nature

www.nature.com/nature

Vol 462 | Issue no. 7272 | 26 November 2009

Democratic fallacy

Japan's effort to make budget allocations by public hearing could be good for the country and for science, but not as currently planned.

The Japanese government is attempting something that seems revolutionary, at least in Japan: to let people outside the bureaucracies observe its budgetary decision-making process and, even more radical, to involve the public in that process.

In hearings that started on 11 November and continue this week, working groups of non-government specialists and public representatives are grilling bureaucrats about 220 major government projects and weighing budget requests against their own estimates of their value. The groups have recommended cuts that are sometimes slight, sometimes deep and sometimes total. Just how much impact those recommendations will have on the finance ministry's 2010 budget decision, usually finalized by the end of December, is not yet clear.

The country's highest science policy-making body, the Council for Science and Technology Policy (CSTP), is responsible for the scientific re-evaluation of the projects — a counter or complement to the 'social evaluation' going on in the working groups. The presence of the finance minister on the Government Revitalization Unit, which oversees the working groups, has most observers convinced that the recommendations will carry hefty influence (see *Nature* **462,** 258–259; 2009).

The process is being applied to dozens of science projects and to some of the most basic scientific-funding mechanisms. It is ruffling the feathers of a good proportion of the Japanese scientific community. Yet if implemented intelligently, it could be a positive development. Transparency and public involvement, both of which are weak in Japan, should certainly be encouraged. They help to ensure that decisions are made not through a deal between the head of an institute and bureaucrats, as sometimes happens in Japan, but rather on scientific need.

Before deciding how much weight to give these recommendations, however, the finance ministry should consider some major flaws in the process.

Can a bureaucrat, in one hour, explain a long-term project worth tens of millions of dollars — especially one that has existed for more than a decade, such as the SPring-8 synchrotron — to a group of 19 people, only a few of them scientists and none of them specialists in the fields in question, and hope that they will understand its importance? Can the judgement of that group, which in the case of the synchrotron was to cut the budget by a whopping 30–50%, be considered conclusive? Can such a process adequately assess the repercussions of what these cuts would mean?

With regard to the country's supercomputer project, which would probably be terminated if the group's suggestions are followed, the group could be right to question the nationalistic phrase "fastest in the world" used to justify the programme. But there could be a case for the project to be rethought, reduced and renegotiated. On a smaller scale, it might still be able to benefit science enough to justify itself. Such negotiation, however, will require more scientific input than the new process allows, as even the CSTP is dominated by bureaucrats, generally without any scientific background, and representatives of industry.

> "If the public is to evaluate scientists' way of doing things, then scientists should be given a chance to defend themselves."

Overall, the working groups have provided potentially useful feedback about the perceived social value of these projects. For large public investments such as those under scrutiny, that is an important perspective, and one that scientists often lose sight of. But dialogue needs to follow — if the public is to evaluate scientists' way of doing things, then scientists should be given a chance to defend themselves.

Perhaps, as one researcher described it, this is "a bad dream" and scientists will wake up in January to find that the government gave the recommendations due consideration before moving forwards. Perhaps, as the process settles into place in the years ahead, researchers will come to see it as an acceptable and surmountable challenge to justify their studies. But, as currently posed, the recommendations risk being the final word in a decision-making process that could have disastrous repercussions for decades to come. ■

Conservative vacuum

Britain's main opposition party needs policies for research and for universities.

Last year, Britain's higher-education sector generated around £33 billion (US$55 billion) of the UK gross domestic product, putting it ahead of the aircraft, advertising and pharmaceutical industries, according to figures published earlier this month by economists at the University of Strathclyde, UK. It is astonishing, then, that with all the Conservative Party's rhetoric on how it intends to drag Britain out of recession, it hasn't formulated policy on universities and research. And yet there is a strong possibility that the Conservatives will be leading the country by June next year.

The next government needs to have a long-term vision for the role of science. The Labour government took a bold and welcome move with the 2004 publication of its 10-year science and innovation framework. Any new government must either carry this torch forwards or light a new one. Through the framework, priorities for science and the direction the government wanted to focus on were set in the appropriate long-term context.

At the same time, government spending on research and development (R&D) was ring-fenced and surged to more than £3 billion

© 2009 Macmillan Publishers Limited. All rights reserved

a year. No one expects spending on this scale to continue in the current tight fiscal environment. Indeed, the state of the economy only makes the case stronger for clear long-term policies, which will help to ensure that spending is wise.

Radical thinking on the future direction of universities will be needed as part of this long-term vision. But Labour's more recent higher-education framework, published at the beginning of this month, is less inspiring than its 2004 document. The framework hints that the Labour government would like to see funding for university research further concentrated in the top universities and in key strategic science areas, but the government lacks the confidence to say what it really wants. There are legitimate questions to be asked about whether Britain needs more than 100 universities all chasing after the same limited pot of research funds, and whether this money would be better spent across fewer of those institutions. At least Labour's whiffs of a stance on these issues are better than the deafening silence of the Conservative Party.

Rightly, both Labour and the Conservatives remain committed to the current dual funding system for the foreseeable future. It is through this system, in which universities win a pot of research funding from the government in line with their demonstrated research excellence, and also competitively gain funding from research councils, that universities have the freedom to plan and invest as they see fit.

Creative thinking is also sorely needed to improve the exploitation of Britain's research base. This issue has long been on the agenda and Labour has taken some strides forwards, including establishing the Technology Strategy Board, which provides competitive funding for high-tech businesses. But significant increases in private investment in R&D are still lacking, and are as important as ever in the long term.

To be fair, the Conservatives have signalled a desire to support the high-tech commercial sectors, establishing a task force led by James Dyson, a British inventor, that will report its recommendations on how to improve UK innovation to the party before the general election. But its agenda is a worrying indication of the party's unsophisticated appreciation of the interplay between science and innovation: there is no reference to the importance of continuing to support the research needed to yield the discoveries on which products and services are based.

As Britain's next general election approaches, Labour can point to a strong record of personal commitment to science and science-based enterprise from its leaders and of supportive actions. So far, the Conservatives, in contrast, are a vision-free zone. ■

Getting what you pay for

The US Food and Drug Administration cannot fulfil its mandate without a serious funding boost.

The US Food and Drug Administration (FDA) is in capable new hands. Its commissioner, Margaret Hamburg, a Harvard-trained physician six months into her tenure, brings to the job both a broad experience in science, public health and biosecurity (see page 406) and an ability to handle multiple, simultaneous demands — a skill she displayed as New York City's youngest health commissioner.

For all her abilities, however, Hamburg is struggling to steer an underpowered ship that is loaded to the gunwales. The 103-year-old agency, based in Silver Spring, Maryland, has never before had so many demands placed on it, nor has its budget ever been so constrained relative to its duties. Between 2001 and 2007, for example, the number of US food-manufacturing plants under the FDA's jurisdiction increased from about 51,000 to more than 65,000, yet the number of staff in its foods programme fell from 3,167 to 2,757. At current inspection rates, any given domestic food company faces a less than one-in-four chance of being inspected once in seven years. And that looks frequent compared with the agency's estimated average inspection rate for foreign manufacturers of medium-risk medical devices: once every 27 years.

It is true that the FDA's funding has been boosted since 1993 by user fees paid by drug- and device-makers. In 2009, such fees amounted to nearly 23% of the agency's $2.7-billion budget. But this influx has, paradoxically, taken the pressure off Congress to fund the many mandates it continues to heap on the agency. For instance, the FDA is expected to monitor the accuracy of direct-to-consumer advertisements by drug companies, and the promotional materials they send to physicians. But in 2008, Congress gave enough money to fund only 55 staff for this job. With some 71,000 industry submissions in 2008, those employees can cope with only a small fraction. Similarly, because drug and device fees are dedicated largely to funding reviews for market approval, other functions at the agency, most notably food safety, have received short shrift.

Calls for more cash inevitably raise red flags in this era of ballooning deficits, but the imbalance between the FDA's means and its responsibilities makes the need inescapable. A bipartisan group including six former FDA commissioners and three former heads of the agency's parent department, the Department of Health and Human Services, has publicly urged Congress to boost the agency's appropriations. So have almost all FDA-regulated industries, including the Grocery Manufacturers Association, the Medical Device Manufacturers Association and most major drug companies, as well as dozens of patient groups.

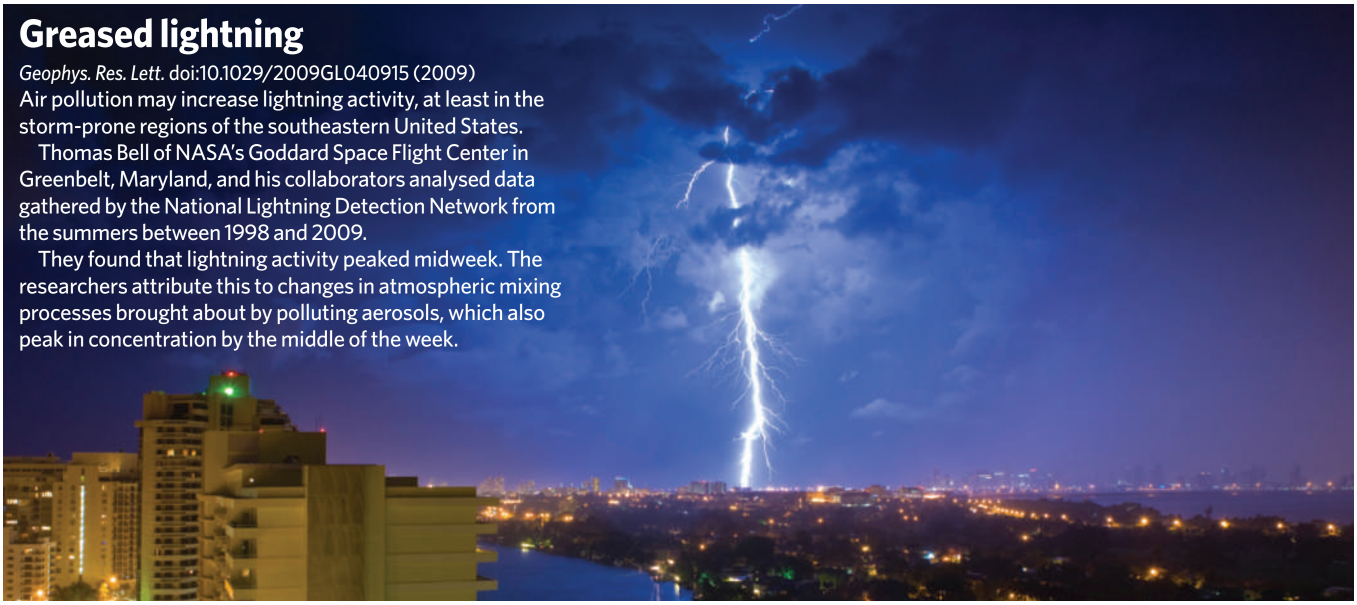
How much extra money is enough? The FDA's science board was asked the same question by Congress in late 2007 after the board issued a scathing report on the agency's eroding scientific capabilities (see *Nature* **450,** 1143; 2007). To set things right, the board concluded last year, Congress would need to add $450 million to the agency's budget in 2010, and $460 million each year between 2011 and 2013.

The administration of President Barack Obama has asked Congress for a further $295 million for the agency in 2010, which would bring its congressional appropriations to $2.3 billion — less than what is needed, according to the science board's estimates, but "a good start", as Hamburg told *Nature* earlier this month. Congress should provide at least that much, and make plans to boost that figure in subsequent years.

Historically, it has taken crises to goad legislators into giving the FDA the money and muscle it needs — a notable example being the poisonous cough syrup that killed more than 100 people in 1937, and led to the 1938 enactment of the Federal Food, Drug, and Cosmetic Act, which still forms the basis of the FDA's authority. Congress shouldn't wait for the next crisis. ■

© 2009 Macmillan Publishers Limited. All rights reserved

RESEARCH HIGHLIGHTS



Greased lightning

Geophys. Res. Lett. doi:10.1029/2009GL040915 (2009)
Air pollution may increase lightning activity, at least in the storm-prone regions of the southeastern United States.

Thomas Bell of NASA's Goddard Space Flight Center in Greenbelt, Maryland, and his collaborators analysed data gathered by the National Lightning Detection Network from the summers between 1998 and 2009.

They found that lightning activity peaked midweek. The researchers attribute this to changes in atmospheric mixing processes brought about by polluting aerosols, which also peak in concentration by the middle of the week.

S. RESNICK/SCIENCE FACTION/CORBIS

PALAEONTOLOGY

Mammoth fungal trail

Science **326,** 1100–1103 (2009)
The collapse of large animal populations, including mammoths and mastodons, in North America may have set off widespread ecosystem changes and occurred before major climatic events that have been put forward as causes of the die-off.

By analysing sediment cores, Jacquelyn Gill of the University of Wisconsin–Madison and her co-workers found that levels of spores of the dung-borne fungus *Sporormiella* began dropping 14,800 years ago, finally falling to a level indicating a megafaunal collapse by 13,700 years ago. By studying levels of fossil pollen and charcoal, the researchers surmised that the extinction of the herbivores led to a larger variety of plant species and a higher frequency of fires.

The team also concludes that neither the cooling period known as the Younger Dryas nor a purported comet impact wiped out the megafauna, as has been hypothesized.

CHEMISTRY

Get into the groove

J. Am. Chem. Soc. doi:10.1021/ja9085512 (2009)
At some point in their careers, most chemists have scratched a beaker to induce crystallization, but how this trick works is still mysterious. Amanda Page and Richard Sear of the University of Surrey in Guildford, UK, have studied the process using computer simulations.

They modelled scratches as wedge-shaped grooves and found that when the angle of the wedge is optimal, the rate of crystal nucleation is orders of magnitude higher in the wedge than on a flat surface, as is seen in experiments.

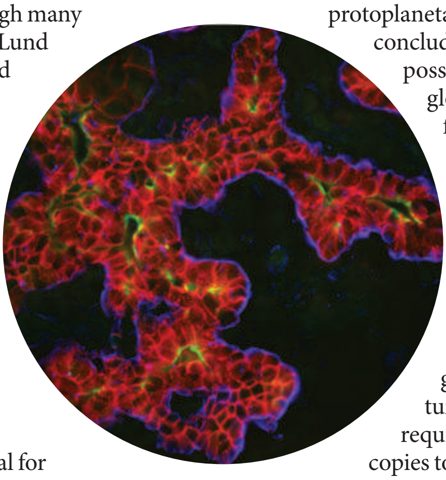
Nucleation is fastest when this angle allows a defect-free piece of crystal to fit perfectly in the wedge. So by tuning a wedge angle to fit a particular crystal polymorph, the creation of this form could be favoured over others, the authors say.

DEVELOPMENTAL BIOLOGY

Down the tube

Cell **139,** 791–801 (2009)
Little is known about how the body forms the tubing that snakes through many organs. Henrik Semb of Lund University in Sweden and his colleagues tracked pancreas development (pictured right) in mice and showed that the same protein signal that controls tube formation also determines how progenitor cells develop to form the surrounding tissue.

The team found that the cell-cycle regulator protein Cdc42 is essential for initiating and maintaining tube development. The protein also helps to create a distinct microenvironment around the forming tubes that controls the specialization of other early cells in the organ.



ELSEVIER

ASTRONOMY

Galaxies aglow

Astrophys. J. **706,** 1020–1035 (2009)
Recent space-based observations of distant galaxies show that many are shining unexpectedly brightly at near-infrared wavelengths.

Erin Mentuch of the University of Toronto in Canada and her colleagues say that disks of material that are the precursors of planets could be responsible for the glow. The team analysed 103 galaxies between 1.9 billion and 5.2 billion parsecs from the Milky Way and found that their light shared similar features to that from nearby stars surrounded by protoplanetary disks. The group concludes that it might be possible to use the excess glow to measure planet-formation rates in distant galaxies.

CANCER BIOLOGY

Dicer blocker

Genes Dev. doi:10.1101/gad.1848209 (2009)
A gene involved in gene silencing is also a tumour suppressor that requires two functional copies to protect against cancer.

Tyler Jacks at the Massachusetts Institute of Technology in Cambridge and his colleagues studied several mouse models of cancer in which one or both copies of *Dicer1* were deleted. This gene encodes a protein

© 2009 Macmillan Publishers Limited. All rights reserved

that is crucial for processing fragments of RNA called microRNAs, which silence other genes.

The authors showed that deleting a single copy of *Dicer1* led to more tumours, lower levels of microRNAs and reduced survival. However, full loss of *Dicer1* blocked tumour formation, presumably because some level of its protein is needed for cell growth or viability.

The team also looked at data for several human cancers. A high proportion of these also had partial, but never complete, loss of the tumour suppressor.

DEVELOPMENTAL BIOLOGY

To be or not to be sperm?

J. Cell Biol. **187,** 513–524 (2009)

When stem cells in the rat testes are at a developmental crossroads, they are able to make their decision independently of their surroundings, according to Zhuoru Wu and her colleagues at the University of Texas Southwestern Medical Center in Dallas.

Progeny of spermatagonial stem cells have two choices: become stem cells or differentiate into sperm. Some models predict that the cells' decision is determined by environmental cues.

But the researchers found that stem cells grown in the same culture medium gave rise to both differentiated cells and more stem cells. Mathematical modelling showed that each cell's decision was biased towards the stem-cell fate 67% of the time.

NEUROSCIENCE

Rats versus mice

J. Neurosci. **29,** 14484–14495 (2009)

Neuroscientists have generally assumed that there is little difference in how adult rats and mice regulate the generation of new brain cells. But a study by Jason Snyder and his colleagues at the National Institute of Mental Health in Bethesda, Maryland, reveals that rats are much more likely to recruit new neurons during learning than mice.

The researchers also showed that adult rat brains contain more young neurons than adult mouse brains, and that these cells mature much faster. In addition, more new neurons are activated in rats during memory tasks. These findings could resolve inconsistencies in the literature about rodent neurogenesis.

MATERIALS

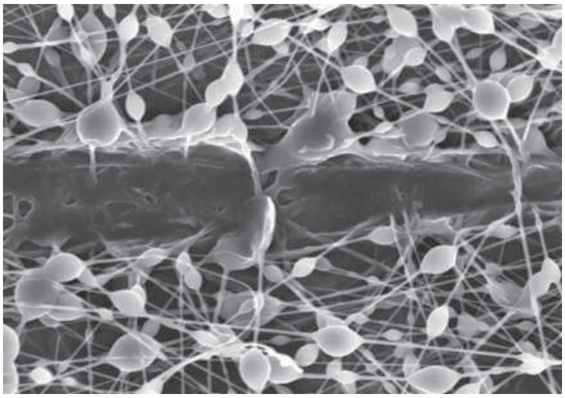
Healed steel

Adv. Mater. doi:10.1002/adma.200902465 (2009)

A polymer coating can enable damaged steel to 'heal' itself, according to Jeong-Ho Park and Paul Braun of the University of Illinois at Urbana-Champaign.

The coating consists of a mat of thin fibres spun using a technique called electrospinning, which creates fibres from a liquid by pulling it through an electric field. Trapped inside pockets along the fibres are bubbles of one or another of two liquid polysiloxane-based healing agents.

When these fibres are electrospun onto a steel surface and that steel surface is cut, the two liquids burst out (pictured below) and mix together in the crack formed during damage. The liquids then polymerize and fill in the gap with a solid substance. In the team's experiments, the healed steel didn't rust for three months, even after initially sitting in salt water for five days.



WILEY-VCH

METEOROLOGY

Can't beat the heat

Geophys. Res. Lett. doi:10.1029/2009GL040736 (2009)

Climate change has left its mark on temperature extremes in the United States. Data collected by weather stations across the country from January 2000 to September 2009 reveal that there have been about twice as many record warm days as record cold days during the period, according to Gerald Meehl of the National Center for Atmospheric Research in Boulder, Colorado, and his colleagues.

The trend is strongest in the western states, where observations roughly match patterns simulated by the centre's climate-system model. For one scenario of future climate change, the model suggests that record warm days across the United States could outnumber record cold days by 20 to 1 by 2050 and by 50 to 1 by the end of the century.

JOURNAL CLUB

Philippe Ciais
Laboratory of Climate and Environmental Sciences, Gif sur Yvette, France

A geoscientist is astounded by Earth's huge frozen carbon deposits.

I believe that the vulnerability of soil carbon to warming is one of the largest sources of uncertainty in the projection of future climate change. If, in a warmer world, bacteria decompose organic soil matter faster, releasing carbon dioxide, this will set up a positive feedback loop, speeding up global warming.

I was stunned to learn, from an article by Charles Tarnocai of Agriculture and Agri-Food Canada in Ottawa and his colleagues, that the global mass of soil carbon needs to be revised upwards by a frightening amount: from the 2,500 billion tonnes of carbon previously accounted for to more than 4,000 billion tonnes (C. Tarnocai *et al. Glob. Biogeochem. Cycles* doi:10.1029/2008GB003327; 2009). This is a result of the previously overlooked presence of vast amounts of peat, Siberian yedoma deposits (organic-rich permafrost) and other frozen carbon stores at high latitudes.

These massive stores deserve special attention because the boreal and arctic regions that house many of them are expected to warm more rapidly than average in the coming decades. Even a small leakage from these stores could cause an explosion in the growth rate of atmospheric CO_2 as well as methane, a potent greenhouse gas emitted by flooded thawed soils.

So what do these findings mean for the role of high latitudes in the Earth system? We need more extensive field observations to monitor the stability of frozen carbon, and studies to measure the decomposition rates of such stores. And we should incorporate these processes into climate models such as those used by the United Nations Intergovernmental Panel on Climate Change. If I had to pick just one new PhD subject right now, exploring this terra incognita of frozen carbon and its impact on climate change would be the one.

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

© 2009 Macmillan Publishers Limited. All rights reserved

NEWS

UK physics council sees grim future

Second financial crisis in two years leaves researchers questioning the council's long-term viability.

Britain's high-energy physicists and astronomers are bracing themselves for budget cuts.

The Science and Technology Facilities Council (STFC), which funds the United Kingdom's astronomy, particle- and nuclear-physics communities, is short by roughly £40 million (US$66 million) in its annual £450-million cash budget. High-energy-physics grants have already been affected, and in a bid to contain costs the council said last week that it would probably withdraw from the multinational Gemini telescope project in 2012.

It is the second such budgetary dilemma for the STFC. The council was formed from the merger of two other councils in April 2007, at the same time as the UK government was undergoing a triennial budget review. "It was a perfect storm" of financial pressures, says Paul Crowther, an astrophysicist at the University of Sheffield. Within months, the newborn STFC announced that it was facing an £80-million budget gap.

The latest problems have made physicists angry once more. "This second crisis makes clear that the STFC is incapable of being run in its current form," argues Brian Foster, a particle physicist at the University of Oxford.

Things are likely to worsen in the coming months. Throughout the autumn, physicists have met to prioritize projects in areas supported by the council. The prioritization will be used to determine how to spend money within the current budget levels, says Terry O'Connor, the STFC's director of communications.

High-energy physicists have already seen their latest grants funded for one year rather than for the standard three to five. That makes it difficult to support postdocs and hire technical staff, says Phil Allport, a particle physicist at the University of Liverpool. And the change comes as scientists are gearing up to study data from the Large Hadron Collider (LHC) near Geneva, Switzerland. "Just as collisions are starting in the LHC, the United Kingdom may not be able to adequately exploit it," Allport says.

GEMINI OBSERVATORY

UK astronomers may lose access to the Gemini North Observatory in Hawaii through cost cutting.

Astronomers are also feeling the pinch. The proposed withdrawal from the Gemini Observatory, a pair of eight-metre telescopes in Hawaii and Chile, echoes a 2007 council announcement that it later cancelled. This time, a review panel of academics made the decision, says Andrew Fabian, an astronomer at the University of Cambridge and president of the Royal Astronomical Society. "We feel that the current package we have with Gemini does not give us a big enough benefit," he says.

Nuclear physicists' dreams for the future are also being affected by the cash problems. UK researchers had hoped that Britain would become a partner in the multinational Facility for Antiproton and Ion Research, now being planned at the GSI Helmholtz Centre for Heavy Ion Research in Darmstadt, Germany. But the budgetary shortfall has left those plans in question, says Bill Gelletly, a nuclear physicist at the University of Surrey in Guildford.

The origins of the shortfall are complex. In 2007, the STFC proposed deep cuts to deal with its financial problems. The UK government responded by allowing the council to borrow money from future years and by providing some support to compensate for currency fluctuations. "The outcry got the attention of people high up," says Crowther, but "it didn't make the problem go away."

Since then, the weakened pound has made it increasingly difficult for the STFC to pay its overseas subscription fees to international facilities such as CERN, the lab that houses the LHC. In addition, repayment of the money borrowed from future years is now due. A 2008 prioritization cut grant renewing by a quarter, but that was not enough to make up the shortfall.

Many who depend on the council for funding can barely contain their anger. "We've had scientific prioritization after scientific prioritization," says George Efstathiou, director of the Kavli Institute for Cosmology at the University of Cambridge. "Why has this organization still not got its programme sorted out?"

O'Connor says the council is doing the best it can at a time when the country's economic future stands at a crossroads. The uncertainty "is not confined to particle physics, nuclear and astronomy", he says, "it's right across the research base — it's right across the economy".

The council is now looking beyond its current three-year spending plan to establish a five-year programme and a ten-year strategy, he says. ■

Geoff Brumfiel

© 2009 Macmillan Publishers Limited. All rights reserved

A. NAYAR

FUTURE DEFORESTATION PREDICTED
Central African nations prepare for Copenhagen.
go.nature.com/eNMuH8

Storm clouds gather over leaked climate e-mails

The online publication of sensitive e-mails and documents from a British climate centre is brewing into one of the scientific controversies of the year, causing dismay among affected institutes and individuals. The tone and content of some of the disclosed correspondence are raising concerns that the leak is damaging the credibility of climate science on the eve of the United Nations climate summit in Copenhagen in December.

The Climatic Research Unit (CRU) at the University of East Anglia (UEA) in Norwich confirmed on 20 November that it had had more than 1,000 e-mails and documents taken from its servers, but it has not yet confirmed how much of the published material is genuine. "This information has been obtained and published without our permission," says Simon Dunford, a spokesman for the UEA, adding that the university will undertake an investigation and has already involved the police.

Many scientists contacted by *Nature* doubt that the leak will have a lasting impact, but climate-sceptic bloggers and mainstream media have been poring over the posted material and discussing its contents. Most consist of routine e-mail exchanges between researchers. But one e-mail in particular, sent by CRU director Phil Jones, has received attention for its use of the word "trick" in a discussion about the presentation of climate data. In a statement, Jones confirmed that the e-mail was genuine and said: "The word 'trick' was used here colloquially as in a clever thing to do. It is ludicrous to suggest that it refers to anything untoward."

"If anyone thinks there's a hint of tweaking the data for non-scientific purposes, they are free to produce an analysis showing that Earth isn't warming," adds Michael Oppenheimer, a climate scientist and policy researcher at Princeton University in New Jersey. "In fact, they have been free to do so for decades and haven't been able to."

"There are apparently lots of people who really do think that global warming is an evil socialist plot, and that many scientists are part of the plot and deliberately faking their science," adds Tom Wigley, a senior scientist at the National Center for Atmospheric Research in Boulder, Colorado, and former director of CRU.

Alleged e-mails containing critical remarks about other climate scientists are merely proof of lively debate in the community, adds Gavin Schmidt, a climate researcher with NASA's Goddard Institute for Space Studies in New York City.

