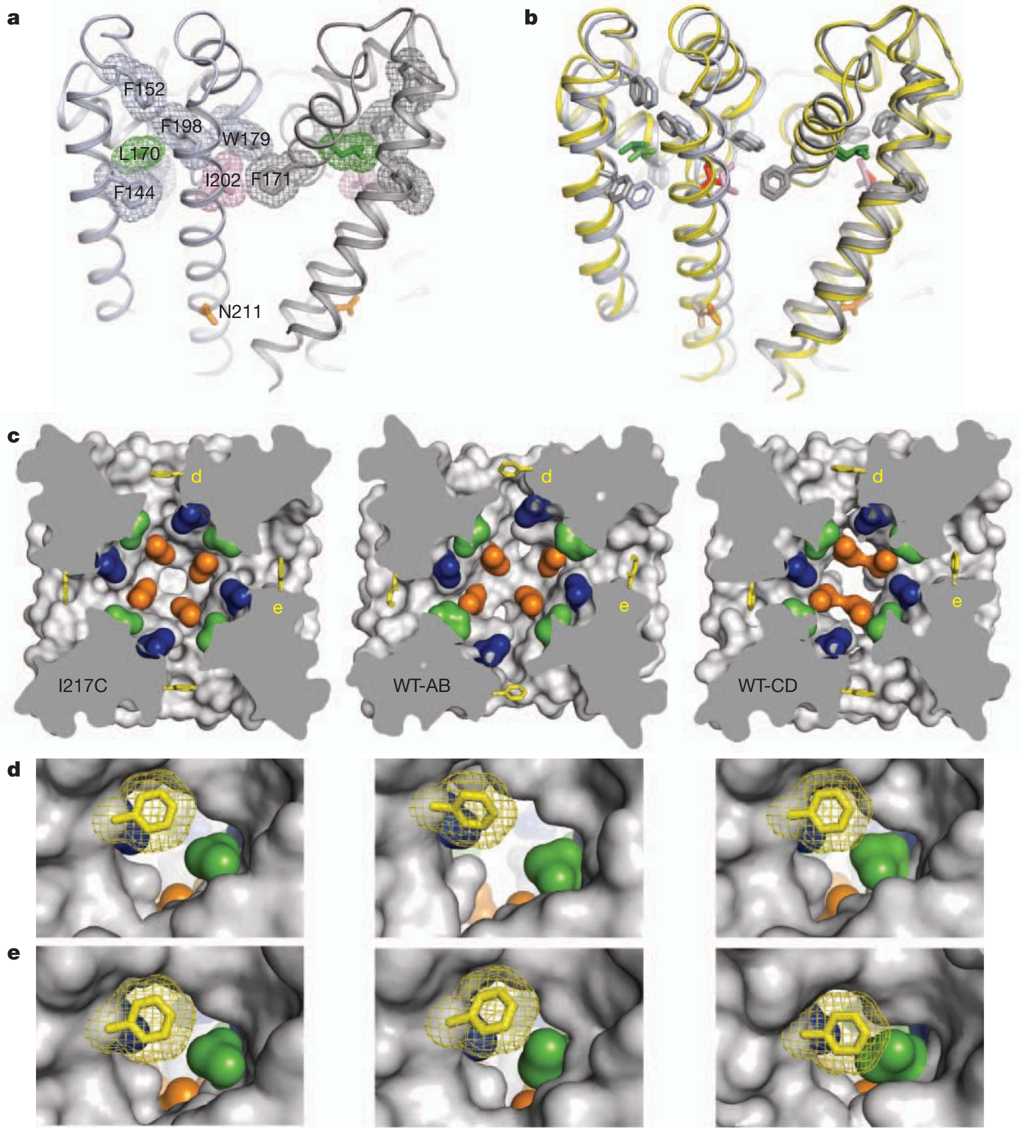


Figure 3 | Conformational changes in the pore module of Na$_V$Ab. **a**, Highly conserved residues in the PM of Na$_V$ channels are shown in stick and mesh representation. Mutations of the Leu 170 and Ile 202 side-chain equivalents (green and pink, respectively) in Na$_V$1.4 alter slow-inactivation in Na$_V$1.4. Asn 211 (S6) is shown for reference. **b**, Superposition of the Na$_V$Ab-I217C (grey) and Na$_V$Ab-AB (yellow) PMs highlights a structurally coupled, evolutionarily conserved set of amino acid residues that communicate between the intracellular activation gate (S6), the central cavity (S5 and S6), and the selectivity filter (P- and P2-helices). **c**, A view through the PM sectioned below the selectivity filter illustrates the lateral pore fenestrations, hydrophobic access to the central cavity, and structural asymmetry in the Na$_V$Ab-AB and Na$_V$Ab-CD pore domains. Phe 203 side chains are yellow sticks. Na$_V$Ab residues implicated in drug binding in vertebrate Na$_V$ channels are coloured: Thr 206 (blue), Met 209 (green) and Val 213 (orange). **d**, **e**, Viewed from the plane of the lipid bilayer, the pore fenestrations provide dynamic access to the central cavity from the hydrophobic membrane phase. Phe 203 side chains are shown as stick and mesh representations (yellow). The yellow labels d and e on the structures in **c** show where the fenestrations illustrated in **d** and **e** are located in the complete structure. Note the positions of Phe 203 illustrated in yellow for orientation.