"There are apparently lots of people who really do think that global warming is an evil socialist plot."

The title of the uploaded file containing the leaked e-mails — 'FOIA.zip' — has led to speculation that the affair may be linked to the deluge of requests for raw climate data that have recently been made under the UK Freedom of Information Act to Jones (see *Nature* 460, 787; 2009). The source of many of those requests is Steve McIntyre, the editor of Climate Audit, a blog that investigates the statistical methods used in climate science. "I don't have any information on who was responsible," McIntyre told *Nature*.

Nevertheless, e-mails allegedly sent by Jones seem to illustrate his reluctance to comply with these requests. "All scientists have the right to request your data and to try to falsify your results," says Hans von Storch, director of the Institute for Coastal Research in Geesthacht, Germany. "I very much respect Jones as a scientist, but he should be aware that his behaviour is beginning to damage our discipline." In a statement, the UEA said: "The raw climate data which has been requested belongs to meteorological services around the globe and restrictions are in place which means that we are not in a position to release them. We are asking each service for their consent for their data to be published in future."

However, von Storch believes that, at least until the affair is resolved, Jones should cease reviewing climate science for the Intergovernmental Panel on Climate Change. ■

Quirin Schiermeier

Indian neutrino lab site rejected

India's particle physicists have lost their battle to build a neutrino laboratory — one of the country's biggest physics projects — under the Nilgiri hills at Singara in the state of Tamil Nadu. The government has upheld conservationists' view that its construction would endanger wildlife in the Nilgiri Biosphere Reserve (NBR), an important tiger and elephant habitat.

The 6.8-billion rupee (US$150 million) India-based Neutrino Observatory (INO) has been mired in environmental controversy since 2006, but physicists were hoping it would be resolved in their favour (see *Nature* 461, 459; 2009). However, on 20 November India's minister of environment and forests, Jairam Ramesh, informed the scientists that they should not proceed at Singara.

Ramesh wrote that he was acting on a "large number of reports" received against the proposed site and the "very weighty reasons" put forward by Rajesh Gopal, head of forestry in his ministry. Ramesh has suggested the project consider instead a site near Suruliyar, also in Tamil Nadu, that does not pose Singara-type problems.

"Everybody in the INO project is disappointed," says project spokesman Naba Mondal, a physicist at the Tata Institute of Fundamental Research in Mumbai. Project scientists had already considered and rejected the potential site at Suruliyar because there were less available data on the characteristics of the rock that would need to be blasted out to create a cavern to host the neutrino detector. "Preparing a new site means a further delay of one year to a project that has already lost four years due to environmental activism," he says.

"A new site means a further delay of one year to a project that has already lost four years."

Conservationists are pleased, however. "We are indeed relieved," says Tarsh Thekaekara, coordinator of the NBR Alliance, the group that spearheaded the campaign against building the neutrino observatory at Singara. The proposed Suruliyar site is also close to the Periyar tiger reserve, although not in a wildlife corridor as the Singara site is.

Thekaekara says that environmentalists near Suruliyar may decide to challenge the new proposal. "We only represent organizations in Nilgiri," he says. "It may happen that some of the members also active in [Suruliyar] will protest if there is a serious threat to nature." Mondal says that work at the new site will start only after all government clearances are in place. ■

K. S. Jayaraman

© 2009 Macmillan Publishers Limited. All rights reserved

Flu-virus prevalence comes under scrutiny

Projects to monitor antibodies seek true extent of H1N1 infection.

Researchers are turning their attention to one of the great unknowns about the ongoing H1N1 influenza pandemic: how many people have been, and are being, infected. The first surveys to monitor for antibodies to the virus are now getting under way, belatedly in some countries such as the United States. The findings could substantially change much of what epidemiologists know about the current pandemic.

"I'm very struck that we don't have even an idea of the magnitude of infection," says Xavier de Lamballerie, a virologist at the University of the Mediterranean Aix-Marseille II in Marseilles, France. "Epidemiologists haven't a clue if it is 5%, 10% or 20% of the population."

Gathering that information is crucial for improving estimates of pandemic spread, severity and mortality, and informing policies such as how to distribute vaccines and antiviral drugs.

Laboratory-confirmed cases of pandemic H1N1 underestimate the true prevalence by several orders of magnitude, as only a tiny fraction of cases can be tested. Instead, public-health agencies such as the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, use proxy measures, including data on the frequency of people reporting influenza-like symptoms to their doctors.

But testing blood samples for antibodies to pandemic H1N1 is the only definitive way to establish how many people have been exposed to the virus and to begin to estimate how this is changing over time. "Arguably, these data are one of the most important quantities," says Marc Lipsitch, an epidemiologist at the Harvard School of Public Health in Boston, Massachusetts, who is working with the CDC on pandemic flu.

Britain, France and Vietnam are among those farthest ahead with such studies for H1N1. In Britain, Andrew Hayward heads Flu-Watch, which for the pandemic has scaled up its seasonal-flu work to a £2.1-million (US$3.5-million) study co-funded by the Medical Research Council and the Wellcome Trust. Instead of its usual 650–850 subjects, the group will investigate 10,000 subjects, including 2,500 in a serology study to look for antibodies against H1N1 in their blood serum.

The UK Health Protection Agency has also launched a £180,000 study. This draws on blood samples collected from hospital patients for other purposes — so what it lacks in terms of targeting well-characterized groups of individuals, it makes up for in speed by using

Japan sets sights on solar power from space

Japanese scientists are once again eyeing an off-world approach to alternative energy — collecting solar energy from satellites in orbit and beaming it down to Earth.

A space-based solar-power satellite — which could gather energy without having to worry about clouds or night-time — has been a dream for decades in both the United States and Japan. But the costs of developing it has meant that support has waxed and waned over the years. Now, however, Japan has a new sense of mission. In June, it released a national space plan calling for a programme to "lead the world in space-based solar power". And earlier this month, scientists, engineers and policy-makers met at Kyoto University to lay out development plans.

The government's commitment "is definitely a milestone and has given tremendous excitement to solar-power satellite researchers", says Hiroshi Matsumoto, a radio scientist and president of Kyoto University.

Researchers are hoping to launch a full-scale system by 2030, but costs need to come down dramatically for it to be economically viable.

"I'm 100% confident this will happen. We need another stable power source."

Few doubt that the project is technically possible. The well-understood process starts with collecting solar energy with photovoltaic cells, transferring that energy to antennas that transmit microwaves, then receiving those microwaves with a 'rectifying antenna' that converts them to electricity. As early as 1975, scientists at the Jet Propulsion Laboratory in Pasadena, California, transferred energy by means of microwaves over a distance of 1.54 kilometres. And in May last year, scientists beamed power over a distance of 148 kilometres, between two Hawaiian islands.

Japan has been investigating solar-power satellites since the 1980s. In 1983 and again in 1993, Matsumoto, working with Kobe University's Nobuyuki Kaya, launched rockets into the ionosphere to investigate what happens to microwaves as they travel through space (H. Matsumoto *Radio Sci. Bull.* 273, 11–35; 1995). In March this year, a group from Kyoto University became the first to use microwaves to send power from the air to the ground when they charged a mobile phone with microwaves transmitted from a blimp-like airship hovering some 30 metres above the ground.

Current scale-up plans call for a series of tests, each with an increasingly larger capacity for power transmission. First, Japan aims to demonstrate ground-based transmission in the kilowatt range, then space-based kilowatt transmission using Japan's Kibo module on the International Space Station or small satellites. By 2020, researchers hope to have a prototype satellite that can transmit in the range of hundreds of kilowatts, and by 2030 a satellite that can transmit a gigawatt. As currently envisioned, the system to launch in 2030 would be a 2-kilometre-wide array of

© 2009 Macmillan Publishers Limited. All rights reserved

HAVE YOUR SAY
Comment on any of our News stories, online.
www.nature.com/news

C. ANAND/AP

Antibody analysis of blood samples is the only way to accurately track the evolving flu pandemic.

existing samples. Survey leaders have collected 1,403 blood samples from before the first pandemic wave, and 1,954 taken in August and September, across all age groups from eight regions in England. "I believe England is the first to obtain such seroincidence data," says Elizabeth Miller, the study's lead researcher.

She declined to comment on the results because they are under review at a journal, but says the data will provide "insight into the extent to which surveillance has underestimated the true burden of infection due to the occurrence of mild or asymptomatic infections". Hayward is also writing up preliminary FluWatch results gathered during Britain's first pandemic wave, and these too are likely to show a very different picture from that provided by surveillance data alone.

In France, the second wave of the pandemic has only just begun, buying precious time for the €300,000 (US$450,000) SéroGrippeHebdo ('Sero Flu Weekly') study, led by the French School of Public Health in Rennes and Paris. This project is recruiting 30,000 pregnant women, and gains speed by piggybacking on existing infrastructure for routine blood sampling of this group. It will publish serology data from 800 of the women weekly in real time, beginning this week. Already, says de Lamballerie, whose lab is doing the testing, "we've got a great baseline — no higher than 5% or so of the study population has already been infected".

"Arguably, these data are one of the most important quantities."

In Vietnam, Peter Horby, a researcher at the National Institute for Infectious and Tropical Diseases in Hanoi, is part of a project testing leftover serum from haematology and biochemistry labs in nine provincial hospitals. He is also switching a seasonal-flu study of 908 people in 269 households in the northern Ha Nam province to study the serology of pandemic flu. Horby says it should yield good data on the true epidemiology of H1N1 in Vietnam.

Other projects are also just getting started. On top of the survey of pregnant women, Antoine Flahault, dean of the French School of Public Health, is seeking support for a French-led international study called CoPanFlu. This would see each partner fully profile 1,000 households during the pandemic, including serology testing at six-monthly intervals over two years.

"One shouldn't underestimate the difficulty in getting these sorts of studies off the ground," says Hayward.

The United States was less prescient, it seems. Academic groups there are still in the process of applying for funds for such surveys, and *Nature* has also learned that the CDC is about to announce a serology study across ten states. During pandemic planning before the current virus arose, the United States extensively discussed the need for such studies but decisions weren't taken, says Donald Burke, dean of the Graduate School of Public Health at the University of Pittsburgh in Pennsylvania: "None of these very important studies were in place when we knew there was going to be a pandemic. It's unfortunate." ■

Declan Butler

solar cells with an array of 1 billion transmitting antennas — each measuring 5–10 centimetres across — on the side facing Earth.

The goal is to make satellites for under ¥1 trillion (US$11 billion) each; it currently costs 100 times that. "It's exciting, but there are many problems to overcome," says Naoki Shinohara of Kyoto University. For one thing, transmission efficiency must rise to 75%, he says; the airship experiment achieved just 40% efficiency, although the technology it uses differs from what a satellite would use.

Rocket launches will also need to be cut to a hundredth of their current cost; options such as reusable rockets are being considered, according to Susumu Sasaki of the Japan Aerospace Exploration Agency (JAXA). At this month's meeting, Tokyo University's Kimiya Komurasaki discussed how a remote microwave source could power rockets. That would reduce the amount of propellant they need to carry and, in theory, mean that rockets used to build a solar-power satellite could carry more antennas and solar cells.

Experiments this year with an airship transmitted enough power to charge a mobile phone on the ground.

RESEARCH INSTITUTE FOR SUSTAINABLE HUMANOSPHERE, KYOTO UNIV.

Matsumoto estimates that it will take ¥2 billion to ¥3 billion to demonstrate solar-power satellite technology on the ground, and ¥10 billion to ¥50 billion to demonstrate it in orbit.

The nation's space plan calls for an "all-Japan" effort to prepare for space-based demonstrations within three years. And as research budgets have been tight in many areas (see *Nature* 462, 258–259; 2009), the industry and science ministries have more than doubled their budget requests for solar-power satellite-related programmes, to nearly ¥1.4 billion. JAXA has pressed for a doubling of its budget for space-based solar power, from ¥250 million to ¥500 million.

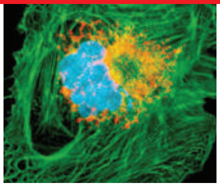
"I'm 100% confident this [technology] will happen," says Shinohara. Unlike wind or Earth-based solar, solar-power satellites in space can gather energy 24 hours a day to provide a reliable source of alternative energy. "We need another stable power source," he says.

Japan looks likely to lead the way, as interest in the United States has waned, says John Mankins, who led the space solar-power programme at NASA. Most efforts in the United States are now in private companies or non-profit organizations. In April, Solaren, a company based in Manhattan Beach, California, signed a contract with San Francisco-based Pacific Gas and Electric to produce 200 megawatts of energy from a solar-power satellite starting in 2016. But Mankins, who co-founded and works at Managed Energy Technologies in Ashburn, Virginia, calls that goal "extremely challenging".

Japan's effort, he says, may lead the way: "The Japanese plan is quite well formulated." ■

David Cyranoski

© 2009 Macmillan Publishers Limited. All rights reserved



NIKON MICROSCOPYU

POSITIVE INTERFERENCE
RNAi boosts production of biological drugs.
go.nature.com/qvalaR

Icelandic genomics firm goes bankrupt

deCODE's demise leaves fate of its valuable genetic database unclear.

After struggling financially for years, the genomics company deCODE, based in Reykjavik, Iceland, filed for bankruptcy on 16 November. The question now is whether other companies looking to commercialize genomics will follow the same path.

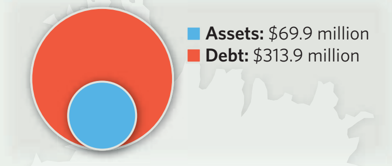
Scientists are already lamenting the prospect of losing deCODE's vast database of genetic and medical information, which includes much of Iceland's population (see graphic). "There is no researcher doing genetics who would not want access to a certain amount of those data," says Manolis Dermitzakis, a geneticist at the University of Geneva Medical School in Switzerland. "It would be a huge loss if the data disappear."

Analysts blame business decisions and a mountain of debt for bringing deCODE down. The worldwide financial crisis also played a part; US$30 million of deCODE's money was managed by Lehman Brothers before Lehman's collapse. But deCODE's chief executive, Kári Stefánsson, says other factors were more important. "The Icelandic economy is in a bad shape, but that is not the reason why we are where we are," he says. "We probably founded the company a few years too early."

deCODE was founded in 1996 to find the genetic roots of common diseases such as schizophrenia, heart disease and cancer, and was wildly successful at this task. "deCODE has carried out many seminal studies, which any medical school would have been proud to see published," says David Goldstein, a geneticist at Duke University in Durham, North Carolina.

But the business of turning genetic discoveries into cash has long been difficult, and many such firms have converted themselves into drug-discovery operations. Unfortunately for deCODE, it could not develop drugs quickly enough to satisfy investors. Stefánsson says that may be because it began its work before cheap, standardized technologies for genome analysis, such as single nucleotide polymorphism (SNP) chips, became widely available.

DECODE BY THE NUMBERS



Proposed sale price: $14 million
Icelanders represented in its database: 140,000
Research publications: 102
Diagnostic scans marketed: 6
Drugs marketed: 0

Kári Stefánsson led a major genetic database.

I. LANGSDON/REUTERS

Some other researchers, however, say that deCODE's scientific approach is to blame. The company worked to mine genetic data for common variations linked to disease through genome-wide association studies (GWAS), and some experts note that these studies have turned up only a small fraction of the variation that causes disease. "The translation to commercial value is just not very direct," says Goldstein, "in part because it is now clear that GWAS is not the tool of choice for unlocking the genetics of most common diseases."

"We probably founded the company a few years too early." — Kári Stefánsson

Stefánsson and others disagree. "Five years ago there was no direct evidence that GWAS studies would provide us with completely new genes or pathways causing disease, and now there are close to 200 findings replicated," says Leena Peltonen, head of human genetics at the Wellcome Trust Sanger Institute, near Cambridge, UK. Stefánsson points to deCODE products such as a cardiovascular test that can detect a nearly twofold increase in the risk of heart attack.

And other companies are beginning to make inroads in drug development guided by genetics. Human Genome Sciences of Rockville, Maryland, transformed itself into a drug-discovery company some years ago after failing to convince other companies to subscribe to its genetic database. This year, the company announced the completion of promising clinical trials for a drug to treat the autoimmune disease lupus.

Many companies are vying to create and own genetic content in different ways, "so in the long run I don't think [deCODE's] problems are systemic", says Isaac Ro, an analyst at Leerink Swann in New York.

But companies that focus exclusively on personal genomics services, such as the one sold by deCODE as deCODEme, might find themselves in more trouble, Ro says. The services are not seen as a medical necessity, diminishing their appeal, particularly in difficult economic times. This year the personal-genomics company 23andMe, based in Mountain View, California, announced two rounds of lay-offs, lost one of its two co-founders and announced a series of product and price restructurings.

"There's no clinical trial supporting the value of these results, so it's really recreational genomics," Ro says. Large academic centres, not consumers, will find value in the rich genetic databases; 23andMe has tried to move into the research market, but because its data come from a self-selected customer population their value is limited, he says.

By contrast, deCODE's database incorporates a wide range of valuable information coupled with biological samples.

deCODE intends to sell most of its assets, including its drug-discovery and development services and the unit that conducts its genetic research, to Saga Investments, a US venture-capital-backed company, unless a better offer is made. The database and biological samples themselves cannot be sold, Stefánsson says, because of legal restrictions on their use. He says that the Wellcome Trust in Britain had approached deCODE to try to fund a non-profit institute to manage the database in Iceland, but was unable to do so.

"The database will never be managed by a foreign organization," he says. "The data are sensitive. We are a proud nation, and the data are not for others to manage." ■

Erika Check Hayden

© 2009 Macmillan Publishers Limited. All rights reserved

GOT A NEWS TIP?
Send any article ideas for *Nature*'s News section to
newstips@nature.com

K. CAMPBELL/GETTY

Famous brain set to go under the knife

Slices from the brain of H.M., a key patient in pioneering memory studies, will be immortalized online.

Neuroanatomist Jacopo Annese will next week begin slicing one of the most precious pieces of tissue in the history of neuroscience: the brain of the famous amnesiac Henry Gustav Molaison, more commonly known by his initials, H.M.

In 1953, Molaison underwent an experimental operation that aimed to treat his severe epilepsy, during which the surgeon removed a part of his brain, including a large chunk of the hippocampus. For the rest of his life, Molaison's ability to form new memories was severely impaired, although he could easily recall memories from his childhood. Hundreds of ensuing psychological studies on him have yielded invaluable insight into memory formation, the separation of different cognitive functions, and the relationship between brain structure and function.

On 2 December, exactly one year after Molaison's death, Annese, of the University of California, San Diego, will begin dividing the brain into roughly 2,400 slices, each thinner than a human hair, and digitizing them. Annese hopes that Molaison's brain will become the first of many in a digital human-brain library at the university.

Annese is one of the few people with the sophisticated equipment needed to slice whole human brains, which is how he came by Molaison's brain. Most labs cut human brains into blocks before slicing them — the fate that befell Albert Einstein's brain.

Annese will mount and stain about every 30th slice for cell nuclei and projections, which will allow him to map the cellular architecture in three dimensions. The remaining slices will be available to the neuroscience community, with researchers able to view the particular slice they want to study before requesting it.

Because demand for certain regions of Molaison's brain is likely to outstrip supply, Annese plans to create a self-moderating forum through which scientists can discuss how to distribute the pieces. "Everyone will be aware that a certain person has this piece of tissue," he says. "It will be their own prerogative for their reputation to do something valuable with it."

ANNESE LAB/UNIV. CALIFORNIA, SAN DIEGO

Jacopo Annese hopes to make thousands of slides and create a three-dimensional brain map.

S. CORKIN/WYLIE AGENCY

"After an experimental operation to treat his epilepsy, Molaison's ability to form new memories was severely impaired."

The tissue will give neuroscientists who worked with Molaison a chance to test hypotheses through detailed anatomical study.

"People can finally look in detail at the precise border of the [brain] lesion, and what the consequences were," says Sandra Witelson at the Michael G. DeGroote School of Medicine at McMaster University in Hamilton, Canada, who studied Einstein's brain and oversees the world's largest repository of normal human brains.

If certain hippocampal structures remained intact, this might explain why Molaison could sometimes recall the names of people who became famous after his injury, such as presidents John F. Kennedy and Ronald Reagan. Studies may also show whether Molaison developed Alzheimer's disease and could hint at whether other brain areas became stronger to compensate for his injury.

Suzanne Corkin, a behavioural neuroscientist at the Massachusetts Institute of Technology in Cambridge who worked with Molaison for 46 years, is "ecstatic" about Annese's project. She says she is looking forward to the chance for others to test their hypotheses against the physical material from Molaison.

But even when accompanied by a mountain of clinical analyses, there is only so much that can be learned from a single brain. "There are still some lingering questions about H.M., but I'd expect that most of what we can learn from him we've already learned," says Adam Gazzaley, a cognitive neuroscientist at the University of California, San Francisco.

To help extend the findings from Molaison's brain, Annese wants to slice, digitize and disseminate many other brains, including some from other people with memory problems. He has already processed the brain of a patient who had some similar cognitive deficits to Molaison after his hippocampal area was damaged by a viral infection.

Annese hopes that the publicity surrounding Molaison's case will bring attention and funding to the brain-library initiative, helping him to overcome the biggest hurdle: storing the data. To make it possible to zoom in on one slide of brain from life size to the cellular level requires 20,000 images, he says — which will require 1–10 terabytes per slide.

He wants to provide 500 or so such slides for each of hundreds of brains. ■

Lizzie Buchen

© 2009 Macmillan Publishers Limited. All rights reserved

Mexico's transgenic maize under fire

Experimental planting scheme has insufficient controls to prevent gene flow to native crops, critcs say.

Mexico doesn't have an adequate system to monitor or protect natural maize (corn) varieties from transgenes, say prominent scientists concerned about the experimental planting of genetically modified crops.

In the past month, Monsanto and Dow AgriSciences have received government permission to plant transgenic maize across 24 plots, covering a total of nearly 13 hectares, in the northern states of Sonora, Sinaloa, Chihuahua, Coahuila and Tamaulipas. The planting of transgenic maize had been prohibited for 11 years in Mexico, where maize was first domesticated.

The experiments are meant to test hardier varieties of the crop, and federal officials say that they are implementing controls to prevent gene flow.

Ariel Álvarez Morales, executive secretary of the Mexican Inter-Secretarial Commission on Biosafety of Genetically Modified Organisms, described the experimental planting as a compliance trial to see how the companies and the plants perform. “We want to see how the planting will work in these conditions,” he says. Plots will be less than half a hectare in area, seed-planting will occur at different times from that of natural varieties, and farmers will be surveyed about the effect on native maize.

In Sonora, where Monsanto has begun planting, transgenic maize is kept 500 metres away from conventional maize fields, says Eduardo Perez Pico, the firm's chief of research and regulatory affairs for the Latin American region.

However, nearly 2,000 scientists have signed a petition to block the experiments. “There is no way to stop gene flow to the native crops,” says signatory Montgomery Slatkin, a geneticist at the University of California, Berkeley. Greenpeace and other groups filed a legal challenge, which the government has rejected.

Activists question Mexico's transgenic maize.

M. TOVAR/AP PHOTO

“If Mexico experimentally plants transgenic maize, it should be done with ideal experiments and a great capacity to monitor them — but we don't have either,” adds José Sarukhán Kermez, a Mexican biologist who has served in top ministerial posts and is a former rector of the Autonomous National University of Mexico (UNAM) in Mexico City.

One facet of the debate surrounds the US firm being used by the Mexican government to train and equip staff at two reference labs for transgene testing in Mexico City. The firm, Genetic ID, is a spin-off by John Fagan of the Maharishi University of Management in Fairfield, Iowa, which favours organic crops and transcendental meditation.

Álvarez Morales says the firm was chosen because of its widely known analytical techniques. But geneticist Elena Alvarez-Buylla, of UNAM's Institute of Ecology in Mexico City, questions whether the company's methods are sensitive enough to detect transgenes after several generations of plant growth. Earlier this year, her group reported that Genetic ID failed to detect transgenes in blinded samples[1]. Genetic ID responded that Alvarez-Buylla's results were due to sample contamination[2], which she challenged[3].

Jay Reichman, an authority on transgenic testing with the US Environmental Protection Agency in Corvallis, Oregon, says that “overall the combined evidence suggests” that at least two transgenes “were present within the plant tissues” in question. In particular, Reichman noted that Alvarez-Buylla showed newly grown test plants believed to harbour transgenes were resistant to herbicide, indicating that they bore transgenes just like commercial seeds modified to be herbicide resistant.

Fagan disputes the criticism. Still, he too is against transgenic planting, citing the potential contamination of native maize: “It is very, very unacceptable.” ■

Rex Dalton

1. Piñeyro-Nelson, A. *et al. Mol. Ecol.* **18,** 750–761 (2009).
2. Schoel, B. & Fagan, J. *Mol. Ecol.* **18,** 4143–4144 (2009).
3. Piñeyro-Nelson, A. *et al. Mol. Ecol.* **18,** 4145–4150 (2009).

Maize genome sequenced

Geneticists have sequenced the genome of maize (corn), one of the world's most widely grown grains, a feat that should accelerate efforts to develop improved crop varieties to meet the world's growing hunger for food, animal feed and fuel.

The genome “is really a tremendous resource”, says John Doebley, a maize geneticist at the University of Wisconsin–Madison. “It gives us a tool for mapping genes that we didn't have.”

The four-year, US$31-million project to sequence maize (*Zea mays*) was led by a US-based consortium of researchers who decoded the genome of an inbred line of maize called B73, an important commercial crop variety. The 2.3-billion-base sequence — the largest genetic blueprint yet worked out for any plant species — includes more than 32,000 protein-coding genes spread across maize's 10 chromosomes. Sections of DNA called transposable elements, which can move around the genome and cause mutations, are the most abundant parts of the sequence.

“What we have here is a crucial part of the instruction manual for how you breed a better corn plant,” says project leader Richard Wilson, director of the Genome Center at Washington University in St Louis, Missouri.

The genome was published last week in *Science* (P. S. Schnable *et al. Science* **326,** 1112–1115; 2009), together with 14 companion analyses in *Science* and other journals.

Elie Dolgin

For more, see go.nature.com/UXHHw4

Correction

In the News Feature ‘The Disappearing Nutrient’ (*Nature* **461,** 716–718; 2009), Amit Roy was misquoted as saying there was a possibility of “market manipulation” with phosphates. His full quote was: “The biggest challenge is that concentration of supply is only in a few hands and there is the possibility of manipulation of supply, demand and prices.” Roy did not mean to imply that there is a possibility of market collusion.

© 2009 Macmillan Publishers Limited. All rights reserved

A TOUGH TONIC

The new head of the US Food and Drug Administration has inherited an agency battered by crises. **Meredith Wadman** asks whether Peggy Hamburg can concoct a cure.

When Margaret 'Peggy' Hamburg was appointed health commissioner of New York City in 1992, the 36-year-old physician and former researcher took a job that few people wanted. A tuberculosis epidemic was roaring out of control in the city, facilitated by rising rates of HIV/AIDS. Rodents plagued city streets and a demoralized health department was facing draconian cuts in its budget as the city government fought a ballooning deficit. Hamburg's predecessor had abruptly resigned early in his tenure, after antagonizing the city's vocal HIV/AIDS community.