this S5 gating hinge may be involved in both pore-opening[4] and inactivation gating. Our Na$_V$Ab structures do not provide evidence for transition of the S4 segment into a 'relaxed' conformation[29]; however, we do observe repositioning of the entire VSD around the PM (Fig. 4d, e). This movement of the VSD with respect to the PM is likely to be required for entry into the potentially inactivated states represented by the Na$_V$Ab-AB and Na$_V$Ab-CD models. Perhaps pivoting of the VSD around the PM at the S5 gating hinge forces collapse of two S6 segments into an asymmetric dimer-of-dimers conformation at the activation gate. Some gating-modifier toxins have binding determinants in both the VSD and neighbouring PM of voltage-gated ion channels (Supplementary Discussion), including Na$_V$ channels[30], suggesting that a strategy has evolved to trap specific gating intermediates by binding toxins at this interface and lock the VSD and PM in fixed relative positions. This gating movement is a potential target for design of next-generation Na$_V$ blocking drugs that could have increased voltage-dependence and improved subtype selectivity.

Our Na$_V$Ab crystal structures provide insight into conformational changes that may underlie the process of slow inactivation, a conserved property of Na$_V$ channels from bacteria to humans (Supplementary Discussion). During the conformational changes that we propose lead to slow inactivation, the Na$_V$Ab channel dramatically alters its central pore, moving from a nearly square arrangement in the selectivity filter, pore-lining S6 segments, and activation gate, to a strikingly asymmetric arrangement in which the four subunits morph into two pairs of conformations. This structural transition has dramatic consequences for all functional elements of Na$_V$Ab, providing new templates for understanding the slow-inactivation process, the effects of disease mutations, and the complex properties of drugs that block mammalian Na$_V$ channels.

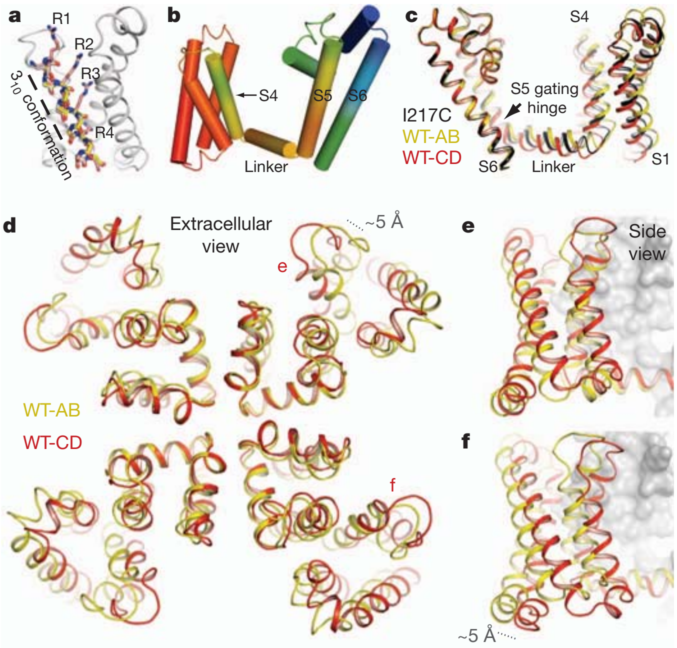


Figure 4 | Structure and coupling of the VSD in Na$_{\rm V}$Ab. a, Superposition of the S4 segment backbone from Na$_{\rm V}$Ab-I217C with WT-AB and WT-CD models (white, yellow and salmon, respectively) shows high structural similarity with similar lengths in 3$_{10}$ conformation. **b**, Crystallographic *b*-factors coloured from low to high (blue to red) along one WT-AB subunit highlight the mobility of the S1–S3 segments relative to the S4 segment. The distribution of *b*-factors further suggests a structural coupling between the S4 segment, the S4–S5 linker, and the S5 segment. **c**, Superposition of the pore domains from Na$_{\rm V}$Ab-I217C (grey), WT-AB (yellow) and WT-CD (red) demonstrates a major hinge point at the base of the S5 segment. **d**, Selectivity filter-based superposition of the Na$_{\rm V}$Ab-AB and Na$_{\rm V}$Ab-CD channels illustrates a rolling motion of the VSDs around the PM during inactivation gating, as viewed from the extracellular side of the membrane. Displacements of the VSDs measure up to ~5 Å. **e**, View from the plane of the membrane highlights the relative vertical and horizontal displacements of the WT Na$_{\rm V}$Ab VSDs during inactivation gating. Chain B, yellow; chain C, red. **f**, Similar view of chain A, yellow; chain D, red. The red labels on the structures in **d** show where the VSDs illustrated in **e** and **f** are located in the complete structure.

METHODS SUMMARY

WT Na$_{\rm V}$Ab was expressed in *T. ni* insect cells, purified using anti-Flag resin and size exclusion chromatography, reconstituted into DMPC:CHAPSO (1,2-dimyristoyl-sn-glycerol-3-phosphocholine: 3-[(3-cholamidopropyl)dimethylammonio]-2-hydroxy-1-propanesulphonate) bicelles, and crystallized over an ammonium sulphate solution containing 0.1 M sodium citrate, pH 4.75. A single anomalous dispersion (SAD) data set from a selenomethionine (SeMet)-substituted protein crystal enabled phase determination and guided initial rigid body refinement protocols. Standard refinement procedures accounting for merohedral twinning were performed. Electrophysiological experiments were carried out on WT Na$_{\rm V}$Ab in *T. ni* cells using standard protocols.

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

Received 7 December 2011; accepted 23 March 2012.

Published online 20 May 2012.

1. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117,** 500–544 (1952).
2. Rudy, B. Slow inactivation of the sodium conductance in squid giant axons. Pronase resistance. *J. Physiol. (Lond.)* **283,** 1–21 (1978).
3. Vilin, Y. Y. & Ruben, P. C. Slow inactivation in voltage-gated sodium channels: molecular substrates and contributions to channelopathies. *Cell Biochem. Biophys.* **35,** 171–190 (2001).
4. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475,** 353–358 (2011).
5. 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).
6. Ren, D. *et al.* A prokaryotic voltage-gated sodium channel. *Science* **294,** 2372–2375 (2001).
7. Catterall, W. A. From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. *Neuron* **26,** 13–25 (2000).
8. Armstrong, C. M., Bezanilla, F. & Rojas, E. Destruction of sodium conductance inactivation in squid axons perfused with pronase. *J. Gen. Physiol.* **62,** 375–391 (1973).
9. West, J. W. *et al.* A cluster of hydrophobic amino acid residues required for fast Na$^+$ channel inactivation. *Proc. Natl Acad. Sci. USA* **89,** 10910–10914 (1992).
10. Eaholtz, G., Scheuer, T. & Catterall, W. A. Restoration of inactivation and block of open sodium channels by an inactivation gate peptide. *Neuron* **12,** 1041–1048 (1994).
11. Nesterenko, V. V., Zygmunt, A. C., Rajamani, S., Belardinelli, L. & Antzelevitch, C. Mechanisms of atrial-selective block of Na$^+$ channels by ranolazine: II. Insights from a mathematical model. *Am. J. Physiol. Heart Circ. Physiol.* **301,** H1615–H1624 (2011).
12. Wang, Y. *et al.* Merging structural motifs of functionalized amino acids and α-aminoamides results in novel anticonvulsant compounds with significant effects on slow and fast inactivation of voltage-gated sodium channels and in the treatment of neuropathic pain. *ACS Chem. Neurosci.* **2,** 317–322 (2011).
13. Pavlov, E. *et al.* The pore, not cytoplasmic domains, underlies inactivation in a prokaryotic sodium channel. *Biophys. J.* **89,** 232–242 (2005).
14. Yue, L., Navarro, B., Ren, D., Ramos, A. & Clapham, D. E. The cation selectivity filter of the bacterial sodium channel, NaChBac. *J. Gen. Physiol.* **120,** 845–853 (2002).
15. Zhao, Y., Yarov-Yarovoy, V., Scheuer, T. & Catterall, W. A. A gating hinge in Na$^+$ channels: a molecular switch for electrical signaling. *Neuron* **41,** 859–865 (2004).
16. Zhao, Y., Scheuer, T. & Catterall, W. A. Reversed voltage-dependent gating of a bacterial sodium channel with proline substitutions in the S6 transmembrane segment. *Proc. Natl Acad. Sci. USA* **101,** 17873–17878 (2004).
17. Irie, K. *et al.* Comparative study of the gating motif and C-type inactivation in prokaryotic voltage-gated sodium channels. *J. Biol. Chem.* **285,** 3685–3694 (2010).
18. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K$^+$ conduction and selectivity. *Science* **280,** 69–77 (1998).
19. Cuello, L. G., Jogini, V., Cortes, D. M. & Perozo, E. Structural mechanism of C-type inactivation in K$^+$ channels. *Nature* **466,** 203–208 (2010).
20. Chen, Y., Yu, F. H., Surmeier, D. J., Scheuer, T. & Catterall, W. A. Neuromodulation of Na$^+$ channel slow inactivation via cAMP-dependent protein kinase and protein kinase C. *Neuron* **49,** 409–420 (2006).
21. Wang, S. Y. & Wang, G. K. A mutation in segment I-S6 alters slow inactivation of sodium channels. *Biophys. J.* **72,** 1633–1640 (1997).
22. Vilin, Y. Y., Fujimoto, E. & Ruben, P. C. A single residue differentiates between human cardiac and skeletal muscle Na$^+$ channel slow inactivation. *Biophys. J.* **80,** 2221–2230 (2001).
23. Zarrabi, T. *et al.* A molecular switch between the outer and the inner vestibules of the voltage-gated Na$^+$ channel. *J. Biol. Chem.* **285,** 39458–39470 (2010).
24. Ragsdale, D. S., McPhee, J. C., Scheuer, T. & Catterall, W. A. Molecular determinants of state-dependent block of sodium channels by local anesthetics. *Science* **265,** 1724–1728 (1994).
25. Yarov-Yarovoy, V. *et al.* Molecular determinants of voltage-dependent gating and binding of pore-blocking drugs in transmembrane segment IIIS6 of the Na$^+$ channel α subunit. *J. Biol. Chem.* **276,** 20–27 (2001).
26. Yarov-Yarovoy, V. *et al.* Role of amino acid residues in transmembrane segments IS6 and IIS6 of the sodium channel α subunit in voltage-dependent gating and drug block. *J. Biol. Chem.* **277,** 35393–35401 (2002).
27. Nau, C. & Wang, G. K. Interactions of local anesthetics with voltage-gated Na$^+$ channels. *J. Membr. Biol.* **201,** 1–8 (2004).
28. Hille, B. Local anesthetics: hydrophilic and hydrophobic pathways for the drug-receptor reaction. *J. Gen. Physiol.* **69,** 497–515 (1977).
29. Villalba-Galea, C. A., Sandtner, W., Starace, D. M. & Bezanilla, F. S4-based voltage sensors have three major conformations. *Proc. Natl Acad. Sci. USA* **105,** 17600–17607 (2008).
30. Wang, J. *et al.* Mapping the receptor site for α-scorpion toxins on a Na$^+$ channel voltage sensor. *Proc. Natl Acad. Sci. USA* **108,** 15426–15431 (2011).