Within three years Hamburg had turned the department around. She rescued it from proposed cuts that would have crippled its public-health lab and its immunization and school-health programmes. She won extra money for rodent control. She launched a needle-exchange programme to combat AIDS and she reversed the tuberculosis epidemic, with a 21% drop in new cases[1] over the course of a programme the United Nations later cited as exemplary.

Hamburg, who became commissioner of the US Food and Drug Administration (FDA) in May, is now hoping to repeat that success at a beleaguered federal agency that has endured a string of recent crises (see 'Peggy Hamburg's FDA fixes'). Looking back on her experiences in New York, she says: "We took what was a faltering agency and restored it to its former position as the premier health department in the country. I do feel that that has a lot of relevance to the FDA now."

The parallels are striking. The FDA, once globally revered as the gold standard in regulation of food and medical-product safety, has lapsed repeatedly in recent years under a string of different leaders and a long stretch without any permanent chief. In 2004, the anti-inflammatory drug Vioxx (rofecoxib) was taken off the market after five years of sales because of cardiovascular side effects that may have caused tens of thousands of deaths in the United States alone[2].

The agency's scientific capabilities eroded to the point that its own science board declared two years ago that "not only can the agency not lead, it cannot even keep up with the advances in science"[3]. The FDA's centre in charge of medical-device approvals had also been rocked by charges that top regulators there overrode centre scientists and approved devices that put the public at risk — charges first aired by the centre's own scientists and given further credence in January in a report from the Government Accountability Office. And a series of food-poisoning outbreaks — including *Salmonella*-tainted peanut butter — have led to persistent questions about the agency's ability to ensure the safety of the country's food[4].

Peggy Hamburg has a difficult job on her hands in restoring public faith in the FDA.

E. VUCCI/AP

Although its $2.7-billion budget for 2010 lags far behind those of its sister agencies in the Department of Health and Human Services, the FDA is charged with policing the safety of goods that account for about one-quarter of the US consumer economy, from pacemakers to toothpaste. And its mandates keep growing. In 2007, the agency received new powers to police the safety of drugs already on the market, and in June, a new law brought tobacco products under its jurisdiction.

Increasing the budget of the 11,000-person agency "is critical", Hamburg says. "Not only are we trying to rebuild important core functions, we are also adding on some new responsibilities and gaining new authorities."

Hamburg has made it clear that she plans to step up enforcement to catch both fraudulent operations that prey on consumers and aboveboard companies that fail to comply with required manufacturing standards. The FDA set a tough tone starting in May by clamping down on websites that were marketing products fraudulently claiming to diagnose, prevent or treat pandemic H1N1 influenza.

Persuasive policy-maker

The qualities that Hamburg brings to the job, say her supporters, include an ability to analyse complex issues and to persuade — rather than browbeat — others to accept her point of view. "Pounding on the desk is not her style," says David Dinkins who, as mayor of New York in the early 1990s, was persuaded by Hamburg to back a controversial needle-exchange programme to combat the spread of

© 2009 Macmillan Publishers Limited. All rights reserved

HIV/AIDS. "She would reason with you. She would set forth a rationale. It just made sense. She seemed to me to always make sense."

PEGGY HAMBURG'S FDA FIXES

Issue	Solution
Lax monitoring of drugs on the market	Hamburg must implement a 2007 law that requires the FDA to set up a database of 100 million patients by 1 July 2012 as a platform for active surveillance of emerging drug safety risks.
Small budget, sprawling mandate	Hamburg will face an uphill battle pushing for increasing funding in times of ballooning federal deficits.
Food-safety inspectors outstripped and enforcement teeth lacking	Legislation currently under consideration would require the FDA to inspect food plants much more frequently. The agency would gain new authority to force food recalls.
Erosion of scientific expertise within the FDA staff	Hamburg has added a 'regulatory science' piece to the agency's 2011 budget request. She pledges to foster exchange programmes with external scientists and other educational opportunities to keep agency scientists up to date on research.

Perhaps most importantly for scientists at the agency, Hamburg, although steeped in public health, has an innate familiarity with scientific culture. She grew up on the Stanford University campus in Palo Alto, California, where her parents were both on the medical-school faculty. As a 15-year-old, she spent a semester in Africa with primatologist Jane Goodall, and before entering medical school she worked in a neuropharmacology lab at the National Institute of Mental Health in Bethesda, Maryland. Later, as a resident in internal medicine, she worked in the lab of neuroscientist Paul Greengard at the Rockefeller University in New York. In 1989 and 1990, she was deputy director of the National Institute of Allergy and Infectious Diseases in Bethesda, where she focused on infectious-disease research and policy.

Hamburg is proud of her time in the research world. She says she was "irritated" by media coverage opining that her public-health background — including four years as a top health adviser during the Clinton administration — made her a more natural fit to lead the Centers for Disease Control and Prevention. "As though that was my only background," she says. "People should recognize that I actually began my career in medicine doing some bench research." She says that the results of research inform her policy decisions, noting that she pushed for the needle-exchange programme in New York because of evidence that needle exchange reduced the rate of HIV transmission.

"She seemed to me to always make sense."
— David Dinkins

At the FDA, however, that kind of science-driven stance has not yet become evident in one contentious area: an emergency contraceptive called Plan B remains unavailable as over-the-counter medication to girls younger than 17 years old, even though FDA scientists in 2004 called it safe and effective for women of all ages and a federal judge urged the agency to revisit the issue in March.

In June, dozens of groups, including the American Academy of Pediatrics and the American College of Obstetricians and Gynecologists, wrote to Hamburg asking her to lift existing restrictions and make Plan B available without a prescription to younger girls. They have received no response.

"We are disappointed at the lack of action," says Kirsten Moore, president and chief executive of the Reproductive Health Technology Project, a non-profit organization based in Washington DC that organized the letter. Judy Leon, an agency spokeswoman, says that because there are ongoing legal challenges to the FDA concerning Plan B, the agency could not comment.

Despite the expectation that Hamburg will favour tighter regulation, industry reviews of her performance have been positive. "It is obvious that she brings clear vision and dedication to the job," says Ken Johnson, senior vice-president at the Pharmaceutical Research and Manufacturers of America in Washington DC, a lobby group for the country's pharmaceutical manufacturers.

Tevi Troy, a visiting senior fellow at the right-leaning Hudson Institute in Washington DC who was deputy secretary at the Department of Health and Human Services in the George W. Bush administration, calls Hamburg "the right kind of person" for the job. Troy is especially pleased with her background in biological security: in New York City, she instituted the first health-system-based anti-bioterror programme in the nation, and from 2001 until her appointment as FDA commissioner she was vice-president for biological programmes at the Nuclear Threat Initiative.

FDA
U.S. Department of Health and Human Services
Food and Drug Administration

The FDA is based in Silver Spring, Maryland.

J.REED/REUTERS

Still, says Troy, he has been made "nervous" by the abrupt departure in August of Dan Schultz from his position as chief of the FDA's device centre. He sees it as "a signal that anyone who even gets a reputation as being willing to listen to industry concerns may have trouble going forward" in a Hamburg-run FDA.

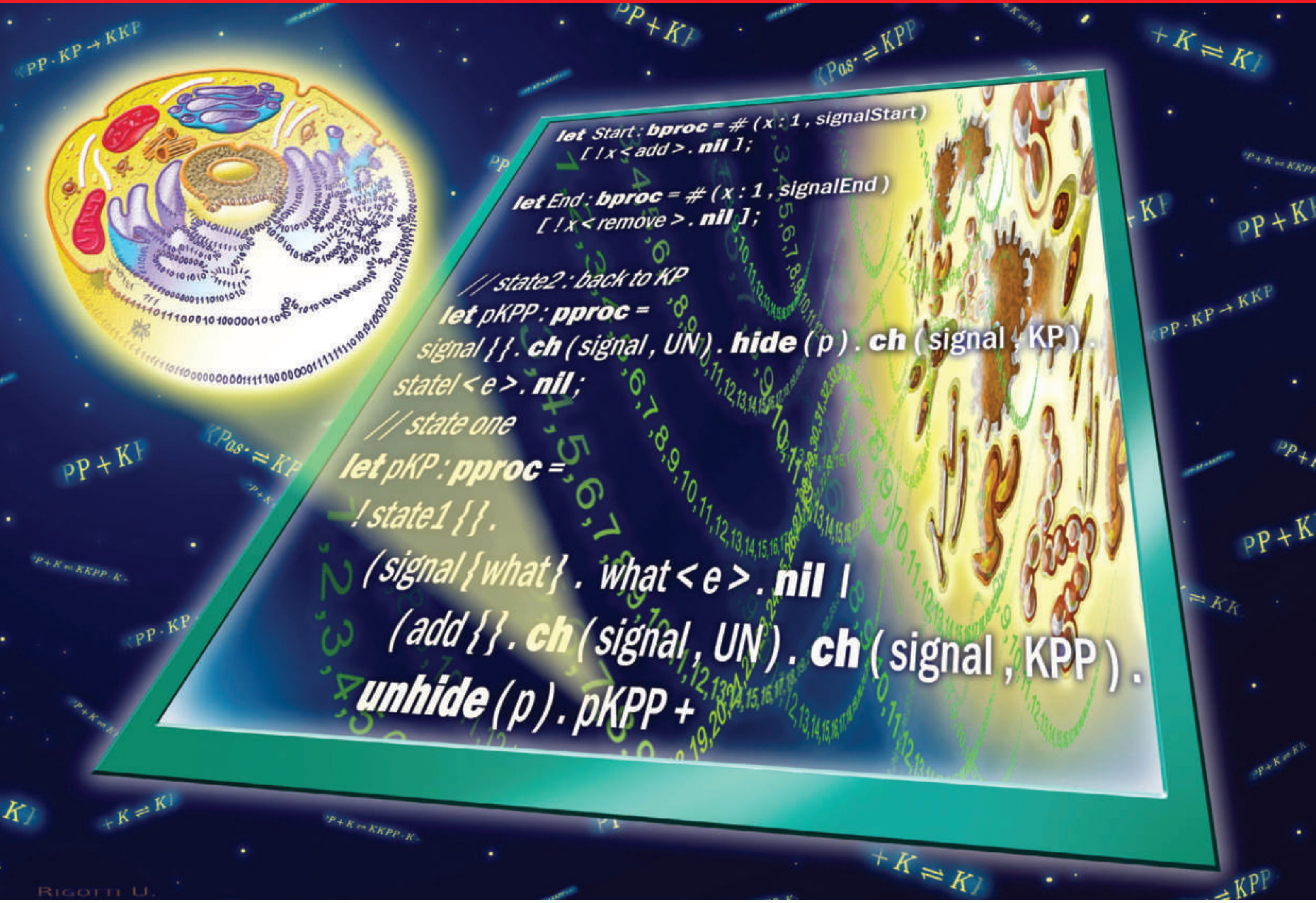
Yet some inside the agency assert that Hamburg has done nothing to change what they call a long-standing bias at the FDA in favour of the drug and device industries. "We have had multiple examples since the change in administration where the work of [FDA scientists charged with monitoring the safety of marketed drugs] was basically filed in the wastebasket," says one scientist who did not wish to be named because of job-security fears.

Hamburg has spent much of her short tenure dealing with the H1N1 pandemic; her agency is responsible for licensing vaccines and overseeing their quality after production. The demands of responding to H1N1 make clear just how much Hamburg's success at the agency will hinge on how well she balances unforeseen crises with her overall goal of reinvigorating the agency. She will be helped, say her supporters, by an almost preternatural ability to remain calm under pressure. Two years ago, Hamburg was in New York to give a speech on biological security when her husband called to say that their house in Washington DC was on fire. "Everyone is okay, but three fire trucks just pulled up," he said. She went ahead and gave the speech, without mentioning that her house was burning. ■

Meredith Wadman is a reporter for *Nature* based in Washington DC.

1. Frieden, T. R., Fujiwara, P. I., Washko, R. M. & Hamburg, M. A. *N. Engl. J. Med.* **333,** 229–233 (1995).
2. Graham, D. J. *et al. Lancet* **365,** 475–481 (2005).
3. Wasman, M. *Nature* **450,** 1143 (2007).
4. Yeager, A. *Nature* **457,** 770–771 (2009).

© 2009 Macmillan Publishers Limited. All rights reserved


let Start : bproc = # (x : 1 , signalStart)
[! x < add > . nil] ;
let End : bproc = # (x : 1 , signalEnd)
[! x < remove > . nil] ;
// state2 : back to KP
let pKPP : pproc =
signal { } . ch (signal , UN) . hide (p) . ch (signal , KP) .
state1 < e > . nil ;
// state one
let pKP : pproc =
! state1 { } .
(signal { what } . what < e > . nil |
(add { } . ch (signal , UN) . ch (signal , KPP) .
unhide (p) . pKPP +
PP + KP
+ K ⇌ KP
PP · KP → KKP
RIGOTTI U.

Biological logic

An intuitive approach to computer modelling could reveal paths to discovery, finds **Lucas Laursen**.

REF. 7

Grabbing one of the three laptops in her office at Microsoft Research in Cambridge, UK, Jasmin Fisher flips open the lid and starts to describe how she and her collaborators used an approach from computer science to make a discovery in molecular biology.

Fisher glances across her desk to where her collaborator, Nir Piterman of Imperial College London, is watching restlessly. "I know you could do this faster," she says to Piterman, who is also her husband. "But you are a computer scientist and I am a biologist and we must be patient."

After a few moments, patience is rewarded: Fisher pulls up a screen of what looks like programming code. Pointing to a sequence of lines highlighted in red, she explains that it is a warning generated by software originally developed for finding flaws in microchip circuitry. In 2007, she, Piterman and their colleagues found a similar alert in a simulation they had devised for signalling pathways in the nematode worm *Caenorhabditis elegans*. Using that as a clue, they predicted and then experimentally verified the existence of a mutation that disrupts normal cell growth[1].

'Executable biology', as Fisher calls what she's demonstrating, is an emerging approach to biological modelling that, its proponents say, could make simulations of cells and their components easier for researchers to build, understand and verify experimentally.

"Modelling in general is regarded sceptically by many biologists."
— Stephen Oliver

The screen full of code doesn't look especially intuitive to a non-programmer. But Fisher toggles to another window that shows the same *C. elegans* simulation expressed graphically. It now looks much more like the schematic diagrams of cell–cell interactions and cellular pathways that biologists often sketch on white boards, in notebooks or even on cocktail napkins. One big goal of executable biology is to make model-building as easy as sketching. Fisher explains that each piece of biological knowledge pictured on the screen, such as the fact that the binding of one protein complex to another is necessary to activate a certain signal, corresponds to a programming statement on the first screen. Likewise, the diagram as a whole — illustrating, say, a regulatory pathway — corresponds to a sequence of statements that collectively function as a computer simulation. Ultimately, she says, this kind of software should develop to a point at which researchers can draw a hypothetical pathway or interaction on the screen in exactly the way they're already used to doing, and have the computer automatically convert their drawing into a working simulation. The results of that simulation would then show the researchers whether or not their hypothesis

© 2009 Macmillan Publishers Limited. All rights reserved

corresponds to actual cell behaviour, and perhaps — as happened in the 2007 work — make predictions that suggest fruitful new experiments.

In the meantime, however, Fisher and her fellow executable-biology enthusiasts have a lot of convincing to do, says Stephen Oliver, a biologist at the University of Cambridge, UK. "Modelling in general is regarded sceptically by many biologists," he points out.

Born-again modeller

Fisher's fascination with this type of modelling started in about 2000. She was studying for her PhD in neuroimmunology at the Weizmann Institute of Science in Rehovot, Israel, when she encountered David Harel, a computer scientist who was applying computational ideas to biology.

Harel wanted to get around the problems encountered in conventional simulations, which use reaction-rate equations and other tools of theoretical chemistry to describe, step by step, how reaction networks and cell interactions change over time. Such simulations can provide biologists with a gratifying level of detail for testing against reality. But the number of differential equations in these models escalates rapidly as more reactions are included, until they become a strain on even the most powerful computers. In one recent model of the networks involving epidermal growth factor, for example, 499 equations were required to describe 828 possible reactions[2]. Even if the computers can handle such a load, the output is often difficult to interpret.

P. MYNOTT

Jasmin Fisher wants to be able to model complex cellular interactions.

Such models quickly become "an impossibly unwieldy black box", says Vincent Danos, a computational biologist at the University of Edinburgh, UK. And if the models have such a hard time simulating the behaviour of a single set of signalling pathways, he adds, then it's hard to imagine they will ever be of much use in systems biology, which might, for example, seek to understand all the pathways in a cell as an integrated whole.

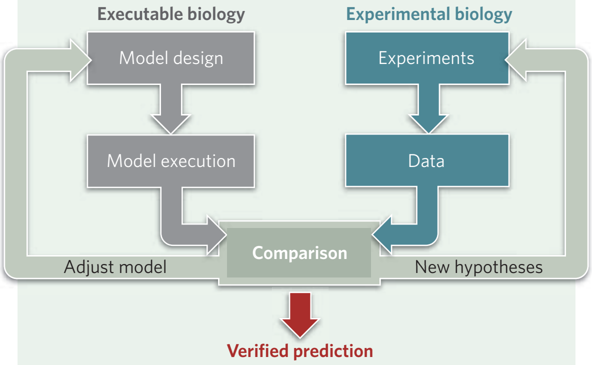
Harel's approach was to represent networks of biological events by a considerably smaller set of logical statements. For example, instead of specifying the number of signal molecules involved in a particular cell–cell interaction, or the sensitivity of the various receptors, a statement might simply say 'when cell X is near cell Y for long enough, cell Y switches from one type of behaviour to another'. And, unlike the conventional equations, the rules tend to be independent of one another — an important part of why the simulations are so much easier to build.

An additional advantage of the logic-based approach was that standard model-checking algorithms — widely used by industry for testing computer hardware — could check whether the statements were logically consistent, and capable of producing the behaviour seen in cells. This analysis would highlight points in the model at which the behaviour was going awry, which in turn might suggest experiments to look for previously unsuspected reactions and molecular species at that point (see graphic).

Fisher became so caught up in the idea that in 2003 she joined Harel's lab as a postdoc. She continued to work in the field during a three-year postdoc appointment under Thomas Henzinger at the computer-science department of the Swiss Federal Institute in Lausanne (EPFL). Piterman, whom she had married in 1998, came to the EPFL as well, and the three of them collaborated with their colleague Alex Hajnal to build the *C. elegans* model.

REFINED VIEW

Executable biology may suggest new hypotheses for testing.



They started by recording all the rules they could find in the literature pertaining to the maturation of a simple, well-studied system of six vulval precursor cells. "I wrote it all down first in a diagram," says Fisher, pointing to a figure in a research article on her desk, "then we formalized all the arrows and feedback loops into the computer program." Because the model needed only rules, not numbers, most of the information was qualitative (for example, this cell is closest to the cell sending the signal so the messenger molecules reach it first).

Lab confirmation

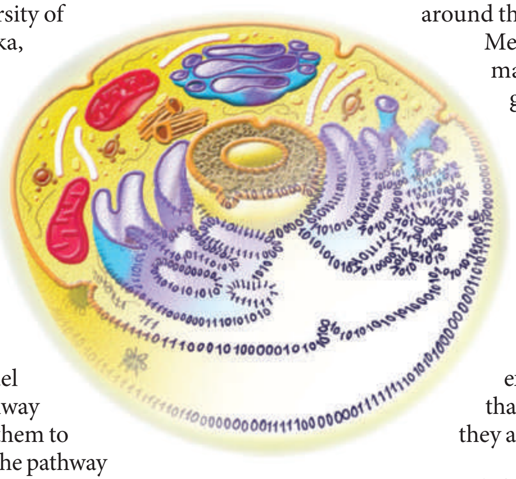
The team knew that genetic mutations could nudge the cells into different roles during maturation, but they wanted to know more about the cascade of signals that dictate the fate of each cell. The model-checker explored the set of 48 mutations known to affect vulval development, which could have up to 92,000 possible outcomes. All but four of the perturbations predicted normal cell fates, so the team concentrated on simulating different timings of those four cases. They found two previously unknown effects. First, a set of inhibitory genes collectively known as *lst* genes have to be activated for vulval cells to convert to their 'primary' fate, meaning that their daughter cells will make up the vulval opening. Second, if another gene was disrupted and signals between the cells weren't timed just in just the right sequence, the cell would adopt a different fate. A laboratory experiment confirmed both predictions.

"We used this qualitative model because we simply didn't have the quantitative knowledge," says Fisher. But now that the approach and its predictions have been verified in the lab, she says, "you can't argue with it".

Since then, Fisher has become one of the world's most energetic proponents of executable biology[3], but she is far from being the only enthusiast. In 2007, for example, biologist John Heath of the University of Birmingham, UK, was trying to model signal transduction pathways

© 2009 Macmillan Publishers Limited. All rights reserved

and protein–protein interactions. “The processes are just really just too complicated to understand using intuition,” he says. He discussed his problem with University of Oxford computer scientist Marta Kwiatkowska, who was then working in the adjacent building at Birmingham, and she gave him a paper on model-checking. “I was reading the opening paragraph on the train and I thought, ‘This is exactly what I want’,” says Heath. In collaboration with Corrado Priami, who leads the Centre for Computational and Systems Biology at the University of Trento in Italy, Heath was soon modelling the gp130/JAK/STAT signalling pathway[4], a well-studied system involved in human fertility, neuronal repair and embryonic stem-cell renewal. Their model reproduced the dynamic behaviour of the pathway as observed in the laboratory, and has allowed them to make testable predictions about which parts of the pathway are most sensitive to mutation or other perturbation. Heath, like Fisher, is now actively promoting executable biology, and has joined with Kwiatowska to publish a review paper on the approach[5].



REF. 7

Another level

Executable biology does have limitations, Fisher acknowledges. At present, for example, such models can handle only one level of narrowly defined biological activity at a time — the level of protein–protein interaction, say, or the level of cell–cell interaction. “We know there is feedback between the levels,” Fisher says, “but we don’t know enough about it” to get a computer to simulate that feedback.

An additional complication is that the different levels are best handled by different computer languages. To model the molecules that travel between cells, for instance, the most natural languages are those known in computer science as ‘process calculi’, which were devised to model information flow through communication webs. But to model the behaviour of an individual cell and its components, as in the various signalling and regulatory pathways, the most natural languages are those based on the theory of interacting ‘state machines’, which was developed to describe how objects transition from one state to another.

> “The model is not an oracle, it is an automation of your understanding.”
> — John Heath

The long-term goal, says Fisher, is to develop more sophisticated and complete simulations that would help researchers explore a wider range of biological phenomena, both by integrating behaviour at the genetic, molecular and cellular levels, and by integrating executable models with more mathematical models. Indeed, as a group of bioengineers led by C. Anthony Hunt of the University of California, San Francisco, pointed out in a response[6] to Fisher and Henzinger’s 2007 review, it’s not an either–or choice between the executable biology and conventional mathematical modelling: both have their uses and limitations, depending on the level of biological activity being simulated.

Fully integrated modelling is still a long way off, admits Fisher. But now that executable-biology predictions have been verified in the lab, the field has begun to attract more attention. Labs worldwide are starting to use executable biology to study systems, and Fisher herself is giving invited lectures on the subject 15–18 times per year around the world.

Meanwhile, she and Piterman are trying to make the software more accessible to biologists, so that researchers can make executable-biology simulations a routine part of their work. Other research groups are working towards the same end. Priami’s group is trying to write interfaces so simple that biologists can fill in tables with their data, specify the rules they want to use in spatially organized diagrams and sit back while the program translates the data into a computer-readable language that can execute a simulation[7]. “We develop languages that allow people to program without knowing they are programming,” says Priami.

Commercial efforts

In another effort to make the executable-biology approach more intuitive, Walter Fontana of the Harvard Medical School in Boston, Massachusetts, has joined with colleagues at the start-up firm Plectix to launch Cellucidate, an online visual interface for biological-pathway modelling that generates statements in an executable computer language called Kappa, which Fontana developed explicitly to model molecular interactions. Cellucidate — available for free during its trial period — allows collaborators to add information to a shared online model and revise it Wikipedia-style, something Fontana says is increasingly important because the empirical facts on which models are based are continually being revised.

Fisher hopes that the excitement will catch on in more groups and suggests that some of the computer-inspired ideas she is testing in her group’s latest *in vivo* experiments, which now extend to fruitflies and yeast cells, should entice more interest in executable biology among lab-based biologists.

But in the end, Fisher emphasizes, the fact that using executable rules could make the models easier to visualize is only an added bonus. Executable biology’s real pay-off is that it can help biologists to understand the complexity of living things, whether at the level of groups of molecules, such as Kappa describes, or at that of signals sent between cells, as in the nematodes Fisher herself studies. And that enhanced understanding, in turn, helps biologists ask new questions, design new experiments and make new discoveries. “But however good the models are, “you still need a good scientist to implement them”, says Kwiatkowska.

“The model is not an oracle,” Heath agrees, “It’s an automation of your understanding.” ■

Lucas Laursen is a freelance journalist in Cambridge, UK.

1. Fisher, J., Piterman, N., Hajnal, A. & Henzinger, T. A. *PLoS Comput. Biol.* **3,** e92 (2007).
2. Chen W. W. *et al. Mol. Syst. Biol.* **5,** 239 (2009).
3. Fisher, J. & Henzinger, T. A. *Nature Biotechnol.* **25,** 1239–1249 (2007).
4. Guerriero, M. L., Dudka, A., Underhill-Day, N., Heath, J. K. & Priami, C. *BMC Syst. Biol.* **3,** 40 (2009).
5. Kwiatkowska, M. Z. & Heath, J. K. *J. Cell Sci.* **122,** 2793–2800 (2009).
6. Hunt, C. A., Ropella, G. E. P., Park, S. & Engelberg, J. *Nature Biotechnol.* **26,** 737–738 (2008).
7. Priami, C. *Commun. ACM* **52,** 80–88 (2009).

© 2009 Macmillan Publishers Limited. All rights reserved

CORRESPONDENCE

The road ahead for brain-circuit reconstruction

As someone who has spent the past 25 years charting brain circuits, I am baffled by the view expressed in your Technology Feature that "sadly, ... pretty much" nothing has happened in my field since the early 1980s (*Nature* **461,** 1149–1152; 2009).

Unlike structural descriptions of the Universe, fossil bones or molecules, neural structure has not been a vote-winner among high-profile journals. However, neurocircuiteers have not been waiting patiently in their backwater for a quarter of a century for the arrival of new molecular, genetic and imaging techniques. They have been describing circuits through a variety of clever physiological and anatomical experiments, coupled with hard theory and analysis.

The new techniques offer no improvement in resolution over those that have been available for more than 50 years. Electrophysiology and light and electron microscopy are still the gold standards in space, time and reach for studying any region of any brain.

The new thing these structural techniques promise is volume. This is great, because it means that the wiring diagrams of small brains such as *Drosophila*'s may become available in a decade or so.

One elephant remains in the room. How do we use the circuit reconstructions (involving exabytes of data) that these high-throughput techniques deliver?

Loading any circuit into the biggest super-simulator available and switching on tells us nothing useful. Like a motorcar without wheels being started up by a Martian, exciting noises may come from the exhaust but it isn't going to take us anywhere. It is coupling these powerful techniques to predictive models of neural circuits that will really allow us to go places.

Kevan A. C. Martin Institute of Neuroinformatics, UZH/ETH,
Winterthurerstrasse 190,
8057 Zurich, Switzerland
e-mail: kevan@ini.phys.ethz.ch

Darwin respected by his religious contemporaries

The Church in England did not generally react so "badly" to Darwin's ideas as readers of your Editorial may be led to believe (*Nature* **461,** 1173–1174; 2009).