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

Acknowledgements We thank B. Hille for comments on a draft of the manuscript and members of the Zheng and Catterall groups for their insight and support throughout this project. We are grateful to the beamline staff at the Advanced Light Source (BL8.2.1 and BL8.2.2) for their assistance during data collection. J.P. acknowledges support from a Canadian Institutes of Health Research fellowship and the support of N. and E. Payandeh. This work was supported by grants from the National Institutes of Health (R01 NS15751 and U01 NS058039 to W.A.C.) and by the Howard Hughes Medical Institute (N.Z.). We dedicate this work to the memory of our colleague Laura Sheard.

Author Contributions W.A.C. and N.Z. are both senior authors. J.P., N.Z. and W.A.C. conceived and J.P. conducted the protein purification and crystallization experiments. J.P. and N.Z. determined and analysed the structure of Na$_{\rm V}$Ab. T.M.G., T.S. and W.A.C. conceived, T.M.G. conducted, and T.M.G., T.S. and W.A.C. analysed the electrophysiology experiments. All authors contributed to writing the manuscript.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 4EKW. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.Z. (nzheng@uw.edu) and W.A.C. (wcatt@uw.edu).

METHODS

Protein expression and purification. To the best of our knowledge[31], Na$_V$Ab represents the only prokaryotic membrane protein to be overexpressed in a eukaryotic expression system for structural studies to date. Na$_V$Ab was cloned into the pFASTBac-Dual vector preceded by an amino-terminal Flag-tag. Recombinant baculovirus were generated using the Bac-to-Bac system (Invitrogen). *T. ni* insect cells were harvested 72 h post-infection and resuspended in 50 mM Tris pH 8.0, 200 mM NaCl (buffer A) supplemented with protease inhibitors and DNase. Following sonication, digitonin (EMD Biosciences) was added to 1% and solubilization was carried out for 1–2 h at 4° C. Clarified supernatant was agitated with anti-Flag M2-agarose resin (Sigma) pre-equilibrated with buffer B (buffer A supplemented with 0.12% digitonin) for 1–2 h at 4 °C. Flag-resin was washed with ten column volumes of buffer B and eluted with buffer B supplemented with 0.1 mg ml^{-1} Flag peptide. Na$_V$Ab was subsequently passed over a Superdex 200 column (GE Healthcare) in 10 mM Tris pH 8, 100 mM NaCl and 0.12% digitonin and peak fractions were concentrated using a Vivaspin 30K centrifugal device. Selenomethionine (SeMet)-labelled proteins were expressed as previously described[4] and purified as above.

Na$_V$Ab crystallization and data collection. Na$_V$Ab was concentrated to ~20 mg ml^{-1} and reconstituted into DMPC:CHAPSO (Anatrace) bicelles according to standard protocols[32]. The Na$_V$Ab-bicelle preparation was mixed in a 1:1 ratio and set up in a hanging-drop vapour-diffusion format over a well solution containing ~2 M ammonium sulphate, 100 mM sodium citrate pH 4.75. Native and SeMet-labelled proteins crystallized under essentially identical conditions. Crystals were passed through solutions containing 2 M ammonium sulphate, 100 mM sodium citrate pH 4.75 and 28% glucose (wt/v) in increments of ~6% glucose during harvesting. As previously suggested[33], the inclusion of nicotinic acid at saturating concentration in the cryogenic solution was found to prolong the lifetime of our Na$_V$Ab crystals in the X-ray beam. Crystals were plunged into liquid nitrogen and maintained at 100 K during all data collection procedures.

More than 1,500 Na$_V$Ab crystals were screened at the Advanced Light Source (BL8.2.1 and BL8.2.2), but most WT Na$_V$Ab crystals did not diffract beyond 3.5 Å. A single anomalous dispersion (SAD) data set collected near the selenium absorption edge ($\lambda = 0.9792$ Å) from a SeMet-labelled crystal was used to determine initial experimental phases and proved to be our best data set. In addition to nicotinic acid treatment[33], special care was taken to minimize exposure times and orient the Na$_V$Ab crystals in order to maximize data completeness and quality.

Structure determination and refinement. X-ray diffraction data were integrated and scaled with DENZO/SCALEPACK[34] and further processed with the CCP4 package[35]. Initial efforts to determine and refine the Na$_V$Ab-WT structure using our previous Na$_V$Ab-I217C model and various protocols led to R_{free} stalling at ~40%, even after accounting for the perfect merohedral twinning characteristic of the Na$_V$Ab-WT crystals. Fortunately, unbiased experimental phases could be obtained using a 3.2 Å SAD-data set collected from a single SeMet-labelled crystal with the PHENIX software package[36]. The Na$_V$Ab-I217C model was manually placed into this experimentally phased map. Despite limited map quality, approximate boundaries to define rigid bodies in subsequent refinement procedures were apparent (at the S5 gating hinge) and led to an immediate ~6% drop in R_{free}. Complete and partial poly-Ala models were used in combination with SAD phases (in PHENIX[36]) to assist with model re-building and side-chain placement in the program O[37]. Partial models were similarly used to assess and confirm the boundaries of our WT model. Both protein chains in the WT-AB channel extend from amino acid 1 to 219; in the WT-CD channel, chain C extends from amino acid 1 to 213, and chain D extends from amino acid 1 to 217. Only fragmented or weak electron density can be seen beyond these modelled S6 residues. At this point, the SeMet-labelled data set was reprocessed to lower redundancy (to ~3) and used as 'native' data, leading to an overall improvement in data and map quality. Electron density for nine lipids and three water molecules were accounted for at this stage of model building. Application of TLS groups[38], as implemented in REFMAC[39,40], led to a ~1.5% drop in R_{free} and further improvement in map quality. Although examined, NCS restraints were never applied during the refinement procedure due to the asymmetry immediately apparent within the Na$_V$Ab-WT model. Tight geometric restraints were maintained throughout refinement, and the overall geometry of the final Na$_V$Ab-WT model is excellent (Supplementary Table 3). Only Arg 68 in the S2-S3 loop (chain A and C) and Ser 93 in the S3–S4 loop (chain A, C and D) appear as outliers in the Ramachandran plot (5 residues of the 880 modelled).

In order to facilitate comparison between our Na$_V$Ab structures (that is, Fig. 2), and because Na$_V$Ab-I217C was originally refined using CNS software[41] to 2.7 Å resolution[4], the final scaled data set and deposited Na$_V$Ab-I217C coordinates were re-refined using similar REFMAC procedures[39,40] described above for WT-Na$_V$Ab (to effective resolutions of 2.7 Å and 3.2 Å). The R/R_{free}, overall map quality, geometry and root-mean-square deviation of all refined Na$_V$Ab-I217C models are highly comparable.

Structure analysis. The geometry of Na$_V$Ab-WT structural models was assessed using PROCHECK[42]. The pore radius was calculated using standard settings in MOLE software[43] for Supplementary Figs 3c and 8. More detailed representations of the pore were obtained using HOLLOW[44] software for Fig. 3d, e and Supplementary Figs 7a, 9 and 10. Structural alignments were performed using LSQMAN[45] and O[37]. Unless otherwise stated, all figures have been prepared with the WT-AB and WT-CD channels independently aligned onto the selectivity filter (residues Thr 175-Leu 176-Glu 177) of the tetrameric Na$_V$Ab-I217C channel model. All structural figures were prepared with PyMol software[46].

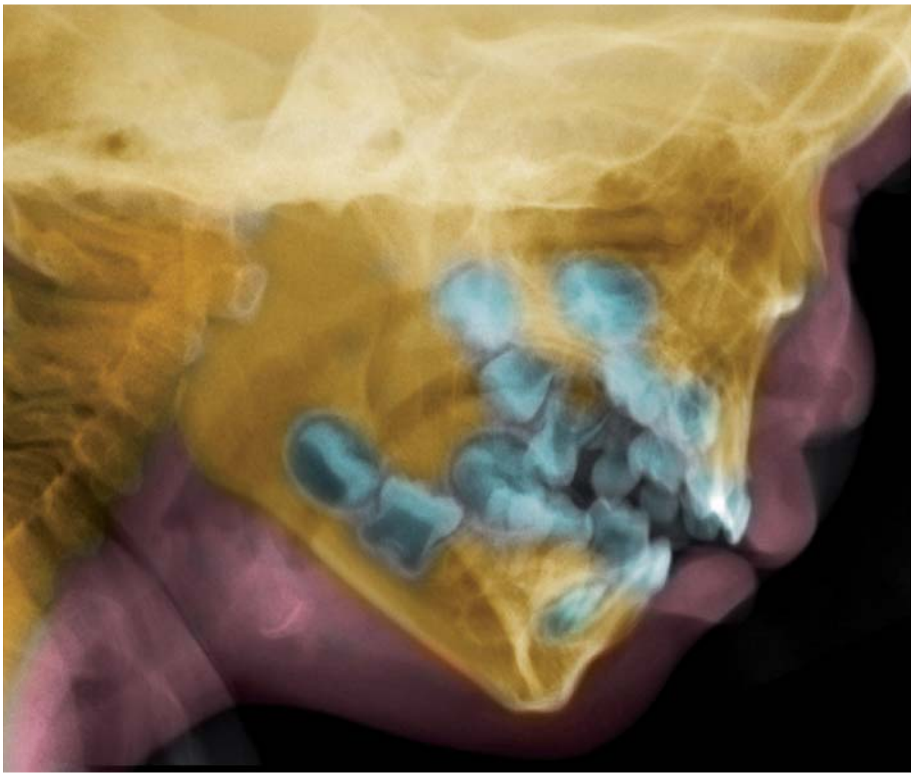
Electrophysiology. Baculovirus containing the WT-Na$_V$Ab construct used for crystallography (that is, containing an N-terminal Flag tag) were used to infect *T. ni* cells. After 24 h, whole cell sodium currents were recorded using an Axopatch 200 amplifier (Molecular Devices) with glass micropipettes (2–5 MΩ). Capacitance was subtracted and series resistance was compensated using internal amplifier circuitry; 80% of series resistance was compensated. The intracellular pipette solution contained (in mM): 35 NaCl, 105 CsF, 10 EGTA, 10 HEPES, pH 7.4 (adjusted with CsOH). The extracellular solution contained (in mM): 140 NaCl, 2 CaCl$_2$, 2 MgCl$_2$, 10 HEPES, pH 7.4 (adjusted with NaOH). Voltage clamp pulses were generated and currents were recorded using Pulse software controlling an Instrutech ITC18 interface (HEKA). Data were analysed using Igor Pro 6.2 (WaveMetrics).