Reverend Charles Kingsley, Regius Professor at the University of Cambridge, UK, wrote in 1863 "God's greatness, goodness and perpetual care I never understood as I have since I became a convert to Mr Darwin's views." The Bishop of Carlisle, Harvey Goodwin, proclaimed after Darwin's funeral in Westminster Abbey "It would have been unfortunate if anything had occurred to give weight and currency to the foolish notion which some have diligently propagated, but for which Mr Darwin was not responsible, that there is a necessary conflict between a knowledge of Nature and a belief in God." In 1884 Frederick Temple, Bishop of Exeter and future Archbishop of Canterbury, wrote "The doctrine of Evolution restores to the science of Nature the unity which we should expect in the creation of God." Aubrey Moore, a leading theologian at the University of Oxford, welcomed Darwinism "as a friend in the disguise of a foe" because it struck at the heart of nineteenth-century deism.

Ironically, in view of later developments, even some of the authors of *Fundamentals* (a series of Christian booklets published in the United States between 1910 and 1915) were happy to see evolution as the method that God used in his work of creation.

The assumption that there must be conflict between evolution and religion was (and is) the result of the distorting "cultural lenses" that you mention. Modern 'creationism' was born only in the twentieth century, largely through the efforts of the Canadian adventist George McCready Price. There has probably been less conflict in England than in most other countries.

None of this is to claim that all religious people view evolution in a positive light, nor that all evolutionists are objective about religion. But we need to remain aware of our cultural lenses.

R. J. Berry Department of Biology,
University College London,
London WC1E 6BT, UK
e-mail: rjberry@ucl.ac.uk

Why some relatives object to organ donation

You question in an Editorial the determination of death for organ-transplant purposes in the United States, where explantation can go ahead once all functions of the entire brain have irreversibly ceased (*Nature* **461,** 570; 2009). Nothing so rigorous is demanded in the United Kingdom.

For successful transplantation, major organs such as the heart, lungs, bowel and liver must be alive. For some 30 years, UK practice has required only bedside tests purporting to show an irreversible loss of ability to breathe and the irreversible cessation of some brainstem functions. Higher parts of the brain may continue to function. As a consultant anaesthetist (now retired), it greatly concerns me that the donor will need some form of paralysis and anaesthesia to control the responses to explantation surgery.

The UK technical definition of death for transplantation purposes is not explained on donor cards or on the donor register, so those who sign up may have a quite different concept of "my death". This may explain the 40–50% refusal rate among relatives when they observe the condition of someone declared dead but still showing signs associated with life.

David J. Hill Eltisley, Huntingdon,
Cambridgeshire PE19 6TG, UK
e-mail: david.hill01@tiscali.co.uk

Readers are welcome to join this debate at Nature Network, go.nature.com/WjUiku.

Brazil's system stops its natural wealth helping science

The pessimism expressed in your Naturejobs feature about the prospects for life sciences in Brazil is justified (*Nature* **461,** 1308–1309; 2009). Unfortunately, the country's science enterprise depends as much on its societal values as on its booming economy and wealth of natural resources.

Brazil's prevailing political and cultural outlook means that its economic growth has not proportionately reduced its chronic poverty and income inequalities. Nor has this growth promoted modernization of its political or financial systems — education, science and technology included.

Poor management of Brazil's abundant natural resources means that, although these account for some 68% of its positive trade balance, the country is left with less than 1% of the money created from its mineral exports. For example, Brazilian iron exports alone totalled US$16 billion last year; however, mining royalties amounted to only $462 million. And with gold royalties at just 1%, Brazil has the world's lowest taxation on gold.

Some sectors are campaigning for new legislation to remedy this situation. A separate strategy will be needed to direct any additional government money towards improving Brazilian science.

Sergio U. Dani Excegen Genetica SA,
Acangaú Valley, CxP 123, 38600-000
Paracatu MG, Brazil
e-mail: srgdani@gmail.com

© 2009 Macmillan Publishers Limited. All rights reserved

OPINION

International spaces promote peace

Lessons are still being learnt from the Antarctic Treaty, adopted 50 years ago this week. It set a visionary precedent for governing regions and resources beyond national jurisdictions, says **Paul Arthur Berkman**.

U. S. A.
To Laurence Gould without whom there would be no Antartica Treaty, with warm regards
Herman Phleger

CARLTON COLLEGE ARCHIVE

US embassador Herman Phleger signing the Antarctic Treaty on 1 December 1959. He later autographed this photo: "To Laurence Gould, without whom there would be no Antarctica Treaty".

This year marks the 50th anniversary of a landmark treaty — the planet's first nuclear arms-control agreement, and the first institution to govern all human activities in a region beyond sovereign jurisdictions. Adopted in Washington DC on 1 December 1959, the Antarctic Treaty recognized that "it is in the interest of all mankind that Antarctica shall continue forever to be used exclusively for peaceful purposes and shall not become the scene or object of international discord".

During the 1960 ratification hearings of the Antarctic Treaty in the US Senate, polar scientist and explorer Laurence McKinley Gould testified that it was "a document unique in history that may take its place alongside the Magna Carta and other great symbols of man's quest for enlightenment and order". This comparison to England's legal charter of 1215, renowned worldwide as a seminal precedent for constitutional law and national democracy, may seem presumptuous. But it is fitting.

Nearly 75% of Earth's surface lies beyond national boundaries. International institutions governing such spaces are still in their infancy, having originated largely in the aftermath of the Second World War, when humankind was inexorably introduced to our global interdependence. Humankind is only gradually awakening to the shared responsibility for governing human activities in these international spaces and for managing the effects of global phenomena such as climate change. At this threshold in our civilization, the Antarctic Treaty offers a unique precedent.

"The Antarctic Treaty demonstrates the strength of science as a tool of diplomacy."

Since 2000, with collaborators around the world, I have been planning an interdisciplinary and inclusive event to celebrate the first fifty years of the Antarctic Treaty. An open Antarctic Treaty Summit will be held from 30 November to 3 December 2009 at the Smithsonian Institution in Washington DC (www.atsummit50.aq). The summit will highlight lessons learned about science–policy interactions in international cooperation and governance. It also will introduce the Forever Declaration — a non-binding affirmation of the Antarctic Treaty legacy, open for signature on 1 December (on the above website) to anyone anywhere with hope for enduring peaceful uses of regions and resources beyond national jurisdictions.

The ice-covered continent of Antarctica is surrounded by oceans and is without indigenous human populations. It could easily have become an area for weapons testing and storage, or been divided up between nations interested in exploiting its resources. The first nation to claim territory in the Antarctic was Great Britain in 1908, followed by New Zealand, France, Australia, Norway, Chile and Argentina. Some claims overlapped. To avoid territorial conflicts and to preserve sovereignty rights, in 1948 the United States issued to the seven claimant nations a secret aide memoire with a draft agreement proposing an international status for the Antarctic area.

The draft focused on the global relevance of science and exploration, as well as on the importance of maintaining international peace and security in Antarctica. This antecedent of the Antarctic Treaty matured under the statesmanship of President Dwight D. Eisenhower, who entered office in 1953 envisioning "a day of freedom and of peace for all mankind".

During the cold-war period of the late 1940s and early 1950s, the United States and Soviet Union raced to create missiles that could deliver nuclear weapons across continents. Few bridges were being considered, much less built, between these superpowers. The treatment of Antarctica, at first, was no exception. At a US National Security Council meeting in June 1954, a territorial solution for the Antarctic was discussed that would "ensure maintenance of control by the United States and friendly powers and exclude our most probable enemies". Curiously, it was rocketry that would also herald cooperation in the Antarctic.

Science for peace

Meanwhile, the International Council of Scientific Unions (ICSU) had begun planning the International Geophysical Year (IGY) for 1957–58 to coordinate geophysical observations on a planetary scale. At their October 1954 meeting in Rome, the ICSU further recommended the development of satellites for the IGY, to advance upper-atmospheric research and provide unparalleled measurements of the Earth system.

Recognizing the inevitability of satellites and ballistic missiles, Eisenhower introduced his 'Open Skies' proposal in Geneva on 21 July 1955, whereby the United States and the Soviet Union would give each other a "complete blueprint of our military establishments" as part of a system of mutual aerial reconnaissance.

© 2009 Macmillan Publishers Limited. All rights reserved

Eisenhower's hope was for "practical progress to lasting peace". But his proposal was rejected by the Soviet Union as an "espionage plot".

The following week, the White House disclosed its first space policy: the United States would launch small Earth-circling satellites during the IGY. Special efforts were made to ensure that this was seen as a peaceful project. The US Navy was chosen to conduct the satellite launch, even though the Army was technologically more advanced in rocketry. In fact, the Army Ballistic Missile Agency was specifically restrained by the White House from firing the fourth stage of the Jupiter-C rocket during a September 1956 test launch for fear of exacerbating the cold war. Instead, the freedom of space was preserved and perhaps because of this, the Soviet Union became the first into orbit with Sputnik in October 1957, followed three months later by the first US satellite.

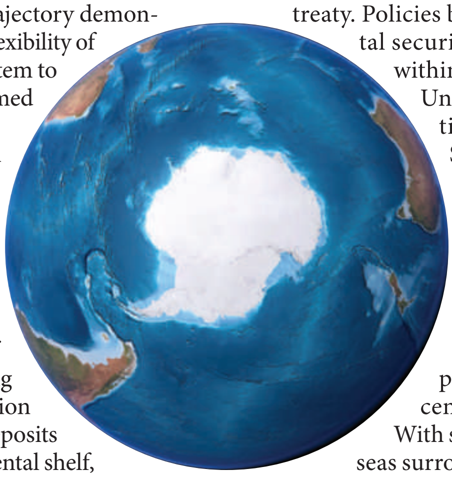
Eisenhower had failed to push through his Open Skies proposal, but there was another front on which he hoped to engage the Soviet Union in peace talks. Building on the momentum of scientific cooperation during the IGY, in May 1958, President Eisenhower invited the Soviet Union and the other ten nations involved with Antarctic research (the seven claimants, plus Belgium, Japan, and South Africa) to seek an effective means of ensuring that the "vast uninhabited wastes of Antarctic shall be used only for peaceful purposes". Over the next 18 months, 60 secret meetings were convened in the United States, culminating in the Conference on Antarctica between 15 October and 1 December 1959, when the Antarctic Treaty was signed.

The Antarctic Treaty is elegant in its simplicity. It has just 14 articles to govern the area south of latitude 60 ° S, covering nearly 10% of Earth's surface. Territorial issues were set aside. "Substantial research" activities became the criterion for nations to consult on "matters of common interest" (species conservation, open inspection, questions of jurisdiction, freedom of scientific investigation, scientific cooperation and peace) and to make decisions by consensus every one or two years. The Antarctic Treaty became the first nuclear-arms agreement, with the unrestricted inspection strategies that Eisenhower had envisioned for Open Skies. With the IGY, science had become a tool of diplomacy.

The first institution to govern a region beyond national boundaries, but without blanket governance, was the 1958 Convention on the High Seas, which formalized several long-standing concepts of international law, including the freedoms of navigation and fisheries as well as the prevention of piracy, pollution and slavery. It was the 1959 Antarctic Treaty, however, that first governed all activities in an international space, demonstrating how common interests could be used to overcome distrust. The Antarctic Treaty became the precedent for the 1968 and 1972 non-armament treaties for outer space and the deep sea, respectively.

Policy building

Once the Antarctic Treaty was in place, the signatories began to build specific policies concerning their common interests, starting with species conservation. With advice from the Scientific Committee on Antarctic Research (an ICSU body), the signatories agreed on measures for the conservation of Antarctic fauna and flora in 1964. A conservation convention for Antarctic seals was adopted in 1972. In 1980, the Convention on the Conservation of Antarctic Marine Living Resources introduced an ecosystem approach for the rational use of species living in the Southern Ocean — an area with global importance because of its extensive biomass. This policy trajectory demonstrates the success and flexibility of the Antarctic Treaty system to reach agreements informed by science.



PLANETARY VISIONS LTD/SPL

It was mineral resources that truly tested the resilience of the Antarctic Treaty consultative process. Following the 1973–74 oil embargo by the Organization of the Petroleum Exporting Countries and speculation about vast oil and gas deposits on the Antarctic continental shelf, new signatories to the Antarctic Treaty expanded exponentially over the next 15 years as nations asserted their interests in potential mineral exploitation. There was intense discussion during this period about how to regulate mineral resource activities, but these negotiations fell apart in the late 1980s. Soon after, the signatories signed the 1991 Protocol on Environmental Protection to the Antarctic Treaty, which prohibits any activity relating to mineral resources other than scientific research. Even for extremely divisive issues, the treaty process was capable of creating resolution.

As US secretary of state Hillary Clinton noted at the April 2009 Antarctic Treaty Consultative Meeting, "the genius of the Antarctic Treaty lies in its relevance today". The Antarctic Treaty model recognizes that solutions to trans-boundary or global issues must be processes involving cooperation, iteration and responsiveness to ever-changing circumstances. This lesson is particularly relevant to managing our changing climate, with perspectives and expectations beyond solutions forged at a single meeting. The challenge for governments and civil society is to envision a science-policy process that will operate over decades and centuries.

The Antarctic Treaty is especially relevant to the Arctic, where stakeholders have thus far avoided shared discussions about peace and security. Amplified climate warming in the polar regions is causing the Arctic Ocean to transition from a permanent ice cap to a seasonally ice-free sea: the most profound environmental state change on Earth. Risks of political, economic and cultural instability are inherent.

Before it becomes ice free and new commercial activities become entrenched, there is opportunity in the Arctic Ocean to establish a process of continuous policy development that explicitly promotes cooperation and prevents discord. This does not require a new treaty. Policies based on environmental security could be facilitated within the framework of the United Nations Convention on the Law of the Sea, in concert with the scientific advice of the Arctic Council and other institutions. An important outcome of this consultative process would be inspired climate adaptation policies with relevance centuries into the future. With statesmanship, the high seas surrounding the North Pole could become the next pole of peace.

The Antarctic Treaty demonstrates the strength of science as a tool of diplomacy, having facilitated peaceful cooperation between adversaries and allies at the height of the cold war. The future of our world requires leaders who can apply all such tools to balance national and common interests. Reflecting on the lasting legacy and lessons of the Antarctic Treaty during its first fifty years, 1 December deserves to be celebrated as a day of "peace for all mankind". ■

Paul Arthur Berkman is head of the Arctic Ocean Geopolitics Programme at the Scott Polar Research Institute, University of Cambridge; chair of the International Board for the Antarctic Treaty Summit; and a research professor at the Bren School of Environmental Science and Management at the University of California, Santa Barbara.
e-mail: paul.berkman@spri.cam.ac.uk

See go.nature.com/xoiQkv for further reading.

© 2009 Macmillan Publishers Limited. All rights reserved

BOOKS & ARTS

Some will go far to catch a falling star

Henner Busemann enjoys a hymn to the passionate collectors who fuelled the science of meteorites.

The Fallen Sky: An Intimate History of Shooting Stars
by Christopher Cokinos
Tarcher/Penguin: 2009. 528 pp. $27.95

Meteorites — rocks that 'fall from the sky' — fascinate and inspire. The origin of these dark, often strangely sculpted boulders that might suddenly dent a ploughed field or demolish a roof has long been disputed. Yet the global consequences of meteorite impacts have only recently been accepted. It took the efforts of a few visionaries — mostly non-academics who had to endure scholastic resistance, sarcasm and slander — to demonstrate that rocks on Earth can come from asteroids, the Moon, Mars and comets; that their fiery crashes created giant craters; and that such bombardment was important for the development of life on Earth.

Christopher Cokinos, a creative-writing teacher at Utah State University, erects a monument to these dedicated pioneers in *The Fallen Sky*. As well as telling their personal stories, he covers a comprehensive range of topics in meteorite science — from the observation and evaluation of fireball trajectories to the discovery of meteorites, their transport, classification, conservation and ownership. He also assesses the rocks' commercial, scientific and spiritual value.

Because of their celestial origin and rarity, meteorite hunting kindles strong passions. Jealousy, personal animosity, commercialization and a struggle for fame and possession have often accompanied the chase. Cokinos retraces the footsteps of the first US meteorite 'addicts', such as the explorer Robert Peary, who transported the Cape York iron meteorites from Greenland in the 1890s, and Daniel Barringer, a mining-company owner who purchased Meteor Crater in Arizona in 1903 in the hope that a large lump of iron could be found there. The desire to possess a meteorite led others into legal disputes. Farmer Ellis Hughes, for example, was convicted of theft after 'recovering' the famous Willamette iron meteorite from private land in 1902. The 14-tonne main chunk of this is now in the American Museum of Natural History, New York.

Cokinos's efforts culminate in three splendid chapters. One concerns pioneer Harvey Nininger, a biologist from Kansas who was the first to systematically trace and collect meteorites in the United States. Cokinos gives a thrilling description of Nininger's epic fight to raise awareness of the importance of meteorites as extraterrestrial rocks, and his education of US rural populations in how to recognize meteorites on their farmlands. Hundreds of his samples, including rare specimens essential to meteoriticists, are now held by the British Museum, and he is considered one of the leaders of modern meteorite research. Yet Nininger's hopes for an appointment as a professor for meteoritics never materialized. Traduced as a collector with merely commercial interests, Nininger struggled to earn enough to support his family yet managed to accumulate an impressive private collection of meteorites.

"Jealousy, personal animosity and a struggle for fame have often accompanied the chase."

Cokinos conveys his excitement at visiting Nördlingen, a picturesque medieval town in the south of Germany, built within a meteorite crater of some 20 kilometres in diameter that formed around 15 million years ago. The town's centuries-old church is made entirely of suevite rock formed by the impact, and offers a full panoramic view of the crater rim from the steeple. Yet it wasn't until 1960 that US scientists Eugene Shoemaker and Edward Chao proved that the Nördlinger Ries crater was created by a meteorite impact.

Finally, Cokinos describes his participation in an annual excursion to Antarctica to collect meteorites — organized by researchers William Cassidy and Ralph Harvey and mountaineer John Schutt, among others. Since the 1970s, scientists supported by the US National Science Foundation have sought meteorites in the perpetual ice, where the strange rocks can be easily recognized and also accumulate as a result of being transported through the ice sheet. Cokinos's detailed description of the pleasures and discomforts of such an extreme expedition is somewhat overdone, though highly recommended to all potential candidates.

The Fallen Sky is an inventive introduction to meteoritics and contains a wealth of scientific, historic and biographic information. It will suit both general readers and planetary scientists who, swamped by detail, might have lost track of the basic motivation for their research. ■

Collector Marvin Killgore with the Fukang meteorite, which could be as old as the Solar System.

M. SEGAR/REUTERS

Henner Busemann is an STFC Aurora Research Fellow in the School of Earth, Atmospheric and Environmental Sciences at the University of Manchester, Oxford Road, Manchester M13 9PL, UK.
e-mail: henner.busemann@manchester.ac.uk

© 2009 Macmillan Publishers Limited. All rights reserved

A wake-up call to educators

Crossing the Finish Line: Completing College at America's Public Universities
by William G. Bowen, Matthew M. Chingos and Michael S. McPherson
Princeton University Press: 2009. 413 pp.
$27.95, £19.95

A. WONG/GETTY

US high-school grades are a better predictor of university graduation rates than admissions-test scores.

In the United States, earning a bachelor's degree is recognized as the most important factor for reducing economic inequality and increasing social mobility. But since the mid-1970s, university graduation rates have stagnated and disparities in educational outcomes have risen both among ethnic minorities and among those with low socioeconomic status. In *Crossing the Finish Line*, authors William Bowen, Matthew Chingos and Michael McPherson analyse these troubling trends and propose solutions to help colleges support their students more effectively.

Using regression analysis, the authors tracked and compared degree-completion rates for different groups based on various criteria. They found that students with low socioeconomic status and those from ethnic minorities — particularly black men and Hispanic students — were least likely to graduate. The authors also tested the predictive limitations of college admissions-test scores, the effectiveness of need-based financial aid and the ease of transferring between institutions. Although focused on the US educational system, these data contain warnings that other countries should heed.

"Degree completion has slowed to unacceptably low levels."

By following the incoming class of roughly 125,000 freshmen entering their first course at 68 US universities in 1999, the authors show that degree completion has slowed to unacceptably low levels. Just 65% of full-time students graduated in four years from the most selective 'flagship' universities, and only half graduated within six years from the least selective public universities. The authors argue that for those students who complete their education, delaying degree attainment from the standard four years to five or six years increases their financial burden and limits their future educational and career opportunities.

Admissions mechanisms, such as the sorting of applicants by universities and colleges and the reliance on standardized tests, dictate which types of institutions students attend. However, student scores on the SAT Reasoning Test (formerly the Scholastic Assessment Test) or the American College Test (ACT) are known to be heavily biased by gender, race and socioeconomic status, such that high test scores and wealth go hand-in-hand, often conferring an advantage on white male students. Interestingly, the authors' analyses revealed that scores from the SAT and the ACT do not predict graduation rates. Instead, high-school grade-point average is the most powerful predictor of both four-year and six-year graduation rates, regardless of the quality of the high school attended. Another surprise was that the scores from tests in individual subjects were able to predict graduation rates: both Advanced Placement tests and SAT Subject Tests (additional exams required by selective universities) were strongly predictive.

The authors found that academically overqualified students who attend less-demanding schools — known as undermatching — have a significantly higher probability of never completing their degrees than comparably qualified students who attend more-selective universities. This is especially prevalent among black men, they note. Many minority students and those of low socioeconomic status undermatch: 59% of students in the bottom quartile of family income do so, compared with 27% of those in the top quartile. In addition, 64% of students whose parents have no college education undermatch, compared with 41% and 31% of those whose parents have college or graduate degrees, respectively.

Difficulty in transferring between colleges also disproportionately affects students from minorities and of low socioeconomic status. For example, students who sought to save money by completing the first two years of their degree at a local community college before transferring to a more expensive public university for the remaining two years had an especially low graduation rate owing to limited transfer opportunities. However, those who did manage to gain later admission into a four-year-institution did well — better, in fact, than first-time freshmen with stronger pre-college credentials who went directly to a four-year university.

The authors' offer several solutions to these worrying trends. They include: early identification of high-performing students from disadvantaged backgrounds and then tracking them to prevent undermatching; greater investment in need-based financial aid to help qualified students of low socioeconomic status to enter a four-year institution directly; and encouraging four-year universities to accept more transfer students.

Crossing the Finish Line serves as a wake-up call to educators and administrators, and provides valuable data that will help universities to invest their resources in nurturing the talents of all their students. It also provides a disturbing glimpse of the far-reaching effects of limited expectations and diminished educational opportunities. ■

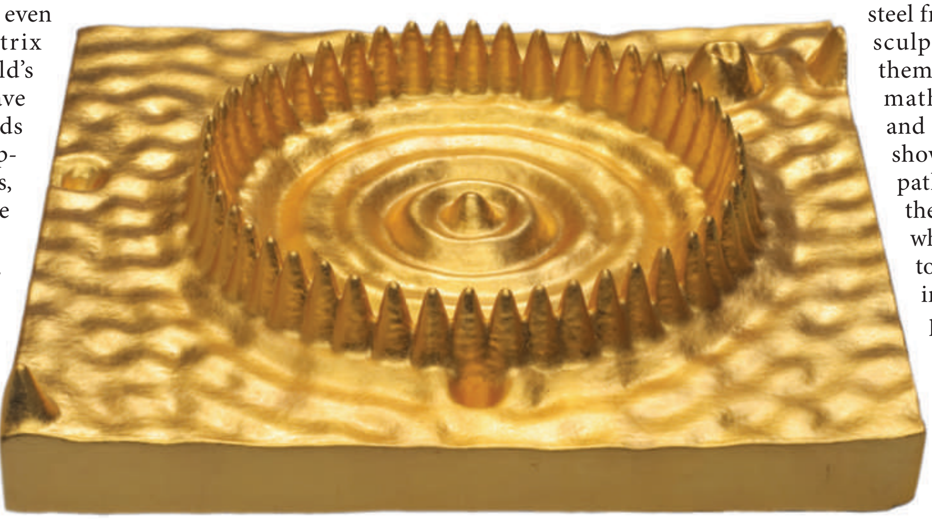
Devorah Bennu is a researcher and writer who writes the blog 'Living the Scientific Life (Scientist, Interrupted)' at ScienceBlogs.com.
e-mail: grrlscientist@gmail.com

© 2009 Macmillan Publishers Limited. All rights reserved

Quantum objects on show

Worlds Within Worlds: Quantum Objects by Julian Voss-Andreae
American Center for Physics, College Park, Maryland
Until 16 April 2010

When asked what his Third Symphony meant, Ludwig van Beethoven is said to have sat down at the piano and begun playing it. Analogously, a physicist might write down Erwin Schrödinger's wave equation as an 'explanation' of quantum theory. But even this formula was Schrödinger's alternative to Werner Heisenberg's even more abstract matrix mechanics. The field's pioneers seem to have concluded that words and images fail to capture quantum concepts, and that equations are all we have left.



Julian Voss-Andreae's gilded wooden *Quantum Corral* depicts quantum waves within a ring of iron atoms.

PHOTO: D. KVITKA/WWW.JULIANVOSSANDREAE.COM

On the contrary, the Oregon-based sculptor Julian Voss-Andreae thinks that art, when free from the demands of literalism, "has a unique potential for indicating aspects of reality that science cannot". He is well placed to judge. Previously a quantum physicist at the University of Vienna, in 1999 he participated in a groundbreaking experiment showing that even objects as 'big' as C_{60} molecules can display the wave-like property of interference. Voss-Andreae's portrayals of quantum objects are now on show in the exhibition *Worlds Within Worlds* at the American Center for Physics in College Park, Maryland.

"There simply are no consistent mental images we can create to understand the world as it is portrayed in quantum physics, because our brains are exquisitely adapted to making sense of the world on our scale," says Voss-Andreae. "I want to increase the audience's capacity to intuit the unfathomable deeper nature of reality by sensually experiencing the works."

This impulse to leap beyond the logical has an obvious appeal to artists, and Voss-Andreae is not the first to find inspiration in modern physics. In the early twentieth century, Surrealist artists such as André Breton and Salvador Dalí were excited by the challenge posed by quantum theory and relativity to conventional notions of causality, time, geometry and objectivity (see *Nature* **453,** 983–984; 2008). Their enthusiasm was rooted mainly in the perceived radicalism of the new physics rather than in an understanding of the science.

Yet even practitioners do not claim to fully understand quantum theory. The disputes about interpretation among the early pioneers such as Albert Einstein, Niels Bohr, Heisenberg and Schrödinger are legendary, but they are still with us. The Copenhagen interpretation — with its wave-particle duality, probabilistic picture and observer-induced collapse of the wave function — is still not universally accepted. And the nature of the transition from quantum to classical behaviour continues to be debated. The failure to unify quantum theory with gravitation leaves open the possibility that the theory is a stop-gap, a mathematical convenience.