31. Koth, C. M. & Payandeh, J. Strategies for the cloning and expression of membrane proteins. *Adv. Protein Chem. Struct. Biol.* **76,** 43–86 (2009).
32. Faham, S. *et al.* Crystallization of bacteriorhodopsin from bicelle formulations at room temperature. *Protein Sci.* **14,** 836–840 (2005).
33. Kauffmann, B., Weiss, M. S., Lamzin, V. S. & Schmidt, A. How to avoid premature decay of your macromolecular crystal: a quick soak for long life. *Structure* **14,** 1099–1105 (2006).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
35. CCP4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
36. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
37. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
38. Painter, J. & Merritt, E. A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D* **62,** 439–450 (2006).
39. Winn, M. D., Murshudov, G. N. & Papiz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374,** 300–321 (2003).
40. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
41. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
42. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231,** 1049–1067 (1993).
43. Petrek, M., Kosinova, P., Koca, J. & Otyepka, M. MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* **15,** 1357–1363 (2007).
44. Ho, B. K. & Gruswitz, F. HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.* **8,** 49 (2008).
45. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52,** 842–857 (1996).
46. PyMOL Molecular Graphics System. Version 1.2r3pre (Schrödinger LLC).

TURNING POINT Polymer chemist describes her path to independence **p.149**

PROFESSIONAL DEVELOPMENT Designation aims to set standards for scientists **p.149**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**



KALLISTA IMAGES/CORBIS

DENTAL SCIENCE

Oral observatory

Studying the mouth, including the diagnostic potential of saliva, is offering opportunities to explore overall health.

BY ROBERTA KWOK

In 2010, Michael Lau received an e-mail from a recruiter seeking candidates for a position at the University of California, Los Angeles (UCLA). Would he be interested, the recruiter asked, in applying for a postdoc related to salivary diagnostics? Lau, who was finishing his biochemistry and molecular biology PhD at the University of California, Riverside, and considering his career options, was intrigued and surprised. “I had no idea that you could actually detect systemic diseases, and oral diseases, using saliva,” says Lau.

The opening was in the laboratory of David Wong, associate dean of research at the UCLA School of Dentistry. Wong’s group had found in saliva potential biomarkers for oral cancer and the autoimmune disease Sjögren’s syndrome, and was searching for others. With his interest piqued, and keen on the potential for practical diagnostic use, Lau successfully applied for the post.

Lau investigates how tumours in different parts of the body might affect the contents of saliva. In March, he co-authored a paper suggesting that tiny vesicles from breast-cancer cells can affect the protein and RNA contents of vesicles released by salivary-gland cells (C. S. Lau and D. T. W. Wong *PLoS ONE* **7,** e33037; 2012), and researching the possible mechanisms in a mouse model. Working in this field has offered Lau ample opportunities to break ground. At a time when most scientists are focused on other bodily fluids, such as blood and urine, this is “an untapped field,” he says.

Many people assume that dental research is limited to teeth and gums. But dental researchers have long considered the mouth to be an indicator of conditions elsewhere in the body. Saliva contains many of the same molecules found in blood, albeit often at much lower levels, and might offer a non-invasive way to test for diseases in a dentist’s office, in the field or even at home. Researchers have also uncovered possible links between gum disease and disorders such as diabetes and cardiovascular disease, creating the potential for research into whether improving oral health could help in the prevention or management of these conditions.

NON-INVASIVE DIAGNOSIS

Resources for scientists interested in the connections between oral and systemic health have grown over the past decade. The National Institute of Dental and Craniofacial Research (NIDCR) in Bethesda, Maryland, invested US$65.6 million into salivary diagnostics research between 2002 and 2011, and the human salivary proteome — an inventory of proteins secreted by salivary glands — was published in 2008 (P. Denny *et al. J. Proteome Res.* **7,** 1994–2006; 2008). The UK Biobank, a project to build a repository of health and lifestyle data and samples from half a million people, has collected about 130,000 saliva samples, and began accepting research proposals from the international scientific community in March. The Human Microbiome Project supported by the US National Institutes of Health (NIH) has sequenced the genomes of about 130 oral bacteria species. And in 2008, a collaboration between researchers in the United States and the United Kingdom launched the Human Oral Microbiome Database (HOMD), which currently contains the genome sequences of about 270 types of microbes that have been found — some of them only occasionally or during infection — in the oral cavity.

If a scientist were choosing a bodily fluid to investigate for disease biomarkers a decade ago, saliva wouldn’t have made that list, says Wong. But technological advances and improved lab protocols have allowed researchers to conduct large-scale biomarker screens and detect, with more consistency, low concentrations of molecules such as proteins and RNA in saliva. Although job candidates ▶

▶ may be cautious — some sceptics question how saliva could reflect systemic disease — other researchers say that now is a great time to enter this fledgling field. "You'll be a big fish in a small pond," says Daniel Malamud, who specializes in oral diagnosis of infectious diseases and is the director of the HIV/AIDS Research Program at New York University College of Dentistry.