Voss-Andreae is therefore either brave or foolhardy to try to represent quantum phenomena tangibly. Perhaps his greatest asset as a former physicist is that he realizes how much we don't know. In some of his works, the inverted commas of analogy are explicit to the knowing eye. *Quantum Corral* (pictured) materializes something that could hardly be less material: the wave-like properties of electrons, first reported in *Nature* in 1927 (C. Davisson and L. H. Germer *Nature* **119,** 558–560; 1927). Here, they are represented in a block of wood that has been milled to the contours of electron density seen in 1993 around a ring of iron atoms on the surface of copper through a scanning tunnelling microscope. The gilded surface reminds physicists that it is the mobility of surface electrons in the metal that accounts for its reflectivity, and the coloration of gold is itself a relativistic effect of the metal's massive nuclei. For art historians, this gilding invokes the crown-like haloes of medieval altarpieces, but could also allude to the way gold was reserved in Renaissance art for the intangible: the other-worldly light of heaven.

Voss-Andreae's works *Night Path* and *Spin Family (Bosons and Fermions)*, with their webs of metal wire or silk thread in solid steel frames, hark back to the sculptures of Naum Gabo, themselves inspired by new mathematical geometries and models. Yet *Night Path* shows a quantum idea: the path-integral approach to the trajectories of light, in which the passage of a photon is considered to be the integral over all possible paths, calculated by slicing up time. Here Voss-Andreae is not trying to produce a textbook representation. Rather, "the paths emerge from one point and then keep opening up", he explains. "I made it to illustrate the 'feel' of it." In the *Spin Family* series, inspired by the quantized spin states of the two classes of fundamental particle, the diaphanous silk thread allows us to visualize superpositions of states while cautioning against too literal a picture of what 'spin' itself represents.

A feeling of intangibility and the subjectivity of points of view pervades *Quantum Man*, a walking figure created from parallel slices of steel in which the particle-like concreteness seen from the front shifts to wave-like near-invisibility when the piece is viewed from the side. This sense of an object on the point of disintegrating is a common trope of recent artistic efforts to capture ideas in physics, from Antony Gormley's *Quantum Cloud* series to Cornelia Parker's *Cold Dark Matter*. Put the pieces together yourself, they seem to say — because that's how the world works anyway. ■

Philip Ball is a freelance writer based in London. His latest books form a trilogy called *Nature's Patterns*.

© 2009 Macmillan Publishers Limited. All rights reserved

NEWS & VIEWS

BIODIVERSITY

Skates on thin ice

Nicholas K. Dulvy and John D. Reynolds

The common skate is not at all common: this large marine fish has 'critically endangered' status. That it turns out to be not one species, but two, is a sharp reminder that good taxonomy must underpin conservation.

Conservation biologists who study obscure species in obscure places are faced with the challenge of protecting species they cannot even name. Such taxonomic headaches evidently blight even charismatic species in the best-studied regions of the world — as reported in *Aquatic Conservation*[1], Samuel Iglésias and colleagues have discovered that one of the largest species of fish in the northeast Atlantic is actually two. The common skate, *Dipturus batis*, which grows to a length of more than 220 centimetres, and has 'critically endangered' status on the IUCN Red List of Threatened Species, is really two species with different body sizes, ages at maturity, teeth, fins and eye colour.

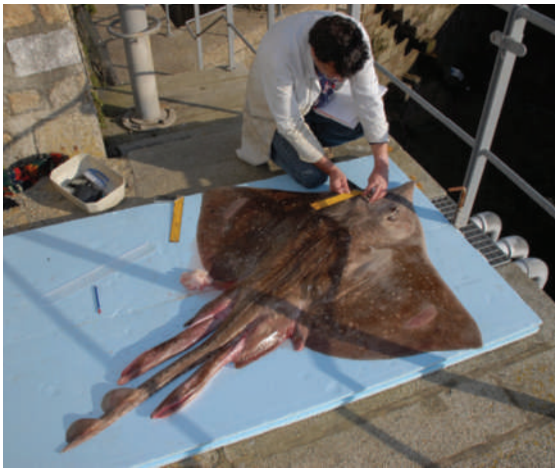

S. IGLÉSIAS

Figure 1 | Big difference. This male fish, with 'claspers' on either side of the tail, is an example of the larger of two species that were previously lumped together as common skate.

These fish were once distributed from Iceland to Morocco and the Mediterranean, and until at least the 1950s they were a prominent component of bottom-trawl catches[2]. Since then they have disappeared from large parts of their ranges. In the seas between southwestern Britain and Ireland, for example, only six individuals were taken between 1988 and 1997 during bottom-trawl research surveys[3], whereas thousands used to be caught accidentally by fisheries using similar gear[2]. French trawlers were still able to catch several hundred tonnes of common skates each year, mainly in deep waters along the edges of the continental shelf in the northeast Atlantic. But behind the scenes, many fisheries scientists wondered how such a large, slowly reproducing fish could sustain these catches.

Iglésias *et al.*[1] have solved the mystery. There are two species, one of which reaches maturity at about 120 cm, the other at 200 cm (Fig. 1). The apparent sustainability of the larger species (one of the biggest skates in the world) was a mirage produced by continued catches of the smaller one, which reaches maturity at a younger age and is almost certainly more productive. Fish-market surveys by Iglésias and colleagues suggest that when landings of common skates from 2005 are reassigned to the correct species, only 140 adults were of the larger species, compared with 8,300 adults of the smaller one. The greater rarity of the larger species is consistent with expectations from life-history theory, whereby large-bodied species have lower potential rates of reproduction and are therefore less able to sustain exploitation[4].

We should have seen this coming. In 1926, R. S. Clark[5] stated: "I have noted frequently the occurrence of mature males with large claspers, and other equally large males with the claspers quite undeveloped. So far I have not given any special study to this phenomenon." A study in 1968 made a similar point[6]. Iglésias and colleagues have added numerous other differences, including eye colour, the orientation of tail spines, dorsal-fin spacing, the colour of prominent eyespots on the skate wings, and the shape of the teeth. Indeed, the authors used genetic analyses to show that these two species are not even each other's nearest relatives (the 'near threatened' long-nosed skate, *Dipturus oxyrinchus*, sits between them on the evolutionary tree). Formal taxonomic description is still in the works. But these are clearly very different animals in ways that matter greatly to fisheries sustainability, because their different life histories should translate into different rates of population growth[7].

Even before the discovery of this taxonomic lapse, conservation measures for these and many other skates, not to mention their relatives the rays, sharks and chimaeras, have been ineffective. A new European Union (EU) requirement is that common skates when caught should be returned to the sea, but this applies only to the North Sea[8]. And in their last known refuge, the western continental shelf edge, large skates and rays remain unprotected. An overdue step could be to extend the EU regulation to all large skates, including the 'common-skate complex', the long-nosed skate and the white skate (*Rostroraja alba*), across their entire historic distribution.

This cautionary taxonomic tale extends, of course, more broadly. In the marine realm alone, a vast store of cryptic biodiversity remains to be discovered. For example, a third of all sharks and rays have been described only in the past 30 years — a new one is described, on average, every month[9]. Systematics underpins our understanding of biodiversity, both marine and terrestrial, yet taxonomic science is at best underdeveloped and at worst in decline or even in crisis[10]. If the nations of the EU cannot stir themselves to provide an adequate footing for taxonomy and conservation management even for commercially valuable fish, the outlook in less favoured parts of the world is indeed grim. ■

Nicholas K. Dulvy and John D. Reynolds are in the Earth to Ocean Research Group, Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.
e-mails: dulvy@sfu.ca; reynolds@sfu.ca

1. Iglésias, S. P., Toulhoat, L. & Sellos, D. Y. *Aquat. Conserv. Mar. Freshwat. Ecosyst.* doi:10.1002/aqc.1083 (2009).
2. Brander, K. *Nature* **290,** 48–49 (1981).
3. Dulvy, N. K. *et al. Conserv. Biol.* **14,** 283–293 (2000).
4. Reynolds, J. D., Dulvy, N. K., Goodwin, N. B. & Hutchings, J. A. *Proc. R. Soc. Lond. B* **272,** 2337–2344 (2005).
5. Clark, R. S. *Fishery Board for Scotland Scientific Investigations* No. 1 (1926).
6. Du Buit, M. H. *Trav. Fac. Sci. Univ. Rennes Sér. Océanogr. Biol.* **1,** 19–117 (1968).
7. Dulvy, N. K. & Reynolds, J. D. *Conserv. Biol.* **16,** 440–450 (2002).
8. Regulation 43/2009 *Offic. J. EU* (2009).
9. Last, P. R. *Mar. Freshwat. Res.* **58,** 7–9 (2007).
10. Systematics and Taxonomy: Follow-up. House of Lords Sci. Technol. Committee go.nature.com/xdtqXI (2008).

© 2009 Macmillan Publishers Limited. All rights reserved

IMMUNOLOGY

A helpers' guide to infection

Thomas Gebhardt and Francis R. Carbone

Killer T cells were thought to patrol the body unhindered, freely gaining access to sites of infection. But it seems that, at least in some body tissues, helper T cells must pave the way for killer T-cell entry.

Cytotoxic T lymphocytes (CTLs) are the killer white blood cells of the immune system, having crucial roles in the defence against a range of viral, bacterial and parasitic infections. During microbial colonization at peripheral body sites, such as the outer layers of the skin and mucosal epithelium, circulating CD8$^+$ CTLs (cells that carry a CD8 receptor molecule on their surface) exit from blood vessels to access and destroy cells harbouring pathogens. The CTL response begins in the lymph nodes that drain the infection site, and involves a concerted effort by various immune cells. These include CD4$^+$ helper T cells, which promote optimal CTL expansion and functional programming.

Once armed and released into the bloodstream, CTLs seem to have free access to various non-lymphoid organs[1,2]. This suggests that the mere presence of these primed killer cells in the circulation might be sufficient for their entry into infected tissues. On page 510 of this issue, Nakanishi *et al.*[3] reveal that this is not the case, or at least not for selected regions of the body*. The authors describe instead a complex pattern of CTL entry into sites of infection that is orchestrated by the same helper T cells that are involved in the killer cells' initial priming.

To demonstrate this principle, Nakanishi *et al.*[3] used a mouse model of herpes simplex virus (HSV) infection of the vagina, and focused on the infiltration of HSV-specific CTLs into the infected tissue. They show that, as expected[4], priming of CTLs is dependent on CD4$^+$ T-cell help. But mucosal invasion by CTLs lagged behind that of helper T cells, suggesting that the helper cells somehow facilitate CTL entry. The authors formally proved this[3] by transferring activated CTLs from infected mice into one of two groups of recipients — mice that were deficient in helper T cells or mice that had normal numbers of these cells. The activated CTLs entered rapidly into HSV-infected tissue only in recipient animals that had helper T cells, implicating these latter cells in the control of CTL infiltration.

The authors observed that the helper T cells behave like pioneers, paving the way for CTL entry into the mucosa by altering the local micro-environment. The secreted immune modulator interferon-γ (IFN-γ) was a central mediator of these environmental changes. Crucially, IFN-γ controlled the synthesis of small chemokine molecules by the vaginal epithelium, which guide immune-cell entry into tissues. The specific chemokines involved were CXCL9 and CXCL10 (which bind to the chemokine receptor CXCR3).

Nakanishi *et al.*[3] exclude a role for a regulatory subset of T lymphocytes, known as Foxp3$^+$ CD4$^+$ T cells, in the direct control of CTL migration. However, Foxp3$^+$ cells have been shown[5] to promote helper T-cell entry into HSV-infected vaginal tissue and, as a consequence, they could indirectly influence killer T-cell migration.

The authors do not speculate on the mechanism of IFN-γ production by helper T cells, but the most likely scenario involves local recognition of HSV antigen by these cells, which is probably presented by dendritic cells that accumulate under the infected epithelium[6]. Infiltrating dendritic cells have been implicated in driving helper T-cell activation[7] and IFN-γ production[8] in non-lymphoid tissues, and thus may be important co-contributors to the complex local immune response that drives optimal CTL recruitment.

Nakanishi and colleagues propose that body tissues can differ in their accessibility to CTLs. For instance, the vagina seems to be restrictive to CTL infiltration, whereas CTLs apparently enter the lung in an unregulated manner[3]. In humans, HSV also causes skin disease, and studies have described[9] sequential T-cell entry into the infected skin similar to that seen in the vaginal tissue. In addition, helper T cells have been shown[10] to facilitate CTL recruitment into tumours implanted in the skin. Thus, the skin and its surrounds may be another site in which helper T cells control CTL entry. Further studies are required to determine the generality of Nakanishi and colleagues' findings[3], and to define the relative contributions of this helper T-cell-directed migration compared with the innate mechanisms known to promote homeostatic T-cell movement out of blood vessels[11].

In the scenario presented by Nakanishi *et al.*[3], helper T cells and killer T cells have inherent differences in their ability to access tissue: migration of CTLs into the site of infection is dependent on CD4$^+$ T-cell help, with the helpers apparently being less restricted in their access. This may reflect differences in their localization within tissues or in the functions of the respective cell populations. For example, helper T cells might remain largely in the submucosal layer of the vagina where they interact with the infiltrating dendritic cells to produce cytokines such as IFN-γ, thereby promoting CTL infiltration. By contrast, CTLs seem to follow the CXCL9/10 gradient, penetrating into the epithelium to actively rid this region of infection (Fig. 1).

It may be possible to exploit Nakanishi and colleagues' findings[3] for therapeutic purposes, such as enhancing CTL infiltration of tumours, or inhibiting harmful cell accumulation during autoimmunity. At the very least, their study argues that the simple presence of CTLs in the circulation can no longer be considered a guarantee of their access to peripheral locations for successful infection control. Instead, proof

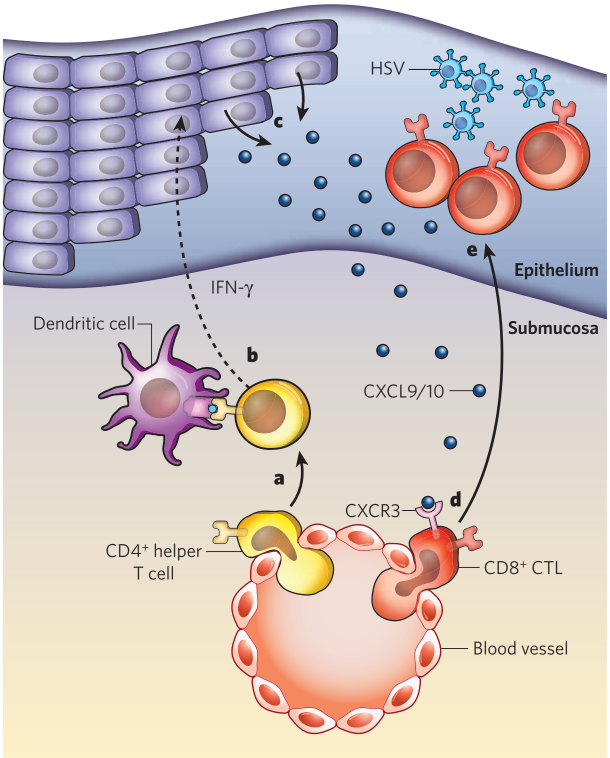


Figure 1 | Model of assisted entry. Nakanishi and colleagues[3] studied a mouse model of infection with herpes simplex virus (HSV). They show that CD4$^+$ helper T cells are required to promote entry of CD8$^+$ cytotoxic lymphocytes (CTLs) into infected vaginal tissue. **a**, Virus-specific helper T cells exit from blood vessels and enter the submucosa of vaginal tissue in response to HSV infection. **b**, These cells probably recognize antigens presented by the dendritic cells that are known to accumulate in the submucosal layer, and they subsequently secrete the immune-modulator interferon-γ (IFN-γ). **c**, In response to IFN-γ, vaginal epithelial cells produce the chemokines CXCL9 and CXCL10. **d**, These bind to CXCR3 receptors expressed on the surface of CTLs. **e**, The IFN-γ-induced chemokine gradient ultimately leads to recruitment of CTLs to the infected epithelium, where they eliminate the virus.

*This article and the paper under discussion[3] were published online on 8 November 2009.

© 2009 Macmillan Publishers Limited. All rights reserved

of actual T-cell infiltration is needed. ■

Thomas Gebhardt and Francis R. Carbone
are in the Department of Microbiology
and Immunology, University of Melbourne,
Parkville, Victoria 3010, Australia.
e-mails: gebhardt@unimelb.edu.au;
fcarbone@unimelb.edu.au

1. Marshall, D. R. *et al. Proc. Natl Acad. Sci. USA* **98,** 6313–6318 (2001).
2. Masopust, D. *et al. J. Immunol.* **172,** 4875–4882 (2004).
3. Nakanishi, Y., Lu, B., Gerard, C. & Iwasaki, A. *Nature* **462,** 510–513 (2009).
4. Jennings, S. R., Bonneau, R. H., Smith, P. M., Wolcott, R. M. & Chervenak, R. *Cell. Immunol.* **133,** 234–252 (1991).
5. Lund, J. M., Hsing, L., Pham, T. T. & Rudensky, A. Y. *Science* **320,** 1220–1224 (2008).
6. Zhao, X. *et al. J. Exp. Med.* **197,** 153–162 (2003).
7. Wakim, L. M., Waithman, J., van Rooijen, N., Heath, W. R. & Carbone, F. R. *Science* **319,** 198–202 (2008).
8. McLachlan, J. B., Catron, D. M., Moon, J. J. & Jenkins, M. K. *Immunity* **30,** 277–288 (2009).
9. Cunningham, A. L. *et al. J. Clin. Invest.* **75,** 226–233 (1985).
10. Marzo, A. L. *et al. J. Immunol.* **165,** 6047–6055 (2000).
11. Butcher, E. C. & Picker, L. J. *Science* **272,** 60–66 (1996).

SOLID-STATE PHYSICS

Silicon spintronics warms up

Michael E. Flatté

Electrical injection and detection of spin-polarized electrons in a silicon chip have now been demonstrated at room temperature, paving the way to the development of low-power semiconductor spintronics circuitry.

Pushing electron charge around a silicon chip drives modern electronics, from supercomputers to mobile phones. Charge distributions have been used for decades to encode digital information, because they are easily tunable as well as surprisingly robust and predictable. Moreover, charges respond rapidly to changes in the voltage of a device, yet quickly relax to the equilibrium configuration required by such voltages. These advantageous properties, however, also mean that a fundamental minimum energy is required to switch a device from an 'on' to an 'off' voltage[1]. As modern chips approach this fundamental limit, the prospect of replacing electron charge with electron spin has become increasingly attractive[2]. On page 491 of this issue, Dash *et al.*[3] describe a dramatic advance in the field of spin-transport electronics (spintronics): the demonstration in silicon at room temperature of the electrical injection and detection of spin-polarized electrons — electrons that all have the same spin orientation.

A decade ago, the demonstration of long-lived coherent spin precession (the rotation of an electron's spin orientation about an axis) in semiconductors at room temperature[4] galvanized research into the use of spintronics for semiconductors[5,6]. To precess coherently, the spins cannot be in thermodynamic equilibrium, and that first experiment[4] showed that the time for which spins can remain out of equilibrium — the spin coherence time — is easily a million times longer than is possible for charge distributions. Switching the spin polarization (Fig. 1), for example from spin 'up' to spin 'down', on the basis of spins out of thermal equilibrium, might then be possible, but would require electrical injection, transport and detection of these coherently precessing spins.

Silicon, the best semiconductor material for charge-based electronics, also seemed a promising choice for spintronics, as the coupling between the electron spin and direction of motion (spin–orbit interaction) in silicon is weak compared with that in many other semiconductors. And as this spin–orbit interaction limits spin coherence times, silicon should have a long coherence time. Successful demonstrations of electrical injection, transport and detection of spins in the semiconductor gallium arsenide[7] were closely followed by those in silicon[8,9], although at low temperatures. Dash *et al.*[3] now achieve spin injection and detection in silicon at room temperature, and control the injected spins by means of a weak magnetic field.

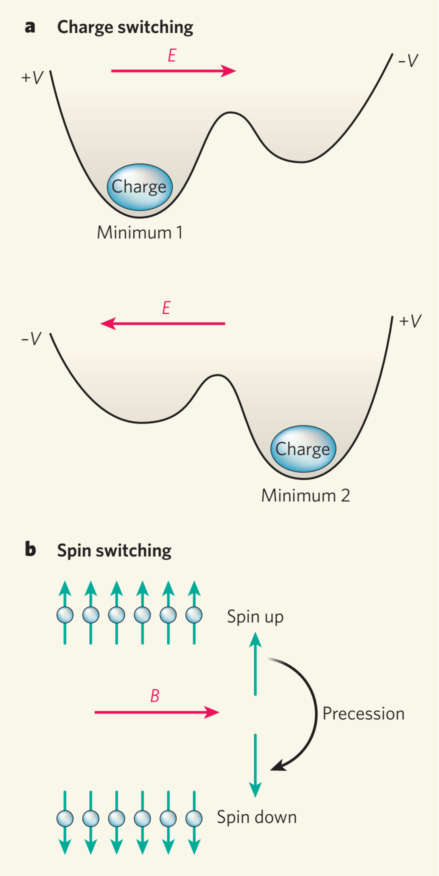


Figure 1 | Charge versus spin switching. a, In conventional electronic switching, which is based on the charge of the electron, an electric field (*E*) moves a charge distribution from one location to another (for example, into the channel of a transistor to permit current to flow). The charge distribution is in thermal equilibrium, and thus an energetic potential barrier (induced by an applied voltage, *V*) that is much larger than thermal energies must be used to keep it in the desired position (potential-energy minimum 1 or 2). **b,** In spintronic switching, which is based on the spin rather than the charge of the electron, the spins remain out of thermal equilibrium for long periods of time, and thus no energetic barrier is required to keep them in the desired polarization (spin up or spin down). Spins can be switched from one polarization to another (for example, spin up to spin down) using a small magnetic field (*B*), which causes the spin orientation to precess coherently. Dash and colleagues[3] demonstrate spin injection, detection and coherent precession in silicon at room temperature — essential steps towards achieving spintronic switching.

The efficiency of the experimental procedure used by Dash and colleagues is considerably higher than that of previous attempts using 'doped' silicon (possessing a background of unpolarized charge carriers in thermal equilibrium). The authors[3] obtained an electron spin polarization of a few per cent, compared with much less than one per cent in previous work with doped silicon[9]; higher percentages have been achieved with undoped silicon[8]. The electrical detection of spin polarization involves measuring the electrical potential at a ferromagnetic contact. The strength of this potential depends on the spin polarization and doping in the semiconductor beneath the contact; the larger spin polarization obtained by the authors in doped silicon led to a much larger detection voltage (millivolts instead of microvolts). Their approach[3] to electrical injection and detection relied on a single electrical contact, which consisted of a ferromagnetic metal electrode and an aluminium oxide interface (barrier) between the silicon and the electrode, for both injection and detection. This interface increases the efficiency of the electrical injection of spin-polarized electrons from the ferromagnet into the silicon[2]. Keeping the current through the contact fixed, the voltage between it and a reference point was measured as a magnetic field was varied. This voltage included a small contribution from the spin polarization under the contact in the silicon, but it was masked by the much larger reference voltage. So how did the authors[3] measure the desired spin polarization in the silicon?

Spins precess in a magnetic field, and if that precession is much faster than the spin coherence time, the spin polarization is greatly reduced. For a sufficiently large magnetic field,

© 2009 Macmillan Publishers Limited. All rights reserved

the spins under the contact randomize and spin polarization is quenched. A measurement of the difference between the voltage at zero magnetic field and at a large magnetic field provided the contribution of the injected spin polarization to the signal at the ferromagnetic contact[3]. The strength of the magnetic field required to quench the injected spin polarization provided a measure of the spin coherence time (as well as the distance over which the electrons remain polarized during that time). Control experiments, which included destroying the spin polarization of the current flowing into the silicon (by having ytterbium in the barrier) or modifying the barrier between the contact and the silicon (by adding caesium), led to negligible spin-injection amplitudes. This demonstrates that the carefully designed contact and barrier are crucial for high-efficiency spin injection.

Dash and colleagues achieved successful spin injection in both 'electron-doped' and 'hole-doped' silicon — the two constituents of complementary metal-oxide-semiconductor (CMOS) technology used in most conventional microelectronic circuitry. One might expect that the resulting spin coherence times measured in doped silicon would correspond to those measured by other techniques. However, the times reported by the authors — 140 picoseconds for electron-doped silicon and 270 picoseconds for hole-doped silicon — are surprisingly short. By comparison, gallium arsenide, which has a spin–orbit interaction tenfold larger than that of silicon, has a room-temperature spin coherence time only threefold smaller[10], about 50 picoseconds. This disconcerting result does not necessarily preclude the use of silicon for spintronic devices, for the distance over which the electrons remain polarized during these times exceeds a couple of hundred nanometres, which is much larger than the expected sizes of devices in modern semiconductor chips. However, it is a surprising result that may require a rethink about the mechanisms of spin decoherence in silicon.

Because of the ubiquitous nature of silicon in modern semiconductor electronics, the demonstration of semiconductor spintronic functionality in silicon at room temperature promises to be a major breakthrough. An observation of room-temperature spin transport between two contacts in silicon, in addition to the injection and detection demonstrated in a single contact by Dash *et al.*, would be a welcome next step. Initial applications may include using spin injection and detection to enhance the performance of predominantly charge-based devices. However, to dramatically reduce the power consumption of modern electronics below the fundamental limit[1], additional advances would be required, especially the control of spin orientation by other means than a magnetic field. ■

Michael E. Flatté is in the Department of Physics
and Astronomy, The University of Iowa,
Iowa City, Iowa 52242, USA.
e-mail: michael_flatte@mailaps.org

1. Landauer, R. *IBM J. Res. Dev.* **5,** 183–191 (1961).
2. Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, 2002).
3. Dash, S. P., Sharma, S., Patel, R. S., de Jong, M. P. & Jansen, R. *Nature* **462,** 491–494 (2009).
4. Kikkawa, J. M. *et al. Science* **277,** 1284–1287 (1997).
5. Wolf, S. A. *et al. Science* **294,** 1488–1495 (2001).
6. Awschalom, D. D. & Flatté, M. E. *Nature Phys.* **3,** 153–159 (2007).
7. Lou, X. *et al. Nature Phys.* **3,** 197–202 (2007).
8. Appelbaum, I., Huang, B. & Monsma, D. J. *Nature* **447,** 295–298 (2007).
9. van 't Erve, O. M. J. *et al. Appl. Phys. Lett.* **91,** 212109 (2007).
10. Meier, F. & Zachachrenya, B. P. (eds) *Optical Orientation* (North-Holland, 1984).

STRUCTURAL BIOLOGY

Highly charged meetings

Anthony G. Lee

When it comes to proteins and their environments, opposites repel. So how is the highly charged, polar helix of a transmembrane ion channel accommodated by a non-polar membrane? Easily, if the charges are buried.

Early in any biochemistry course, students are told that charged amino acids are not happy in hydrophobic (water-repelling) environments. Because the basic unit of biological membranes — the lipid bilayer — has a hydrophobic core, it follows that the α-helices of membrane-bound proteins should rarely contain charged amino acids. But there are exceptions, of which voltage-gated potassium channels form one class. On page 473 of this issue, Krepkiy *et al.*[1] show that, contrary to textbook teachings, the highly charged α-helix present in these ion channels is fully compatible with a normal lipid bilayer.

Voltage-gated potassium channels are homotetramers — they assemble from four identical monomers, each of which contains six membrane-spanning α-helices. Two of the helices (S5 and S6) from each monomer come together in the tetrameric structure to form the pore through which potassium ions move across the membrane (Fig. 1). The remaining helices (S1 to S4) form voltage sensors, one for each monomer. A vital role is played by helix S4, which contains four or five positively charged amino-acid residues. It is these positive charges that allow the channel to sense a change in electrical potential across the membrane; subsequent movement of S4 leads to opening of the channel. But how can such a charged

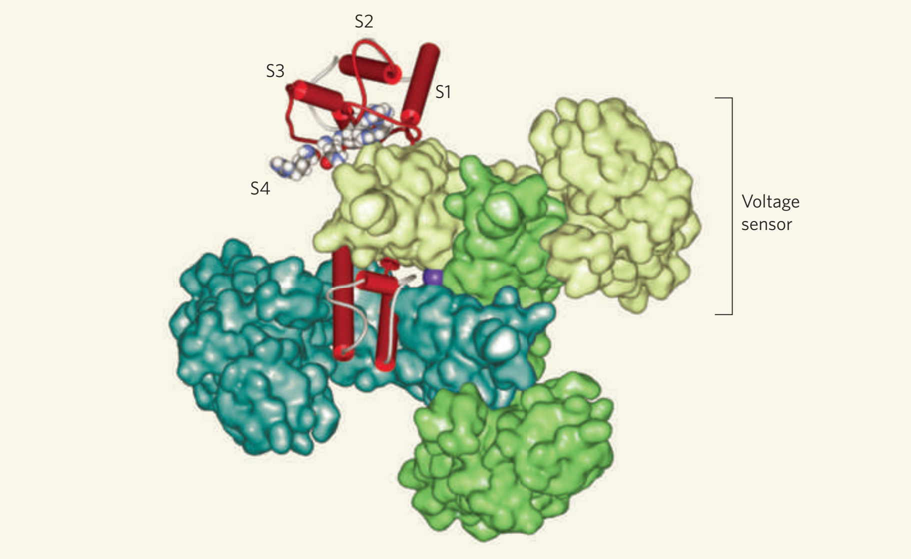


Figure 1 | Structure of a voltage-gated potassium channel. Voltage-gated potassium channels form pores in cell membranes through which potassium ions pass. The channel depicted here is a Kv2.1/Kv1.2 chimaeric channel[9], viewed as though looking down on the membrane from the extracellular side (membrane not shown). Three of the four subunits are shown as surface plots in different shades of green, whereas the fourth subunit is shown in cartoon format. Most of the α-helices in the fourth subunit are depicted as red cylinders, but the first five positively charged amino-acid residues in helix S4 are shown in space-filling format. A potassium ion moving through the central pore is shown in purple. The channel has four voltage sensors (each composed of helices S1–S4) that are loosely attached to the central pore. The charged residues on S4 point in towards the other helices of the voltage sensor, and most are located in a water-filled cavity (not shown). Krepkiy and colleagues[1] report that this means that the highly charged voltage sensor is fully compatible with the lipid bilayer that surrounds it in the native membrane.

© 2009 Macmillan Publishers Limited. All rights reserved

helix be accommodated in the hydrophobic environment of a lipid bilayer?

Crystal structures of voltage-gated potassium channels have shown that the voltage sensors are only loosely attached to the central pore (Fig. 1, overleaf). In the first published structure[2], which is recognized to be a distortion of the naturally occurring structure, the positive charges in S4 are partly exposed on the outer surface of the voltage sensor and are possibly in contact with the lipid bilayer. To see what effect such an exposed helix might have on a membrane, a computer model was generated of an isolated S4 helix in a lipid bilayer[3]. The model showed that the helix greatly distorts the bilayer, leading to the formation of hydrogen-bonded networks of water and lipid phosphate groups about each charged residue in the helix. The model also showed that the thickness of the bilayer's hydrophobic core close to the helix falls sharply from its normal value of 27 Å to about 10 Å. Such effects would be very unusual for a protein in a membrane because of the high energetic cost of distorting the bilayer[4], and because the hydrophobic match between an undistorted lipid bilayer and a membrane protein is usually rather good[5].

But does a complete voltage sensor, made up of helices S1 to S4, have the same effect on a lipid bilayer as an isolated S4 helix? The answer turns out to be no. Using neutron-diffraction techniques, Krepkiy and colleagues[1] studied the properties of the bilayer surrounding a voltage sensor from a bacterial ion channel. They show that the bilayer remains intact and that it is only about 3 Å thinner than its normal thickness.

The authors' neutron-diffraction experiments measured the average thickness of the bilayer across the whole membrane plane. It is therefore possible that the bilayer close to the voltage sensor is thinner than the measured average. To find out whether this is the case, Krepkiy *et al*. performed molecular-dynamics simulations of their system. These models did indeed suggest that the thickness of the bilayer decreases by as much as 9 Å close to the voltage sensor. This conclusion must be treated with caution, however, because the authors' simulation also predicted that the unperturbed bilayer is about 4 Å thicker than the experimentally determined thickness[6] (such models often overestimate the thickness of the bilayer). An earlier, coarse-grained molecular-dynamics simulation suggested that the effects of voltage sensors on bilayer thickness are small[7].

The crystal structure of the potassium channel shown in Figure 1 suggests a simple reason for the small effect that potassium-channel voltage sensors have on membranes: the charges on helix S4 are not actually exposed to the lipid bilayer. Instead, they are buried within the sensor structure, either occupying water-filled cavities or interacting with negatively charged residues in the S2 helix[8]. To confirm this, the authors performed further experiments that showed that the presence of the voltage sensor results in no measurable change in the amount of water in the lipid bilayer's core. The sensor is hydrated, however, as revealed by Krepkiy and colleagues' nuclear magnetic resonance studies[1]. The water molecules are probably located in crevices in the sensor, where they can hydrate some of the positively charged residues in S4, ensuring that these residues remain charged and so able to detect changes in potential across the membrane.

Overall, Krepkiy and colleagues' study is rather comforting — maybe the exceptions to the rules taught to biochemistry students aren't really exceptions after all. As with other membrane-bound proteins[5], voltage-gated potassium channels must have evolved so that the packing preferences of the helices in the voltage sensor cause the sensor to adopt a structure that nicely matches that of the surrounding lipid bilayer. In this way, the hydrophobic lipid and the hydrophilic, charged sensor can meet without either having to change very much. ■

Anthony G. Lee is in the School of Biological Sciences, University of Southampton, Southampton SO16 7PX, UK.
e-mail: agl@soton.ac.uk

1. Krepkiy, D. *et al. Nature* **462,** 473–479 (2009).
2. Jiang, Y. *et al. Nature* **423,** 33–41 (2003).
3. Freites, J. A., Tobias, D. J., von Heijne, G. & White, S. H. *Proc. Natl Acad. Sci. USA* **102,** 15059–15064 (2005).
4. Marsh, D. *Biophys. J.* **94,** 3996–4013 (2008).
5. Lee, A. G. *Biochim. Biophys. Acta* **1612,** 1–40 (2003).
6. Nagle, J. F. & Tristram-Nagle, S. *Biochim. Biophys. Acta* **1469,** 159–195 (2000).
7. Bond, P. J. & Sansom, M. S. P. *Proc. Natl Acad. Sci. USA* **104,** 2631–2636 (2007).
8. Chakrapani, S., Cuello, L. G., Cortes, D. M. & Perozo, E. *Structure* **16,** 398–409 (2008).
9. Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. *Nature* **450,** 376–382 (2007).

ASTROPHYSICS

Assortment in the Galaxy

Judith G. Cohen

Observations of star clusters in the Milky Way defy the view that the constituents of these systems are almost invariably chemically alike. The outlying clusters could be the tattered relics of once larger systems.

In its halo of dark matter, our Galaxy hosts a family of about 150 globular star clusters (GCs). Conventional wisdom holds that they are compact, roughly spherical systems of high stellar density, each containing about 5 million stars held together by gravity. Undergraduates are taught that these classic laboratories for studying stellar evolution each contain a single population of stars of uniform age and chemical composition. More than 30 years ago, it became clear that the most luminous of these clusters, ω Centauri, was the exception to the rule: the system contains stars with a range of iron abundances (Fe metallicity) that vary by more than a factor of 30 (refs 1, 2). In this issue, Lee *et al.*[3] (page 480) and Ferraro *et al.*[4] (page 483) report the discovery of two other GCs that harbour stars containing different proportions of iron and other heavy elements.

Variations among the light elements within individual GCs were also discovered several decades ago. But, unlike heavy elements, light elements can be made during the course of normal stellar evolution in intermediate-mass stars through the fusion of hydrogen at high temperatures[5]. The resultant 'ash' could be mixed into the surfaces of these evolved stars, ejected by gentle winds and then mixed into the gas in the young cluster. A second generation of stars could then be formed, giving rise to the observed variation in light-element content within a GC.

Heavier elements — including calcium, iron and beyond — are mostly produced in stellar explosions known as supernovae. Because material is violently ejected from supernovae at a very high velocity and the gravitational binding energy of present-day GCs is low, in the current conditions supernova gas ejecta would escape from the cluster. The only way in which such energetic gas, rich in heavy elements, could have been retained would be if the mass of the GC was much higher in the past than is typical today. If we find a GC showing variations in those heavy elements, suspicion naturally arises that it is the remnant of a formerly accreted small galaxy, as was suggested for ω Centauri. This is not wild speculation; there are indications that the GC called M54 will probably be the only remnant structure from the Sagittarius dwarf galaxy to survive the galaxy's ongoing violent disruption by the Milky Way.

We now have much better tools with which to search for variations in age and elemental abundance within individual Galactic GCs. These tools operate at a level of accuracy that we could only dream of a decade ago. Lee *et al.*[3] demonstrate definitively that there is a spread in the abundance of calcium within the massive GC M22, which has been a suspect for many years. They find that the population of red-giant branch stars — stars in which the core has ceased to burn hydrogen but the outer shell is still doing so — in the system splits

© 2009 Macmillan Publishers Limited. All rights reserved

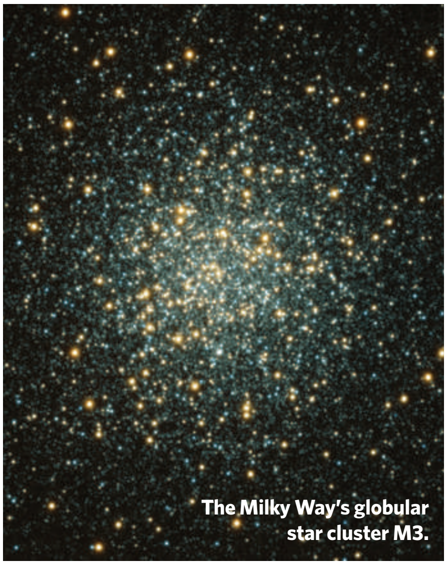
into two subpopulations of different calcium abundances. Two very recently completed spectroscopic studies[6,7] detect star-to-star variations in iron abundance in M22 for smaller samples of red giants. It seems that M22 will join M54 as the only remnant of the disruption of an entire dwarf galaxy in the halo of the Milky Way.

Lee and colleagues[3] go further, claiming that they can detect multiple stellar populations with smaller but still statistically significant variations in calcium abundance in more than half of the systems in their sample of 37 GCs. This is the most interesting and controversial part of their paper because, if they are correct, many GCs — not just a few outliers — must be pathetic remnants of much more massive systems that were accreted by the Milky Way halo during its formation. Although the authors' case for the system NGC 1851 seems reasonably secure, their claims for other GCs seem to be only marginally significant, and will require further confirmation. A previous investigation[8] has already ruled out variations exceeding 12% in Fe metallicity for the majority of the eight GCs that have been studied in detail by Lee *et al.*, demonstrating yet again that there is a high degree of uniformity in the abundance of Fe in most GCs throughout the stellar population.

Analysis of the current generation of high-quality images of GCs, whether taken by the Hubble Space Telescope or with ground-based telescopes equipped with adaptive-optics systems, has allowed exquisite data to be gathered for thousands of stars, and has enabled the discovery in GCs of subtle phenomena that previous studies missed. The GC NGC 1851 was found to have two branches of subgiant stars where there should just have been one[9]. And Piotto and colleagues[10] found that main-sequence stars — those in which energy is created through the fusion of hydrogen in the star's core — in the GC NGC 2808 are divided into three separate branches.

To this collection of abnormalities we can now add the discovery of two subgroups of horizontal branch stars (those that are powered by the fusion of helium in the core) in the GC Terzan 5 that is presented by Ferraro and colleagues[4]. This particular anomaly has never previously been seen in a Galactic GC. The authors[4] have also obtained spectra of a few horizontal branch stars in Terzan 5 that demonstrate that Fe metallicity varies by about a factor of three within this GC. So Terzan 5 must be yet another tattered remnant of a once much more massive system.

Potential causes for the bizarre behaviour of these GCs include helium-content variations (which must exist as a result of the same hydrogen-burning process that gives rise to variation among the observed light elements, but helium is extremely difficult to detect), age differences, and variations among the heavy elements. Another possibility, which was previously suggested[11] to explain the peculiar case of NGC 1851, is extremely large variations among the light elements (particularly carbon, nitrogen and oxygen, the most abundant of these). All of these possibilities can also occur in combination, adding to the confusion. We know that age variations within GC systems are small, but of the order of 10%[12]. D'Antona and Ventura[13] suspect that, in some cases, very high helium abundances (up to 40%) are required to reproduce some of the observed irregularities. This is almost twice the primordial abundance of helium produced in the Big Bang, the relic of which is found in present-day, metal-poor stars, and there is no direct observational evidence to support such a high helium abundance in any GC.



The Milky Way's globular star cluster M3.

S. KAFKA & K. HONEYCUTT, INDIANA UNIV./WIYN/NOAO/NSF

As we look closer and with more precision, the model of the GCs in the Milky Way as simple, single stellar population systems is being severely challenged. Are the anomalies, which seem to be turning up with increasing frequency, confined only to the most massive of the Galaxy's GCs? Exactly how common and how big such deviations from uniformity are among the Milky Way's GCs, and how they relate to stellar streams in the halo, is a hot topic. ■

Judith G. Cohen is at the Palomar Observatory, California Institute of Technology, Pasadena, California 91125, USA.
e-mail: jlc@astro.caltech.edu

1. Freeman, K. C. & Rodgers, A. W. *Astrophys. J.* **201,** L71–L74 (1975).
2. Norris, J. & Bessell, M. S. *Astrophys. J.* **201,** L75–L79 (1975).
3. Lee, J.-W., Kang, Y.-W., Lee, J. & Lee, Y.-W. *Nature* **462,** 480–482 (2009).
4. Ferraro, F. R. *et al. Nature* **462,** 483–486 (2009).
5. Denisenkov, P. A. & Denisenkova, S. N. *Astron. Tsirk.* **1538,** 11 (1989).
6. Da Costa, G. S. *et al. Astrophys. J.* **705,** 1481–1491 (2009).
7. Marino, A. F. *et al. Astron. Astrophys.* (in the press).
8. Carretta, E., Bragaglia, A., Gratton, R., D'Orazi, V. & Lucatello, S. *Astron. Astrophys.* (in the press).
9. Milone, A. P. *et al. Astrophys. J.* **673,** 241–250 (2008).
10. Piotto, G. *et al. Astrophys. J.* **661,** L53–L56 (2007).
11. Ventura, P. *et al. Mon. Not. R. Astron. Soc.* (in the press).
12. Marín-Franch, A. *et al. Astrophys. J.* **694,** 1498–1516 (2009).
13. D'Antona, F. & Ventura, P. *Mon. Not. R. Astron. Soc.* **379,** 1431–1441 (2007).

50 & 100 YEARS AGO

50 YEARS AGO

The appearance of Radiocarbon Supplement Vol. 1 of the *American Journal of Science* marks an important step forward in the publication of radiocarbon dates. In the past, date lists have appeared at irregular intervals in a number of journals, making it difficult for potential users of the dates to keep themselves fully informed of all the work in this field ... The editors ... are to be commended on this project, which provides a single, relatively inexpensive, annual publication specifically for radiocarbon dates and associated measurements. This first volume contains 13 date lists and one paper ... devoted entirely to measurements on samples of known age. This aspect of radiocarbon dating research ... yields information on the past and present distribution of radiocarbon in the carbon exchange reservoir, and this is of particular importance when one is concerned with the attainment of the highest possible accuracy in radiocarbon dates.

From *Nature* 28 November 1959.

100 YEARS AGO

A flying-fish flew on to the lower deck last night about 8.30p.m. The deck is 20 feet above the water-line, and the railing is 4 feet 6 inches above the deck, but it is possible for it to have flown through the railing; the fish measured 17¼ inches from tip of nose to tip of tail. I forgot to weigh it before it was cooked. It was the largest flying fish I have ever handled. Could any reader of NATURE kindly inform me what is the largest size known? We were about fifty miles north of Teneriffe when it came on board. The species up here appear to be larger than those in the tropics and near South America. I have seen large ones in the Gulf of Aden, but never caught one, though I am inclined to think this was a larger species. The longest flyers always appear to be the largest fish: the longest flight I have seen has been about 400 yards.

From *Nature* 25 November 1909.

© 2009 Macmillan Publishers Limited. All rights reserved

OBITUARY

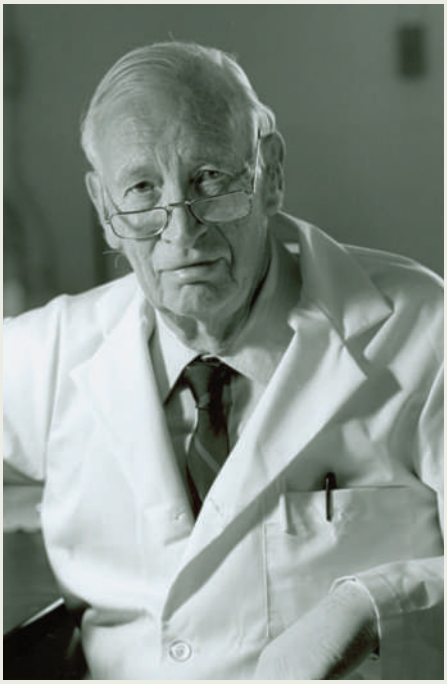
Paul C. Zamecnik (1912–2009)

Trailblazer in the study of protein synthesis.

In April 1940, as low-flying German planes dropped leaflets over Copenhagen announcing that Denmark was now occupied, a young American on a travelling fellowship at the Carlsberg Laboratories realized that he would have to leave. But that young man, Paul Zamecnik, had already absorbed much from the lab's director, the biochemist Kaj Linderstrøm-Lang. Back in the United States, Zamecnik's second try for a post with Max Bergmann at the Rockefeller Institute in New York City was successful, and in due course he took charge of his own group at the Massachusetts General Hospital in Boston. There he revolutionized the study of protein biosynthesis through the use of a cell-free (*in vitro*) system, thus revealing the enzymatic activation of amino acids, the ribosome as the site of peptide-bond formation and the existence of transfer RNA. Zamecnik died at his home in Boston on 27 October.

Zamecnik grew up in Cleveland, Ohio, and at 16 went to Dartmouth College in Hanover, New Hampshire. Graduation from Harvard Medical School in 1936 was followed by an internship back in Cleveland and then a residency at Mass General. His interest in research was apparent early on, as was a strong degree of self-determination. Indeed, even as a medical resident, Zamecnik travelled to New York City to seek a position with Bergmann, who replied that he took only organic chemists. But Zamecnik was prescient to try, for at the time this was virtually the only US laboratory working on protein synthesis.

In 1942, Zamecnik returned from the Rockefeller to Mass General, where, under the freedom granted by Joseph Aub, he collaborated on such projects as a study with Fritz Lipmann of the mechanism of α-toxin produced by *Clostridium* bacteria. But by now, Zamecnik was emerging as a scientific leader in his own right, and a magnet for talent. One to join him was Robert Loftfield. With the recent availability of a long-half-life radioisotope of carbon, Loftfield used the hydrogen-cyanide-based Strecker synthesis to produce the radiolabelled amino acid ^{14}C-alanine. Employing a cell-free system, and with key contributions by Philip Siekevitz, Zamecnik's group obtained the first definitive data showing protein synthesis in a test tube, including compelling evidence that the labelled amino acid was located internally in the product chain. Soon Mahlon Hoagland joined Zamecnik's group, and using the same system discovered the enzymatic activation of amino acids via the formation of acyl anhydrides with adenylate, conceptually sparked by Hoagland's previous postdoc work with Lipmann, upstairs at Mass General.



MERCER PHOTOGRAPHY/UMASS MED. SCHOOL

While Hoagland was working on amino-acid activation, Zamecnik was puzzling over an odd observation in his own experiments. He had seen that ^{14}C-ATP became covalently bound to endogenous RNA in his system, hinting that RNA synthesis might be occurring (this was well before the discovery of RNA polymerase, the enzyme now known to produce RNA). As a control, he had run a separate reaction with ^{14}C-valine and observed that it too became attached to RNA before ending up in protein. Hoagland and others in the lab worked through this puzzle and discovered that a low-molecular-weight RNA fraction behaved as an intermediate in the movement of amino acids from their ATP-activated state into protein.

This was the discovery of transfer RNA, which had been brilliantly predicted by Francis Crick about two years earlier in a conference talk and an unpublished short manuscript that had not reached the Zamecnik group. Crick was said to be elated by the arrival of hard data, whereas Zamecnik had never considered trusting anything less.

In the 1960s, Zamecnik turned his attention to Rous sarcoma virus (RSV), the RNA genome of which is 'reverse transcribed' into DNA for virus replication. His lab started sequencing the region just inside one end, the 3′ end, of the genome. Across the Charles River, at Harvard's main campus, Walter Gilbert and Allan Maxam were sequencing in from the 5′ end using their faster method. The results were soon at hand in both camps — the sequences at both ends were the same, and had the same polarity. From this it became apparent that, during reverse transcription, the 5′ end of the DNA product strand would be complementary to the template RNA's 5′ end and might form a circle with it.

From this, Zamecnik astutely envisaged that blocking circularization might be an antiviral approach. He used a 13-base-long oligodeoxynucleotide complementary to the terminal repeats of RSV to inhibit the translation of viral messenger RNA in a cell-free system and, more momentously, in RSV-infected cells (*Proc. Natl Acad. Sci. USA* **75,** 280–284, 285–288; 1978). These two papers by Zamecnik and M. L. Stephenson launched the era of antisense DNA, which became widely adopted as a powerful tool for experimentally silencing gene expression, almost two decades before the advent of gene silencing by exogenous small interfering RNAs. Meanwhile, commercial efforts began (and continue) to move antisense DNA into clinical application.

Zamecnik's many honours included the US National Medal of Science in 1991 and, in 1996, only the second Albert Lasker Award for Special Achievement in Medical Science to be conferred. Shakespeare (whom Zamecnik could quote extensively from memory) wrote in *Hamlet*:"What a piece of work is a man ... how express and admirable!" Zamecnik was express: he was supremely articulate, and also in a hurry (but too much the gentleman to ever show it). And he was admirable: he always attracted crowds at scientific or social events, not least because of his talent as a storyteller; and as a colleague of his for many years, I often saw him conversing with janitors and other 'sub-faculty' staff, by whom he was held in the same high regard as he was by his scientific peers. He was a most likeable man, a lab-bench perfectionist, and a high-affinity group leader. Medicine, at least on the wards, was not for him, and molecular biology has been the better for that.

Thoru Pederson
Thoru Pederson is in the Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.
e-mail: thoru.pederson@umassmed.edu

© 2009 Macmillan Publishers Limited. All rights reserved

Cover illustration
Artwork by N. Spencer

BIOMATERIALS

Biomaterials research has come of age. Since antiquity, humans have been taking whatever substances are at hand — natural materials, glass, metals or polymers — and using them to replace body parts that have been damaged by disease or injury. But it is only recently, with the advent of molecular biology, that the field has become interdisciplinary, enabling materials scientists to design materials that impart a specific biological function. The field of biomaterials is also broadening as we improve our understanding of how the physical sciences can help to explain biology and indeed of how biological principles, mechanisms and molecules can be applied in the design of materials for non-biological applications.

This Insight explores areas of research in which recent advances in basic biology are driving materials scientists to think differently when developing new materials. Biomaterials science encompasses a huge body of research work, and the Overview article captures the big picture of the field. The Reviews focus on several of the many emerging areas of research that have yielded exciting advances and should continue to do so in the foreseeable future. The design of materials to interact with stem cells and with the immune system is explored, as well as how the fields of arachnid and plant biology could lead to new biologically inspired technologies. We are pleased to highlight that the authors of these Reviews span the physical and biological sciences, celebrating the interdisciplinary nature of the topics covered.

Finally, a Perspective article discusses the challenges that biologists are presenting to materials scientists to design tools for better biodiagnostics. The extreme sensitivity of future devices could, in turn, lead to new challenges for biologists, in how biological processes and disease states are quantified.

Rosamund Daw, Senior Editor, *Nature*
Stefano Tonzani, Associate Editor, *Nature Communications*

Editor, *Nature*
Philip Campbell
Publishing
Nick Campbell
Claudia Deasy
Insights Editor
Karl Ziemelis
Production Editor
Davina Dadley-Moore
Senior Art Editor
Martin Harrison
Art Editor
Nik Spencer
Sponsorship
Amélie Pequignot
Reya Silao
Production
Jocelyn Hilton
Marketing
Elena Woodstock
Emily Elkins
Editorial Assistant
Emma Gibson

OVERVIEW

REVIEWS

PERSPECTIVE

nature insight

© 2009 Macmillan Publishers Limited. All rights reserved

Inspiration and application in the evolution of biomaterials

Nathaniel Huebsch[1,2] & David J. Mooney[1,3]

Biomaterials, traditionally defined as materials used in medical devices, have been used since antiquity, but recently their degree of sophistication has increased significantly. Biomaterials made today are routinely information rich and incorporate biologically active components derived from nature. In the future, biomaterials will assume an even greater role in medicine and will find use in a wide variety of non-medical applications through biologically inspired design and incorporation of dynamic behaviour.

Humankind's use of materials to augment or repair the body dates to antiquity, when natural materials such as wood were used in an attempt to structurally replace tissues lost to disease or trauma (Fig. 1a). Historically, selection of material was based on availability and the ingenuity of the individual making and applying the prosthetic. In the early part of the twentieth century, naturally derived materials began to be replaced by synthetic polymers, ceramics and metal alloys, which provided better performance, increased functionality and more reproducibility than their naturally derived counterparts. These advances led to a pronounced increase in the range of use and the efficacy of biomaterials, as a result of which millions of lives have been saved or improved by devices such as vascular stents, dental restoratives, artificial hips (Fig. 1b) and contact lenses. On the basis of their application, biomaterials were defined as types of material used in a medical device, and the academic foundation of the field lay in materials science and classical engineering. Materials were desired to perform largely mechanical functions: to prevent biological rejection, which hampered device performance and patient health[1], it was preferable that they be 'inert' and not interact with the biology of the host organism. Early research and fortuitous accidents linking materials chemistry to biological response provided a rational basis for developing biologically inert substrates and provided a scientific foundation for biomaterials as an intellectually distinct discipline[1,2].

The molecular biology revolution of the 1970s and advances in genomics and proteomics in the 1990s and 2000s, however, significantly affected the ways in which biomaterials are designed and used. As specific molecules have been implicated in clinically important processes (for example bone morphogenetic protein in osteogenesis), they have been incorporated into materials as bioactive components[1]. Such combination products, which interface directly with cells and tissues through well-defined molecular pathways to direct biological responses, already represent the state of the art of commercial products such as drug-eluting vascular stents (Fig. 1c). One of these products, Medtronic's INFUSE Bone Graft device (Fig. 1d), which combines synthetic components with bone morphogenetic protein, accounted for more than US$760 million in sales in 2007 (ref. 3) and is probably near the billion-dollar mark today.