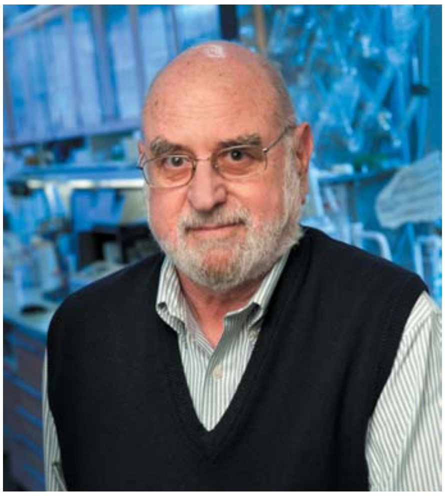
Tests on oral fluids already exist for hormones, illegal drugs and HIV. But the funding from the NIDCR over the past decade has kick-started the field, and teams are now investigating the use of salivary biomarkers for conditions ranging from Alzheimer's disease to heart attacks.

INTRIGUING LINKS

Researchers who work in salivary diagnostics are scattered across dental, medical, biology and engineering departments. PhD graduates in molecular biology, biochemistry, developmental biology and genetics are valued; and statisticians and bioinformaticians are also needed to distinguish significant biomarkers from noise in studies of tens of thousands of genes or proteins. Aims in basic research and device-development often entail technological tasks suited to biomedical engineers, who might develop assays for target molecules, and microfluidics experts, who work out how saliva will flow through a given device. People with experience in electrochemistry, nanotechnology, microfabrication and polymer science also have a role. For example, Fang Wei, a biosensor researcher at the UCLA School of Dentistry, is developing a platform for monitoring the contents of vesicles in saliva in real time. Dental training isn't necessary. Many of Wong's postdoctoral fellows have no dentistry skills at all when they first arrive at the lab, but pick up what they need on the job. Lau taught himself salivary biology, and was helped by lab members who did have dental training.

"I had no idea that you could actually detect systemic diseases, and oral diseases, using saliva."

Investigators should consider a wide range of agencies for funding options. The NIDCR is interested in salivary biomarkers for oral or head and neck cancers, and mechanisms to explain how an organ elsewhere in the body can affect the composition of saliva, says Penny Wung Burgoon, director of the NIDCR's salivary biology and immunology programme. But US researchers interested in systemic disease might have better luck seeking funding at the individual NIH institute that oversees their disease of interest. And last year, the US Defense Advanced Research Projects Agency (DARPA) in Arlington, Virginia, called for research proposals for diagnostic tests that can be easily used in the field to provide on-demand care, such as bioterror-pathogen testing for soldiers. Interest in salivary diagnostic research in Europe lags behind that in the United States, but investigators could apply for grants from funding bodies that cover specific diseases, says Gordon Proctor, a salivary biologist at King's College London Dental Institute.



Daniel Malamud says now is the perfect time to become a oral-diagnostics researcher.

Companies designing and developing tests value biochemists, immunologists or those with skills in molecular testing, says Stephen Lee, executive vice-president and chief science officer at OraSure Technologies in Bethlehem, Pennsylvania. Ronald McGlennen, medical director of OralDNA Labs in Brentwood, Tennessee, expects opportunities to arise for medical technologists — who would typically have a bachelor's degree in medical technology or have completed a relevant training programme — to refine protocols for handling and processing saliva samples. But jobs in this area may be limited because big pharmaceutical and diagnostic companies are waiting for further evidence that tests using saliva are comparable with those that use blood, and that they can meet regulatory standards, says Paul Slowey, chief executive of Oasis Diagnostics in Vancouver, Washington. "A lot of people are sitting on the fence," he says.

In addition to salivary diagnostics, scientists are investigating associations between gum disease and systemic disorders, raising questions about whether improving oral health could help in the prevention of these conditions. Researchers have long known that frequent gum abscesses can be an indicator of diabetes, and it has been suggested that there may be an association between a healthy mouth and improved control of diabetes. A link between poor oral health and cardiovascular disease and with pregnancy complications has also been suggested, but there is no clear evidence of whether gum disease actually contributes to these disorders.

Investigators seeking to establish associations with diseases need to avoid turning projects into fishing expeditions. "You can make a link with ingrown toenails if you want to," says Mark Bartold, director of the Colgate Australian Clinical Dental Research Centre at the University of Adelaide. Researchers need first to consider plausible reasons that a mouth infection or inflammation might affect another disease, he says.

Microbiologists, molecular geneticists and medical researchers could apply their expertise to this area. Researchers have found the DNA of oral bacteria in plaques that build up in blood vessels and in the synovial fluid of joints, raising the possibility that these microbes or their products may help to trigger heart attacks, stroke or prosthetic joint failure. Researchers at the Forsyth Institute in Cambridge, Massachusetts, plan to sequence another 100–200 microbe genomes in the next eight years. With genome data and good research tools, scientists can make connections between oral bacteria and disease more rapidly, says Floyd Dewhirst, an oral microbiologist at the Institute. Researchers can explore how these microbes interact with each other and with humans, including how they might affect systemic diseases.

DENTAL POTENTIAL

Yet Dwayne Lunsford, director of the NIDCR's microbiology programme, warns that because links between oral bacteria and systemic disease are still controversial, early-career investigators should be cautious. If peer-review groups are sceptical, they may score a grant application poorly, he says.