The increasing importance of biomaterials in our society over the past decades can be tracked in a number of ways, including the growth of biomaterials both as an academic discipline and as an important industry. There has been a precipitous increase in scientific publications in the biomaterials field over the past 30 years, and although biomaterials was historically a focus of study in a very small number of schools, the field

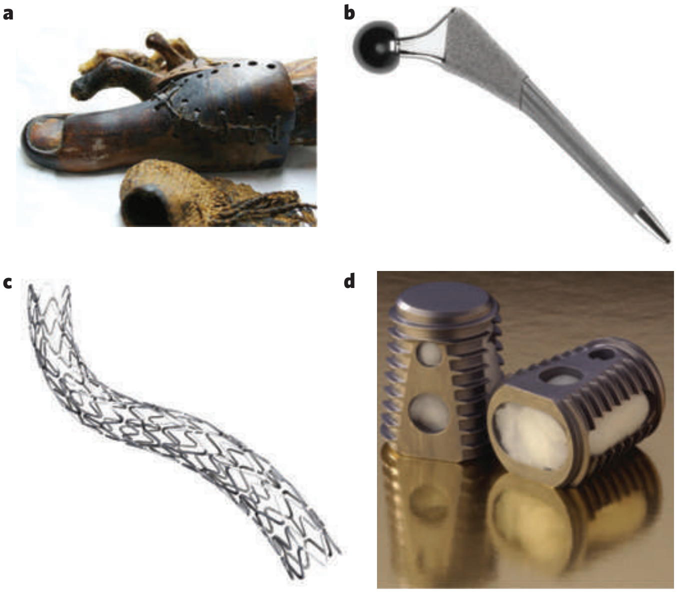


Figure 1 | History and growth of biomaterials as a field and industry. **a**, Prosthetics fashioned from natural materials: wooden toe, circa 1065–740 BC, used as a prosthetic to replace an amputated toe and identified in an anthropological excavation of the Thebes West tombs, Egypt. (Image courtesy of J. Finch, KNH Centre for Biomedical Egyptology, University of Manchester, UK, and The Egyptian Museum, Cairo.) **b**, The SYNERGY hip implant is an example of a state-of-the-art prosthetic device that uses synthetic materials fabricated and engineered to meet performance demands. (Image courtesy of Smith & Nephew, London.) **c**, **d**, Commercially available combination products with both synthetic components and biological activity. **c**, The TAXUS Express[2] Atom Stent, a metal stent from which paclitaxel is eluted into small coronary vessels to prevent restenosis (cell-mediated narrowing of the vessels). (Image courtesy of Boston Scientific Corporation, Massachusetts.) **d**, The INFUSE Bone Graft device, a combination product that uses both traditional prosthetic components (a steel cage) and a tissue-engineering approach (a bovine type I collagen sponge from which recombinant human bone morphogenetic protein 2 is eluted) to provide stability while spinal tissues are being regenerated. (Image courtesy of G. K. Michelson and Medtronic, Burlington, Massachusetts.)

[1]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, 319 Pierce Hall, Cambridge, Massachusetts 02138, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [3]Wyss Institute for Biologically Inspired Engineering, Cambridge, Massachusetts 02138, USA.

© 2009 Macmillan Publishers Limited. All rights reserved

has expanded markedly in the past 20 years. In the United States alone, there are now more than 75 departments of biomedical engineering (12 existed in 1975), with more than 16,000 enrolled students in 2005 (compared with ~3,000 in 1979). Biomaterials is a major field of study in these programmes and is increasingly being emphasized in other engineering departments. In addition, biomaterials are a critical component of a number of industries, including medical devices, dental restoratives and drug delivery, and are increasingly used in technological applications such as *in vitro* diagnostics. Together, these applications generate a market of about $200 billion per year in the United States (as of 2007), with a robust annual growth rate of ~9% (refs 4, 5).

The ability to engineer biological activity into synthetic materials greatly increases the number of their potential uses and improves their performance in more traditional applications. Moreover, the increasing appreciation of the functionality and complexity of biological systems has caused biomaterials researchers to again consider nature for design inspiration. Unlike most man-made materials, materials used in living systems are frequently multifunctional and dynamic, and are built using 'bottom-up' fabrication methods. Both the materials themselves and the biophysical processes involved in their formation are inspiring the design and synthesis of new types of synthetic material that are potentially useful in a wide range of medical and non-medical applications. This widening of the classic view of biomaterials demands an intellectual shift in how these materials are defined. Distinct aspects of this transition in the biomaterials field, and the potential impact on medicine and other industries, are our focus here. We review the current state of the biomaterials field in terms of several major areas of application and design principles, and then we describe emerging and future trends in biomaterials.

Current goals and trajectory of the biomaterials field

The field of biomaterials is in the midst of a revolutionary change in which the life sciences are becoming equal in importance to materials science and engineering as the foundation of the field. Simultaneously, advances in engineering (for example nanotechnology) are greatly increasing the sophistication with which biomaterials are designed and have allowed fabrication of materials with increasingly complex functions. Such sophisticated materials are often designed to mimic a subset of the physicochemical properties of natural materials. Increasingly, nature inspires not only the materials themselves but also the means by which they are made. Whereas synthetic materials are typically engineered on the scale of millimetres or larger and then milled to have micrometre-scale or nanometre-scale features, natural materials are constructed on these smaller scales by self-assembly, a bottom-up means of fabrication that facilitates the construction of information-rich, complex structures in a highly reproducible manner with minimal energy input[6].

Knowledge gained from fundamental studies is being used in conjunction with fabrication methods such as self-assembly to design biomaterials that interface with the biology of the host. This is typically done by means of binding interactions with cell surface receptors[7], to regulate the maintenance, regeneration or even destruction of specific tissues in the body. Key aspects of this line of research include the following: the rich information content of new materials that mimic cellular and extracellular materials, with particular emphasis on presentation of signals in a controlled spatiotemporal manner; provision of non-chemical (for example electrical or mechanical) signals to elicit structural changes in the material or to manipulate cell fate directly; the finding that the physical properties of the materials are probably just as important as their chemistry in terms of the biological response they elicit; and the notion that materials can be designed to regulate host biology at a distance, either by controlling cell trafficking or by trafficking of the material itself in the body.

Biomimetic medical materials and devices

Historically, biological interactions with the host were regulated by the layer of serum proteins adsorbed nonspecifically on surfaces of synthetic materials. A considerable body of research exists on how surface chemistry and topography affect the adsorption of extracellular matrix (ECM) proteins and the presentation of cell-adhesion ligands[2]. However, it is difficult to engineer the surface of materials to adsorb a precise mixture and arrangement of ECM proteins, and those which initially adsorb may be denatured or displaced. Hence, a current theme in biomaterials design is the combination of synthetics that resist nonspecific protein adsorption and molecular components that regulate host biology in a well-defined manner[1].

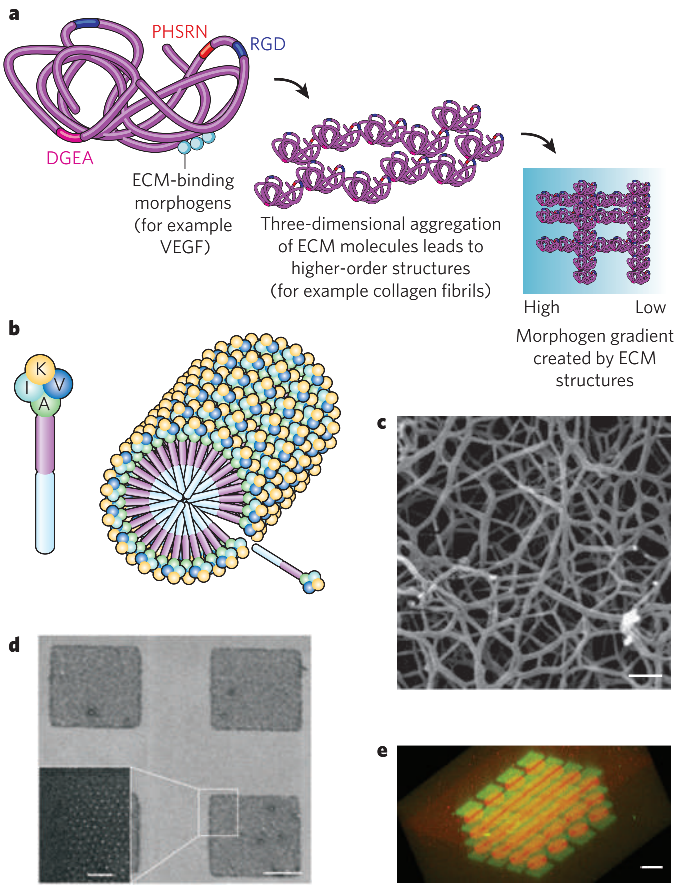

a
PHSRN
RGD
DGEA
ECM-binding morphogens (for example VEGF)
Three-dimensional aggregation of ECM molecules leads to higher-order structures (for example collagen fibrils)
High
Low
Morphogen gradient created by ECM structures
b
K
V
I
A
c
d
e

Figure 2 | Information-rich biomimetic materials. a, Schematic of the natural ECM across different spatial scales. The ECM contains a variety of peptide epitopes (coloured rectangles, labelled with amino-acid sequences of the epitopes) that facilitate integrin-mediated adhesion and other receptor-linked functions. These epitopes are organized in a specific pattern on the nanometre scale within each protein molecule (left) and on the micrometre scale in fibrillar and other structures (centre). The ECM may also regulate the diffusion of soluble proteins, mediating gradients of morphogens between cells on larger length scales (millimetres) (right); the blue colour scale represents one such gradient, with the concentration (from high to low) of morphogen (for example vascular endothelial growth factor (VEGF)) proportional to intensity. **b–e**, Synthetic mimics of the information-rich natural ECM. **b**, **c**, Schematic (**b**) and scanning electron micrograph (**c**) analysis of modularly designed peptide amphiphiles that self-assemble into nanofibres presenting a high density of neural-progenitor-binding epitopes (labelled with the amino acids K, V, I and A). Scale bar, 300 nm. (Image reproduced, with permission, from ref. 9). **d**, Scanning electron micrograph analysis of a surface containing micropatterned islands presenting RGD (adhesive) ligands (white dots) with precisely controlled nanometre-scale spacing. Scale bars, 1 μm (right) and 200 nm (inset, left). (Image reproduced, with permission, from ref. 10.) **e**, A micrograph of a biomaterial modified to present gradients or other complex spatial patterns of morphogens. The fluorescent dyes Alexa Fluor 488 maleimide (green squares) and Alexa Fluor 546 maleimide (red circles) are placed to demonstrate the spatial precision with which bioactive moieties (for example morphogens) could be patterned. Scale bar, 60 μm. (Image reproduced, with permission, from ref. 11.)

Inspiration for the design of new biomaterials has been derived from structure–function analysis on various length scales of the extracellular materials that cells use to organize themselves into tissues (Fig. 2a).

© 2009 Macmillan Publishers Limited. All rights reserved

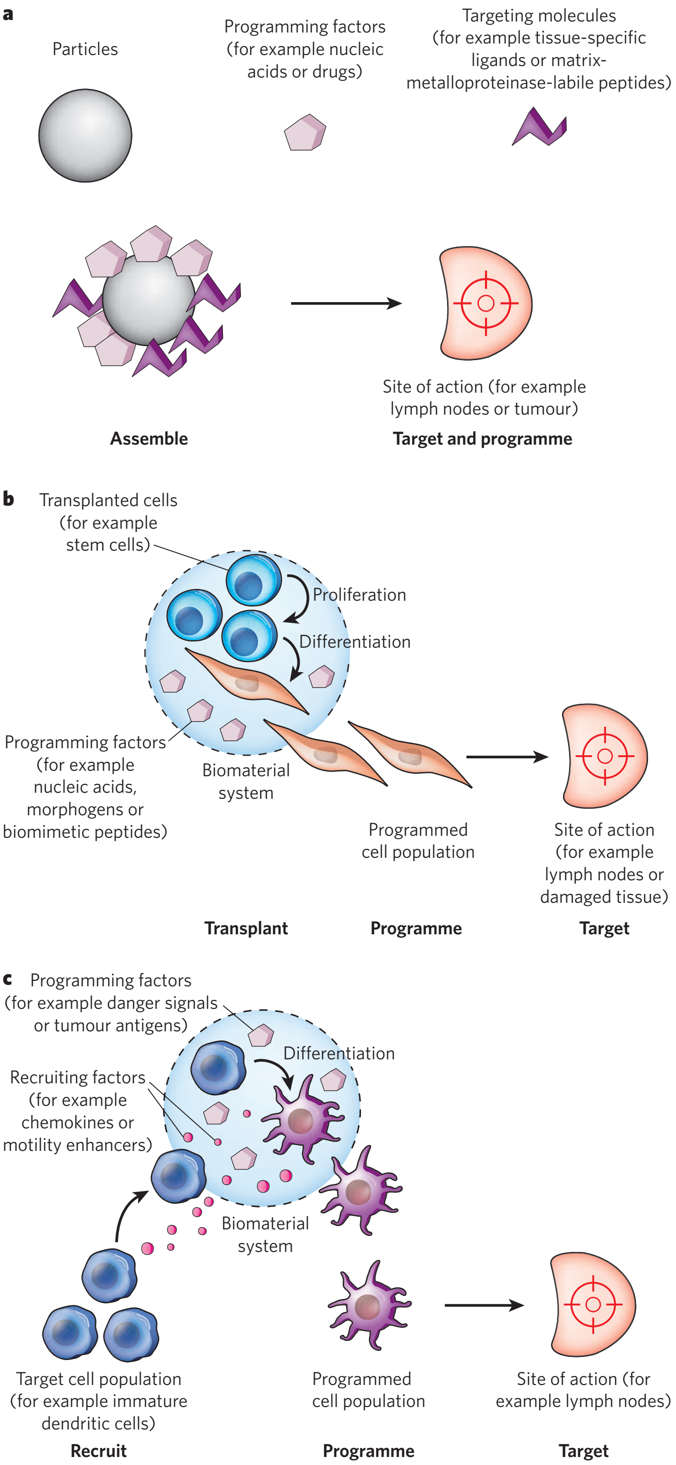

a
Particles
Programming factors (for example nucleic acids or drugs)
Targeting molecules (for example tissue-specific ligands or matrix-metalloproteinase-labile peptides)
Site of action (for example lymph nodes or tumour)
Assemble
Target and programme
b
Transplanted cells (for example stem cells)
Proliferation
Differentiation
Programming factors (for example nucleic acids, morphogens or biomimetic peptides)
Biomaterial system
Programmed cell population
Site of action (for example lymph nodes or damaged tissue)
Transplant
Programme
Target
c
Programming factors (for example danger signals or tumour antigens)
Differentiation
Recruiting factors (for example chemokines or motility enhancers)
Biomaterial system
Target cell population (for example immature dendritic cells)
Programmed cell population
Site of action (for example lymph nodes)
Recruit
Programme
Target

Figure 3 | Regulating biology at a distance: designing materials to target or mimic the niches of specific cell populations. a, Schematic of microparticles or nanoparticles (grey) whose assembly enables them to target specific anatomical or cellular regions of the body. This targeting occurs on the basis of the particles' size, shape and presentation of molecules (targeting molecules, purple) that are complementary to specific features of target cell populations. The particles subsequently manipulate cell fate locally through programming factors (pink). **b,** Schematic of an implantable biomaterial that mimics certain aspects of stem-cell niches in that it activates transplanted progenitor cells to proliferate and programs them to differentiate into cells that migrate into damaged tissues to participate in regeneration. **c,** Schematic of an implantable biomaterial system that mimics the microenvironment of an infection, allowing the recruitment, programming and subsequent targeting of activated antigen-presenting dendritic cells to the lymph nodes to participate in a potent antitumour response.

There has been great progress in elucidating functional domains within large ECM molecules and in using synthetic peptides to mimic key epitopes; perhaps the most common example is the grafting of integrin-binding peptides (for example RGD) onto hydrogel-forming polymers and other non-fouling substrates to facilitate cell adhesion[8]. Both the epitopes themselves and their spatial organization on micrometre and nanometre scales influence the fate of cells with which they interact[9]. Epitope spatial organization may be controlled on micrometre and nanometre scales through fabrication of self-assembling materials that present the epitope (Fig. 2b, c) or through direct patterning of the epitopes onto materials that otherwise present an inert background[10] (Fig. 2d). This type of patterning can also be performed on a larger length scale to mimic the ability of the natural ECM to create morphogen gradients[11] (Fig. 2e).

It is possible to use genetic engineering to increase the structural complexity of self-assembling materials and peptides further. Such an approach has been used to generate entirely new proteins that combine modules of different natural ECM molecules to obtain novel functionality, and even to incorporate non-natural amino acids that extend the range of chemical properties (for example to include the ability to undergo photocrosslinking) of cell-synthesized materials[12]. In some cases, synthetic biomaterials mimic nature not by influencing cells directly through receptor-binding epitopes but indirectly, by regulating the rate of matrix-metalloproteinase-mediated degradation and cellular invasion[13] or by initiating and regulating the formation of bioactive inorganic structures (for example mineralized bone or shell[14]). Both approaches have proved useful in augmenting the formation of mineralized tissues for dental and orthopaedic applications.

The potential utility of information-rich biomaterials that directly manipulate target biological systems is perhaps best exemplified by recent progress in the stem-cell field (see page 433), whereby key cues that regulate stem-cell biology are increasingly being incorporated into sophisticated biomaterials. Fundamental challenges in this field include the ability to expand stem cells *ex vivo* without using feeder layers, and enhancing the survival of transplanted stem cells and reproducibly regulating their fate in the body. Biomaterials are being used to define precisely the stem-cell microenvironment to meet these challenges[15], typically through the provision of high densities of cell-adhesion ligands[9], morphogens and other chemical cues[16], both to direct cell fate *in vitro* and to provide a template for the formation of new tissues by transplanted stem cells[17]. The precisely controlled spatiotemporal presentation of morphogens guiding development has inspired the design of biomaterials in which sequences of morphogens[18] and spatial gradients of morphogens[11,19] can be presented to guide these processes. Notably, soluble morphogens can exhibit enhanced biological activity when they are presented in insoluble form by tethering to biomaterials[20], providing another means of using ECM mimics to regulate cell fate.

Regulating biology at a distance

Although biomaterials are typically used to guide the behaviour of cells transplanted with the material or cells in the tissue into which the material is implanted, it has also become apparent that biomaterials can be designed to manipulate specific cell populations that reside in the host at a significant distance from the implant site. This can be done either by targeting the material to specific cells or anatomical locations or by controlling the trafficking of target cell populations (Fig. 3). Recent demonstrations of biomaterials as regulators of the immune system (see page 449) illustrate these two extremes well. Polymeric nanoparticles can be designed for non-invasive delivery into the body[21] and for trafficking through the lymphatic vessels to target T cells in the lymph nodes[22]. Similarly, nanoparticles are being designed to exploit the chemical and physical differences between normal and tumour-associated vasculature in order to concentrate the particles selectively within or near tumours, allowing subsequent drug-induced cell death[23] (Fig. 3a). Materials can also be designed to regulate outward migration of transplanted stem cells, or their differentiated progeny, to populate damaged tissues and promote regeneration efficiently[24] (Fig. 3b). Alternatively, biomaterials

© 2009 Macmillan Publishers Limited. All rights reserved

may program specific cell populations, without transplantation, by recruiting the population of interest (for example by releasing cytokines capable of recruiting immune cells) and subsequently activating these cells once resident in the material. This approach has been used to generate potent, antitumour responses by recruiting and programming immune cells *in situ*[25] (Fig. 3c). Because of their potential to target disease sites that are not yet clinically detectable, materials that regulate host biology at a distance show great promise for treating systemically disseminating diseases such as cancer.

Importance of physical variables in biomaterials design

The chemical composition of biomaterials has been the focus of their design for the past few decades, but there is growing appreciation of the importance of other properties, including topological, mechanical and electrical cues, in guiding a biological response. The features of particulate biomaterials on the length scale of individual molecules and cells (tens of nanometres to tens of micrometres) have significant effects on how cells perceive, interact and ingest the material, which affects the efficacy of materials used as drug carriers or vehicles targeting specific cells and tissues in the body[21]. Regardless of the chemical composition, the cellular response *in vitro* and *in vivo* can drastically alter depending on the mechanical properties of biomaterials[26]. Although the mechanisms responsible for these effects are only beginning to be understood, an underlying hypothesis in this area of research is that mechanosensing is an active cellular process involving dynamic interplay between the ECM and motor proteins coupled to the cytoskeleton. Biomaterials are being used both to study how cell phenotype is regulated by this crosstalk and as fundamental tools to characterize this dynamic interplay[27]. The ability of cells and natural biopolymers to sense, transmit and respond to mechanical signals is increasingly providing inspiration for new types of sensor, actuator and shape-control material (see page 442).

In addition to mechanical properties and size, external and environmental cues such as temperature and electromagnetic fields are increasingly being used to modulate the performance of biomaterials, often by dynamically altering their structure. Hydrogels, for example, can be designed to change their swelling behaviour and degree of nonspecific protein adsorption in response to temperature[28] or binding to specific ligands[29]. Despite intensive investigation into chemical structure–function relationships in hydrogels, the physics governing macromolecular transport within these materials, and their ability to resist protein adsorption, is still not completely understood and presents opportunities for future tuning of biomaterial performance. Studies have also demonstrated that drug delivery from biomaterials can be manipulated using remotely applied electromagnetic fields[30]; the same types of field can mediate the *in situ* assembly of scaffolds for tissue engineering[31]. Ion flows caused by electromechanical stimulation can probably modulate regeneration[32], suggesting that electrochemical signals could be used in the future to alter cell fate directly and, by manipulating biomaterial structure and presentation of chemical epitopes[33], indirectly. In the future, biomaterials may be engineered not only to respond to external fields and forces but also to generate these physical stimuli.

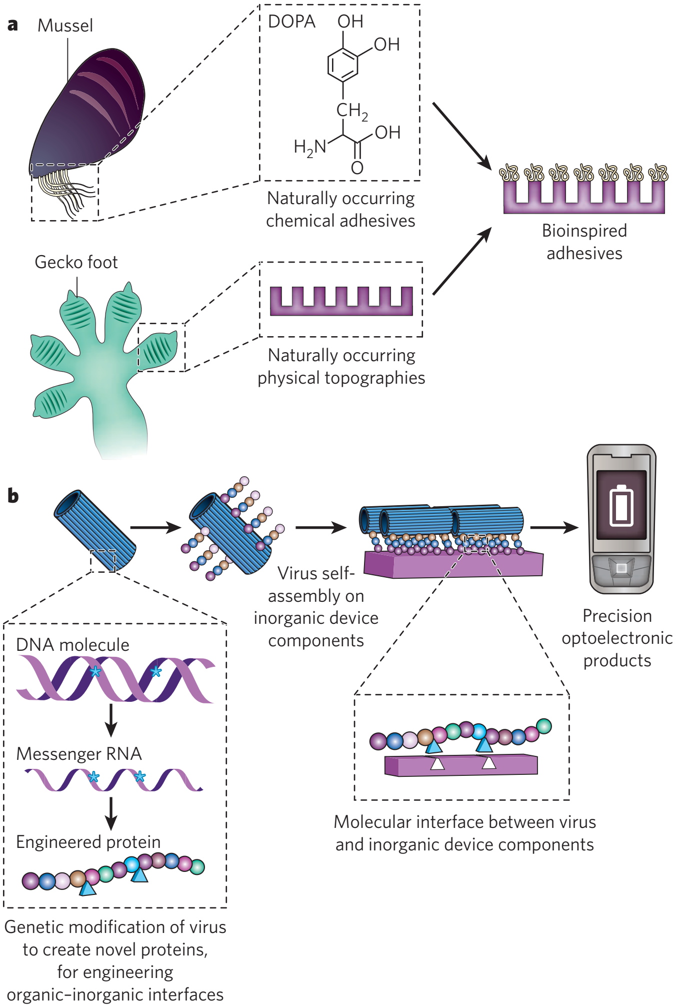

a
Mussel
DOPA
OH
OH
CH$_2$
OH
H$_2$N
O
Naturally occurring chemical adhesives
Bioinspired adhesives
Gecko foot
Naturally occurring physical topographies
b
DNA molecule
Messenger RNA
Engineered protein
Genetic modification of virus to create novel proteins, for engineering organic–inorganic interfaces
Virus self-assembly on inorganic device components
Precision optoelectronic products
Molecular interface between virus and inorganic device components

Figure 4 | Using components of biological organisms and materials in novel applications. a, The chemical and physical properties of materials used by organisms to facilitate surface adhesion can be mimicked, allowing the generation of synthetic coatings that modify surface chemistry or prevent biofouling. For example, 3,4(OH)$_2$-phenylalanine (DOPA), a naturally occurring chemical adhesive used to facilitate the adhesion of mussels to surfaces in wet environments, has been combined with the physically patterned nanopillar topography found in the toes of geckos, which facilitates strong adhesion in dry environments, to produce novel adhesives that work in both wet and dry environments. b, The molecular templating of whole viruses allows high-precision, multiscale patterning of electronic devices. Genetic modification of the organism (left) is used to engender bimolecular organic–inorganic interactions that lead to the coating of viruses with desired inorganic materials and their macromolecular assembly (centre). Low-cost, high-precision energy-storage systems (right) are one potential application of this concept[46].

Application of biomaterials beyond medical devices

Biomaterials have crucial roles in the fabrication of devices for biological screening, in basic science studies and in a variety of non-medical fields. Investigations into new diagnostic materials and devices are being driven by several factors: the ever-increasing recognition of the need for early diagnosis and intervention in human disease, particularly at low cost; the need for better *in vitro* screens for drug efficacy and toxicity; the potential dangers of food and water contamination; and the potential catastrophic results of biological warfare[34]. These biomaterials are designed on multiple length scales to present and organize arrays of molecules and cells for mechanistic studies and drug screening[35].

New approaches to biomaterials fabrication, often incorporating physical as well as chemical fabrication techniques, have paved the way for new approaches to diagnostics. As in the design of biomimetic medical devices, a crucial aspect of this work is the ability to make information-rich materials that assay multiple targets and allow multiple outputs (see page 461). A major feature of these approaches is the ability to capture rare cell populations[36]. Similarly, materials that change their optical or electrical properties in response to specific biological stimuli have been used to eliminate the need both for traditional probing tools (for example fluorescence) in diagnostics[37] and for basic investigations of cell–matrix interactions[38].