Although some research into salivary diagnostics and the links between oral health and systemic disease takes place in medical schools or conventional biology or engineering departments, biologists should not disregard dental-school faculty positions as a possible career destination. For example, the UCLA School of Dentistry has hired a proteomics researcher to work specifically in salivary diagnostics.

Many dental schools are looking for basic-research scientists, says Chris Overall, a proteomics researcher at the University of British Columbia Faculty of Dentistry in Vancouver, Canada. These institutions can give researchers access to patients, providing them with a better understanding of clinically relevant questions.

Researchers who apply for dental faculty positions may find job more easily than those who aim for basic biology departments. "It's challenging for dental institutions to find people of the calibre that we're looking for," says Laurie McCauley, a dentist and bone biologist at the University of Michigan School of Dentistry in Ann Arbor. For applicants with a track record in fields relevant to dentistry, McCauley calls dental faculty positions "a candidate's market". ■

Roberta Kwok *is a freelance writer based in Burlingame, California.*

TURNING POINT
Rachel O'Reilly

This month, Rachel O'Reilly became a professor at the University of Warwick, UK, at the age of 34. O'Reilly, who uses polymer nanoparticles to mimic natural processes and structures, did a postdoc at IBM Research – Almaden in San Jose, California, and then began a fellowship at the University of Cambridge, UK. But it was a fellowship from the UK's Engineering and Physical Sciences Research Council (EPSRC) that allowed her to really establish her independence.

What has been your most important achievement?
In 2009 I received a career-acceleration fellowship from the EPSRC. The fellowship gave me a salary and a research team, and, because it is a five-year programme, allowed me to be ambitious. My laboratory now has 20 people, and in the past three years we have done a lot. We have published something like 15 papers. The fellowship changed everything. I was able to gain support from three companies, which are supporting students and postdocs in my group. The university was keen to support me further by promoting me to professor.

How did you find the transition from postdoc to fellow?
It was hard. As a postdoc I had people to talk about my ideas with. Then suddenly as a research fellow you get thrown into an office on your own and told to be an academic, build your own lab and research programme. You don't really know what you are doing. I found it very lonely.

Why did you choose to move to Warwick?
The Royal Society's Dorothy Hodgkin fellowship, which I embarked on in Cambridge, was a fantastic opportunity. But I didn't feel that I would be able to grow and develop my research the way I wanted to if I stayed. I think one of the reasons I got the EPSRC fellowship was because I said in the application, 'I want to move to this department and make this big step'. I was able to come to the department at Warwick a little bit more senior, which helped with my confidence. At Warwick, I am in a beautiful materials-science building with amazing facilities. You have to go to the place that is best for you, not where other people think is best.

How did working with Karen Wooley, a successful female chemist, during your postdoc influence your career?
As an undergraduate I was taught by only one female academic, and during my PhD at Imperial College London there were no female professors. When I met Karen, I realized it was the first time I had ever met a female professor. Working with her helped me to become the sort of scientist I wanted to be. Science is full of alpha males who are naturally self-assured. Karen showed me that you don't have to have that outwards confidence. People tell me I am a good collaborator because I am positive and I don't try to take over. A lot of our work is collaborative.

You spent some time working alone before collaborating with other researchers. Why?
You have to be able to develop the technology and the methods yourself first, so that you can then go on to make a contribution as a collaborator. That is probably harder for younger people. You can get approached to collaborate while you are still developing ideas and methods, and then end up unable to produce what you need for the collaboration to work. I held off for a few years to actually make sure I could say, 'Yes, we can make that; yes, we can do that'.

What has been your most difficult challenge?
Having grants and papers rejected are probably the things I've found hardest. You have to be able to accept failure. I still feel that when I am talking about my research ideas I am baring part of my soul. It is a very personal thing. I have learned to realize that a rejection may be the result of not communicating my ideas effectively, and not because I am a failure. I try to use negative criticism as constructively as possible. ■

INTERVIEW BY KATHARINE SANDERSON

PROFESSIONAL DEVELOPMENT
UK scientist standards

The Science Council in London has launched two early-career professional designations — registered scientist and registered science technician — through seven of its member societies. The designations, along with the existing chartered scientist register, aim to set standards that recognize competence across the scientific workforce. Applicants must show relevant scientific understanding and proficiency, and pursue continuing professional development through education and training. The designations should make registrants more attractive to employers, says Nicola Hannam, the council's director of education and skills. "This shows that you've reached a certain level and that you have transferable skills," she says.

PHD CANDIDATES
Defining employees

Doctoral candidates should be considered employees, not students, argues the European Council of Doctoral Candidates and Junior Researchers (Eurodoc) in Brussels. According to the council's 17 May statement, PhD candidates' access to benefits and training is jeopardized if they are not designated as employees. In some cases, PhD trainees cannot access their institution's intranet and research databases, notes Zaza Nadja Lee Hansen, Eurodoc's career development workgroup coordinator. Universities in some European countries consider PhD candidates neither students nor employees, Hansen says, but adds that other nations, including Germany and Austria, recognize them as university employees.

IMMIGRATION
Foreign-talent loss

The UK government should not include foreign students in its immigration cap, says Universities UK (UUK) in London. Doing so limits foreign talent entering the country, argues the organization, which represents 134 universities. International students are included in the current net annual migration of more than 250,000 people, more than double the government's target of 100,000. Students should be left out of the tally, says UUK in a statement in May, because many leave the country after completing their studies. Students are also facing more restrictions, including limits on their stay and a requirement for proof of academic progress.