The increasing appreciation of the roles of insoluble signals from the ECM and physical forces in regulating cell fate has led to the use of biomaterials to construct physiologically relevant *in vitro* model systems. Perhaps the fastest-growing application of biomaterials for *in vitro* model systems is in the area of three-dimensional cell culture. Although matrix biologists have appreciated for some time that three-dimensional

© 2009 Macmillan Publishers Limited. All rights reserved

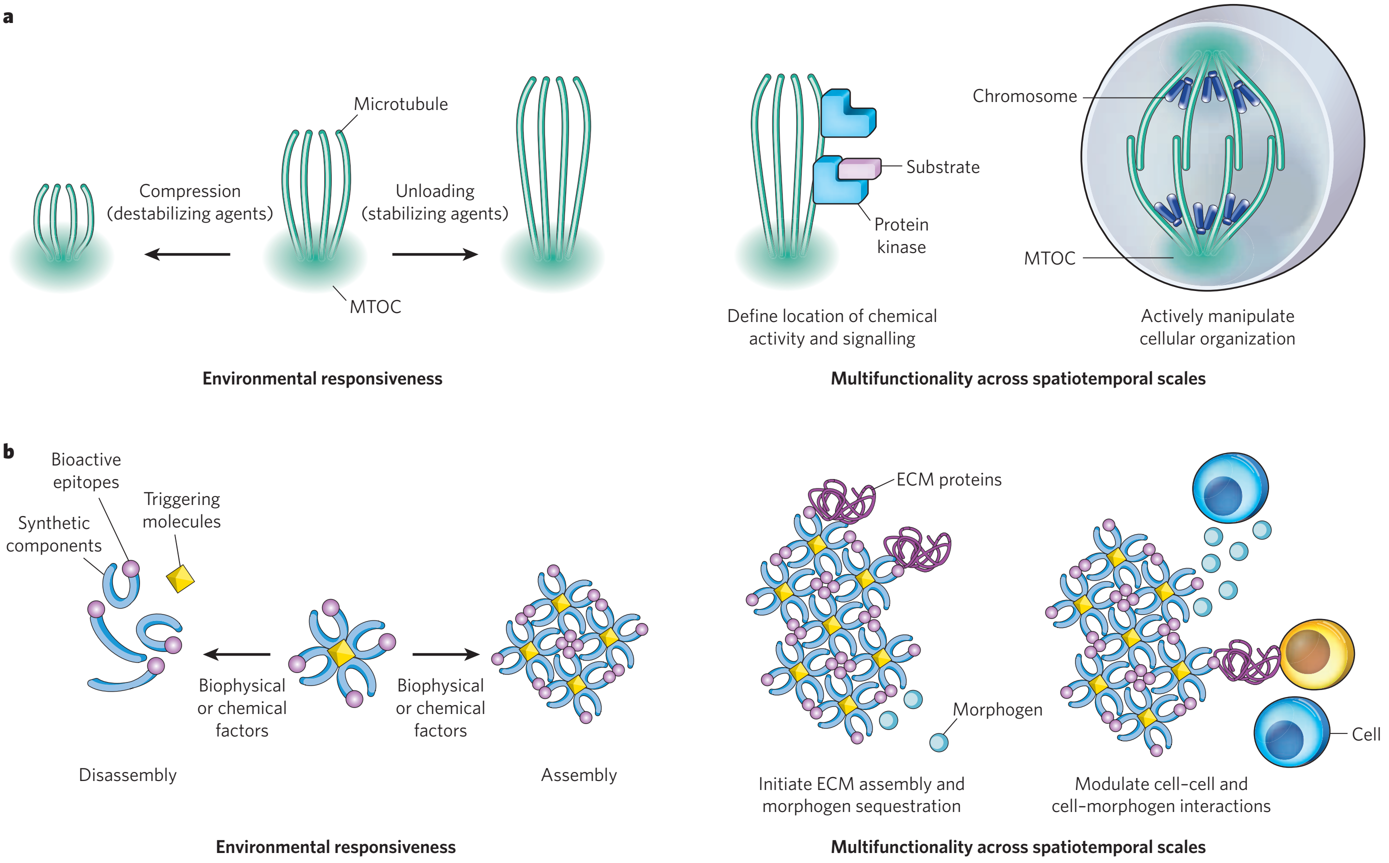


Figure 5 | The future: rethinking how inspiration is drawn from biology, and applying biological design principles to new areas. a, Microtubules are an example of a multifunctional natural material with a dynamic structure, responsiveness to multiple environmental cues (for example chemical stimuli and mechanical forces) and the ability to regulate a variety of events at the molecular and cellular levels. The dynamic nature of microtubules allows these polymers, which are assembled from relatively simple monomeric components (tubulins) at microtubule-organizing centres (MTOCs), to operate with a high degree of functional complexity in cells. Their dynamic assembly and disassembly in response to both chemical and mechanical stimuli from the environment is shown (left). Alterations in microtubule assembly can regulate a variety of cellular events (right), ranging from protein-mediated signalling (for example by regulating the localization of a protein kinase), which occurs on the nanometre scale, to mechanical control over cellular structure and organization during mitosis, which occurs on the micrometre scale. **b**, Potential biomedical devices inspired by microtubules. Structurally simple synthetic polymers can be designed to self-assemble in response to triggering molecules (for example ECM proteins or ions) and further assemble or disassemble in response to other environmental cues (for example pH, matrix metalloproteinases and physical forces) (left). Assembled polymers can subsequently regulate signalling between and within target cell populations to bring about biologically complex changes in their local environment. For example, they may bind to ECM proteins, initiating assembly or restructuring of the ECM and thereby manipulating cells locally (right). At the same time, they may sequester morphogens, generating gradients that alter cell behaviour over longer distances. As is the case for the natural materials that inspired their design, these synthetics undergo reorganization in response to changes in the local environment, subsequently altering the ways in which they interact with cells. This allows relatively simple materials to carry out the complex functions of both integrating multiple inputs from the environment and providing multiple outputs (cell-interactive stimuli) to regulate local biological events.

matrix culture provides more accurate *in vitro* models of *in vivo* phenomena (for example angiogenesis) than does two-dimensional cell culture[39], it has been difficult to distinguish effects that result from changes in the dimension of the microenvironment from effects that stem purely from the chemical changes required for three-dimensional culture. Biomaterials-based platforms that decouple dimensionality and physical properties from matrix chemistry, and engineering approaches to the analysis of natural ECM, are profoundly altering our understanding of a variety of biological processes, including tumour formation[40,41] and early development[42], and are also providing more physiologically relevant *in vitro* model systems for drug screening. In combination with the ability to scale down biological experiments greatly using array-type approaches[43], the ability of biomaterials to provide organized, physiologically relevant three-dimensional structures may fundamentally change how mechanistic questions in biology on cellular and tissue scales are approached, much in the same way that screening technologies such as gene arrays are affecting molecular biology. Such changes may be especially important in understanding how chemical inputs are systematically integrated, knowledge that would aid in efforts to develop network-type models of cell signalling for drug development.

Although bioinspired materials have had an increasing role in diagnostic devices and basic science, the fastest area of growth may be the application of such materials to fields outside medicine and biology. For example, just as fragments of the ECM facilitate mammalian cell adhesion to synthetic materials, functional chemical epitopes and physically patterned topographies used to facilitate surface adhesion in organisms such as mussels and geckos may improve the performance of synthetic adhesives in both medical and industrial applications[44] (Fig. 4a). Likewise, the materials used in nature for sensing are inspiring the development of biomimetic sensors; for example, salient features of the compound eyes of insects have been replicated in completely artificial materials designed to recapitulate the function of these natural sensors[45]. Although self-assembly is useful in constructing nanoscale devices, the assembly of synthetic polymers alone may be insufficient to provide the hierarchical organization such devices require. However, this organization can be accomplished by the molecular templating of whole microorganisms that have been genetically engineered to facilitate functionalities these

© 2009 Macmillan Publishers Limited. All rights reserved

organisms lack in nature (for example adhesion to metallic surfaces, for making devices such as batteries[46]) (Fig. 4b).

Biomaterials of the future

Advances in biomaterials will include the development of more functional medical materials and the expanded use of biomaterials into new fields of application. However, the future may also present an opportunity for practitioners in the field to rethink fundamentally the way in which inspiration is drawn from biology. Understanding the way in which complex dynamic behaviours are accomplished in nature may lead to the design of novel materials that mimic nature not through presenting active motifs replicated exactly from biological molecules but rather through reproducing the functional behaviour of these biological materials to obtain properties that are currently unavailable (Fig. 5). The application of the molecular templating of viruses to optoelectronic device fabrication is one early example of such an approach[46].

One focus of research on the new generation of bioinspired materials will probably be the development of 'smart', multifunctional nanoparticles or implants for use in our bodies. These complex materials would integrate multiple inputs from chemical and physical stimuli to determine their behaviour (Fig. 5b). Such materials could target desired anatomical regions, monitor health, and report on and actively intervene in biological crises. Biological systems have already inspired the development of cell-programming matrices based on our abstract understanding of dynamic biological processes such as infection, and these matrices accomplish their task with a small subset of key molecular stimuli[27]. New *ex vivo* biosensors capable of predicting disease are also likely to result from our understanding of living materials, as are new energy-storage devices, optical materials and other devices. Materials that selectively interact with specific cell populations, for use in diagnostic or therapeutic applications, may even be created by understanding and ultimately harnessing the dynamic cues provided by specific cell types (for example stem cells) to modify *in situ*, or assemble *in situ*, complex devices or materials from simple input templates.

A critical intellectual step in biomaterials design is the recognition both that biological polymers and organisms can be used as models of, or templates for, multifunctional, dynamic devices and that components of natural systems can be used for purposes other than that which they serve in nature[46]. This requires an abstract understanding of the biophysical properties imparted by certain molecular structures. This understanding is being applied in the context of self-assembling natural materials such as DNA, which originally was considered solely as an information-storage system but recently has inspired the development of new types of nanomaterial with precisely defined structures[6], as well as self-assembling synthetic polymers (inspired by the highly regulated base-pairing of DNA).

Some of the best-characterized self-assembling molecules in cell biology are the filamentous polymers that form the cytoskeleton. These information-rich polymers provide structural support that changes in response to environmental cues, and they also form a nexus for chemical signalling, by defining the location for synthetic activity, and regulate the movement of materials within the cell (Fig. 5a). What is perhaps most striking about these polymers from a materials standpoint, however, is that they can exhibit such a high degree of functional complexity while being relatively simple in composition. For example, microtubules consist of monomers that self-assemble in response to both chemical cues and mechanical loading[47] and rapidly disassemble and reassemble, providing distinct functions (Fig. 5a). Multifunctional synthetic materials with a subset of these functions are currently being developed. Nature is also inspiring micrometre-sized and nanometre-sized robots powered by stimulus-responsive soft actuators to augment bottom-up fabrication technologies[48]. Likewise, DNA may inspire the construction of actuators that mimic this biopolymer's dynamic assembly and responsiveness to environmental cues[49].

Beyond the devices and materials themselves, biological inspiration may revolutionize the methods used to produce and transform raw materials in the chemical and materials industries. For example, living plants can process, in huge quantities, a much greater variety of liquids and materials than are produced by humankind commercially, but they do so without the energy cost or waste streams typical of our chemical industry (Fig. 6). Applying lessons from nature may not only allow the synthesis of new chemicals but also significantly reduce the costs and environmental impacts associated with the manufacturing of current chemicals and drugs.

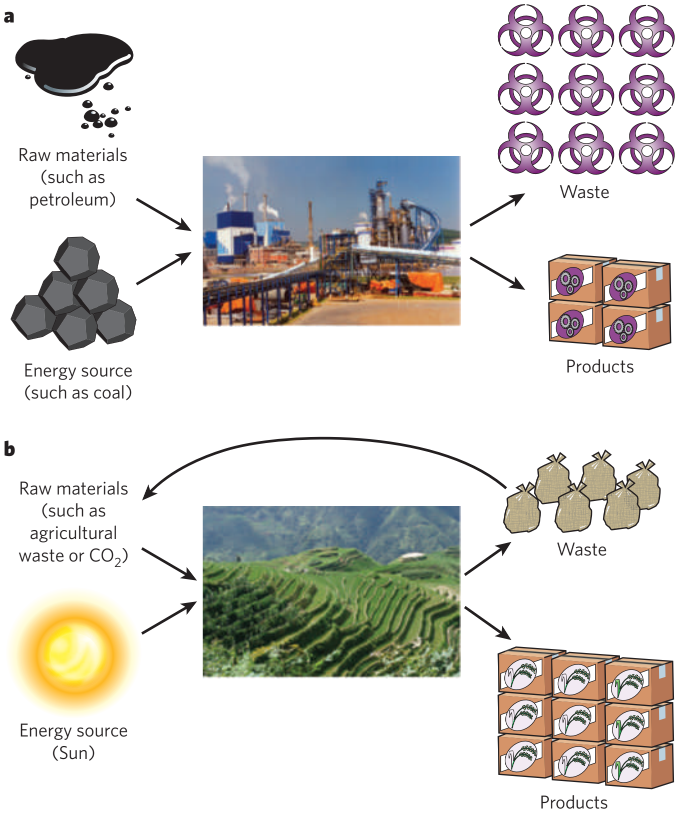

a
Raw materials (such as petroleum)
Energy source (such as coal)
Waste
Products
b
Raw materials (such as agricultural waste or CO$_2$)
Energy source (Sun)
Waste
Products

Figure 6 | The future: drawing inspiration from nature to rethink how materials and pharmaceuticals are manufactured. Schematic of a typical factory used for materials manufacturing, with associated inputs of raw materials and energy, and output waste streams (**a**). Schematic of a rice terrace, with its associated inputs and outputs (**b**). The relative sizes of manufactured products and the associated wastes shown represent the scale of waste streams and input materials in the respective schemes. The ability of natural systems to use renewable energy sources (for example solar energy) effectively and to recycle waste streams when generating products is inspiring novel approaches to manufacturing.

Accomplishing this transformation in the biomaterials field will require an improved understanding of how cells receive information from materials and how key signalling pathways process this information to dictate biological responses[50]; it will not suffice simply to make materials and empirically test for their effects on cell or host responses. In addition, gaining an abstract understanding of how the basic building blocks of biological systems are coordinated and integrated, in a manner analogous to the unit operations approach that revolutionized the chemical industry in the twentieth century, is likely to be an important step. This will require the development and application of new tools from biology, engineering and the physical sciences. Likewise, biophysical models of the materials themselves and their interaction with cells will also be necessary. The biomaterials field, both as an academic pursuit and as an industry, is quickly becoming unrecognizable in terms of its current definition. The field will need to be redefined to encompass materials that direct biology and those whose design and functions are inspired by natural materials; future generations of biomaterials are likely to be critical components in many facets of modern society. ■

© 2009 Macmillan Publishers Limited. All rights reserved

1. Ratner, B. D. & Bryant, S. J. Biomaterials: where we have been and where we are going. *Annu. Rev. Biomed. Eng.* **6,** 41–75 (2004).
   This is an excellent, comprehensive review of the history of the biomaterials field.
2. Anderson, J. M., Rodriguez, A. & Chang, D. T. Foreign body reaction to biomaterials. *Semin. Immunol.* **20,** 86–100 (2008).
3. Mroz, T., Yamashita, T. & Lieberman, I. The on- and off-label use of rhBMP-2 (INFUSE) in Medicare and non-Medicare patients. *Spine J.* **8,** 41S–42S (2008).
4. Shahani, S. *Advanced Drug Delivery Systems: New Developments, New Technologies.* Report No. PHM006F (Business Communications Company, 2006).
5. King, R. G. & Donohue, G. F. Estimates of medical device spending in the United States. *AMSA* <http://www.amsa.org/AMSA/libraries/committee_docs/king_paper_medical_device_spending.sflb.ashx> (2007).
6. Shih, W. M., Quispe, J. D. & Joyce, G. F. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature* **427,** 618–621 (2004).
7. Shin, H., Zygourakis, K., Farach-Carson, M. C., Yaszemski, M. J. & Mikos, A. G. Attachment, proliferation, and migration of marrow stromal osteoblasts cultured on biomimetic hydrogels modified with an osteopontin-derived peptide. *Biomaterials* **25,** 895–906 (2004).
8. Massia, S. P. & Hubbell, J. A. Covalently attached GRGD on polymer surfaces promotes biospecific adhesion of mammalian cells. *Ann. NY Acad. Sci.* **589,** 261–270 (1990).
9. Silva, G. A. *et al.* Selective differentiation of neural progenitor cells by high-epitope density nanofibers. *Science* **303,** 1352–1355 (2004).
10. Arnold, M. *et al.* Activation of integrin function by nanopatterned adhesive surfaces. *ChemPhysChem* **5,** 383–388 (2004).
11. Wosnick, J. H. & Shoichet, M. S. Three-dimensional chemical patterning of transparent hydrogels. *Chem. Mater.* **20,** 55–60 (2008).
12. Carrico, I. S. *et al.* Lithographic patterning of photoreactive cell-adhesive proteins. *J. Am. Chem. Soc.* **129,** 4874–4875 (2007).
13. Lutolf, M. P. *et al.* Repair of bone defects using synthetic mimetics of collagenous extracellular matrices. *Nature Biotechnol.* **21,** 513–518 (2003).
    This paper describes how materials can be designed to mimic key aspects of natural ECM (for example enzyme-mediated degradation) and function as templates for tissue regeneration.
14. Shin, K., Jayasuriya, A. C. & Kohn, D. H. Effect of ionic activity products on the structure and composition of mineral self assembled on three-dimensional poly(lactide-co-glycolide) scaffolds. *J. Biomed. Mater. Res. A* **83,** 1076–1086 (2007).
15. Li, Y. J., Chung, E. H., Rodriguez, R. T., Firpo, M. T. & Healy, K. E. Hydrogels as artificial matrices for human embryonic stem cell self-renewal. *J. Biomed. Mater. Res. A* **79,** 1–5 (2006).
16. Benoit, D. S., Schwartz, M. P., Durney, A. R. & Anseth, K. S. Small functional groups for controlled differentiation of hydrogel-encapsulated human mesenchymal stem cells. *Nature Mater.* **7,** 816–823 (2008).
17. Shin'oka, T. *et al.* Midterm clinical results of tissue-engineered vascular autografts seeded with autologous bone marrow cells. *J. Thorac. Cardiovasc. Surg.* **129,** 1330–1338 (2005).
18. Richardson, T. P., Peters, M. C., Ennett, A. B. & Mooney, D. J. Polymeric systems for dual growth factor delivery. *Nature Biotechnol.* **19,** 1029–1034 (2001).
19. Phillips, J. E., Burns, K. L., Le Doux, J. M., Guldberg, R. E. & Garcia, A. J. Engineering graded tissue interfaces. *Proc. Natl Acad. Sci. USA* **105,** 12170–12175 (2008).
20. Fan, V. H. *et al.* Tethered epidermal growth factor provides a survival advantage to mesenchymal stem cells. *Stem Cells* **25,** 1241–1251 (2007).
21. Tsapis, N., Bennett, D., Jackson, B., Weitz, D. A. & Edwards, D. A. Trojan particles: large porous carriers of nanoparticles for drug delivery. *Proc. Natl Acad. Sci. USA* **99,** 12001–12005 (2002).
22. Reddy, S. T. *et al.* Exploiting lymphatic transport and complement activation in nanoparticle vaccines. *Nature Biotechnol.* **25,** 1159–1164 (2007).
23. Park, J. H. *et al.* Systematic surface engineering of magnetic nanoworms for *in vivo* tumor targeting. *Small* **5,** 694–700 (2009).
24. Silva, E. A., Kim, E. S., Kong, H. J. & Mooney, D. J. Material-based deployment enhances the efficacy of endothelial progenitor cells. *Proc. Natl Acad. Sci. USA* **105,** 14347–14352 (2008).
25. Ali, O. A., Huebsch, N., Cao, L., Dranoff, G. & Mooney, D. J. Infection-mimicking materials to program dendritic cells *in situ*. *Nature Mater.* **8,** 151–158 (2009).
    This paper describes how biomaterials can be designed to regulate host biology at a distance by recruiting, locally programming and subsequently dispersing target cell populations to produce potent biological responses.
26. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126,** 677–689 (2006).
    This paper demonstrates the importance of physical properties of biomaterials in controlling cellular response.
27. Tan, J. L. *et al.* Cells lying on a bed of microneedles: an approach to isolate mechanical force. *Proc. Natl Acad. Sci. USA* **100,** 1484–1489 (2003).
28. Park, T. G. & Hoffman, A. S. Synthesis and characterization of pH- and or temperature-sensitive hydrogels. *J. Appl. Polym. Sci.* **46,** 659–671 (1992).
29. Podual, K., Doyle, F. J. & Peppas, N. A. Glucose-sensitivity of glucose oxidase-containing cationic copolymer hydrogels having poly(ethylene glycol) grafts. *J. Control. Release* **67,** 9–17 (2000).
30. Edelman, E. R., Brown, L., Taylor, J. & Langer, R. *In vitro* and *in vivo* kinetics of regulated drug release from polymer matrices by oscillating magnetic fields. *J. Biomed. Mater. Res.* **21,** 339–353 (1987).
31. Alsberg, E., Feinstein, E., Joy, M. P., Prentiss, M. & Ingber, D. E. Magnetically-guided self-assembly of fibrin matrices with ordered nano-scale structure for tissue engineering. *Tissue Eng.* **12,** 3247–3256 (2006).
32. Adams, D. S., Masi, A. & Levin, M. H^+ pump-dependent changes in membrane voltage are an early mechanism necessary and sufficient to induce tail regeneration. *Development* **134,** 1323–1335 (2007).
33. Lahann, J. *et al.* A reversibly switching surface. *Science* **299,** 371–374 (2003).
34. Martinez, A. W., Phillips, S. T. & Whitesides, G. M. Three-dimensional microfluidic devices fabricated in layered paper and tape. *Proc. Natl Acad. Sci. USA* **105,** 19606–19611 (2008).
35. Khetani, S. R. & Bhatia, S. N. Microscale culture of human liver cells for drug development. *Nature Biotechnol.* **26,** 120–126 (2008).
36. Nagrath, S. *et al.* Isolation of rare circulating tumor cells in cancer patients by microchip technology. *Nature* **450,** 1235–1239 (2007).
37. Stern, E. *et al.* Label-free immunodetection with CMOS-compatible semiconducting nanowires. *Nature* **445,** 519–522 (2007).
38. Gupta, V. K., Dubrovsky, T. B. & Abbott, N. L. Optical amplification of ligand–receptor binding using liquid crystals. *Science* **279,** 2077–2080 (1998).
39. Madri, J. A., Pratt, B. M. & Tucker, A. M. Phenotypic modulation of endothelial cells by transforming growth factor-β depends upon the composition and organization of the extracellular matrix. *J. Cell Biol.* **106,** 1375–1384 (1988).
40. Fischbach, C. *et al.* Cancer cell angiogenic capability is regulated by 3D culture and integrin engagement. *Proc. Natl Acad. Sci. USA* **106,** 399–404 (2009).
41. Ghajar, C. M. *et al.* The effect of matrix density on the regulation of 3-D capillary morphogenesis. *Biophys. J.* **94,** 1930–1941 (2008).
42. Xu, M. *et al.* Encapsulated three-dimensional culture supports the development of nonhuman primate secondary follicles. *Biol. Reprod.* **81,** 587–593 (2009).
43. Khademhosseini, A., Langer, R., Borenstein, J. & Vacanti, J. P. Microscale technologies for tissue engineering and biology. *Proc. Natl Acad. Sci. USA* **103,** 2480–2487 (2006).
44. Lee, H., Scherer, N. F. & Messersmith, P. B. A reversible wet/dry adhesive inspired by mussels and geckos. *Nature* **448,** 338–341 (2007).
45. Jeong, K. H., Kim, J. & Lee, L. P. Biologically inspired artificial compound eyes. *Science* **312,** 557–561 (2006).
46. Nam, K. T. *et al.* Virus-enabled synthesis and assembly of nanowires for lithium ion battery electrodes. *Science* **312,** 885–888 (2006).
    This paper discusses the engineering of non-medical materials through the templating of viruses. The precisely tuned patterns of spatial features of the natural organism promise distinct performance advantages.
47. Needleman, D. J. *et al.* Higher-order assembly of microtubules by counterions: from hexagonal bundles to living necklaces. *Proc. Natl Acad. Sci. USA* **101,** 16099–16103 (2004).
48. Sidorenko, A., Krupenkin, T., Taylor, A., Fratzl, P. & Aizenberg, J. Reversible switching of hydrogel-actuated nanostructures into complex micropatterns. *Science* **315,** 487–490 (2007).
49. Omabegho, T., Sha, R. & Seeman, N. C. A bipedal DNA Brownian motor with coordinated legs. *Science* **324,** 67–71 (2009).
50. Kyriakides, T. R. *et al.* The CC chemokine ligand, CCL2/MCP1, participates in macrophage fusion and foreign body giant cell formation. *Am. J. Pathol.* **165,** 2157–2166 (2004).

Acknowledgements We thank D. Ingber for discussions. We acknowledge funding from the US National Institutes of Health (National Institute of Dental and Craniofacial Research). N.H. is supported by a National Science Foundation Graduate Research Fellowship.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to D.J.M. (mooneyd@seas.harvard.edu).

© 2009 Macmillan Publishers Limited. All rights reserved

Designing materials to direct stem-cell fate

Matthias P. Lutolf[1], Penney M. Gilbert[2] & Helen M. Blau[2]

Proper tissue function and regeneration rely on robust spatial and temporal control of biophysical and biochemical microenvironmental cues through mechanisms that remain poorly understood. Biomaterials are rapidly being developed to display and deliver stem-cell-regulatory signals in a precise and near-physiological fashion, and serve as powerful artificial microenvironments in which to study and instruct stem-cell fate both in culture and *in vivo*. Further synergism of cell biological and biomaterials technologies promises to have a profound impact on stem-cell biology and provide insights that will advance stem-cell-based clinical approaches to tissue regeneration.

Stem cells are defined by their ability to self-renew and produce specialized progeny[1,2]. Consequently, they are the most versatile and promising cell source for the regeneration of aged, injured and diseased tissues. Embryonic stem cells, induced pluripotent stem cells and adult stem cells are obtained from three different sources and have different advantages (Fig. 1). However, despite the remarkable potential clinical applications of each of these stem-cell populations, their use is currently hindered by hurdles that must be cleared[3] (Table 1). These obstacles may appear daunting, but nature has strategies to surmount them *in vivo*. Thus, a major goal is to develop new culture-based approaches, using advanced biomaterials, that more closely mimic what the body already does so well and promote differentiation of pluripotent cells or propagation of specialized adult stem cells without loss of 'stemness'.

An increasing emphasis on design principles drawn from basic mechanisms of cell–matrix interactions and cell signalling has now set the stage for the successful application of biomaterials to stem-cell biology. This application has the potential to revolutionize our understanding of extrinsic regulators of cell fate, as matrices can be made using technologies that are sufficiently versatile to allow recapitulation of features of stem-cell microenvironments, or niches, down almost to the molecular detail[4].

In the body, adult stem cells reside within instructive, tissue-specific niches that physically localize them and maintain their stem-cell fate[5,6] (Fig. 2). The key function of stem-cell niches is to maintain a constant number of slowly dividing stem cells during homeostasis by balancing the proportions of quiescent and activated cells. On insult (that is, injury, disease or damage), stem cells exit the niche and then proliferate extensively, self-renew and differentiate to regenerate the tissue. Within the niche, stem cells are thought to be exposed to complex, spatially and temporally controlled biochemical mixtures of soluble chemokines, cytokines and growth factors, as well as insoluble transmembrane receptor ligands and extracellular matrix (ECM) molecules (Fig. 2). In addition to understanding this biochemical signalling regulatory network, it is key to appreciate the biophysical properties of the niche, including matrix mechanical properties and architecture (topographical cues), to elucidate the role of niche elements[7,8].

To shed light on the mechanisms that regulate stem cells, approaches that allow the study of stem-cell function in response to isolated components of a complex system — that is, models that simplify it — are crucial. Biomaterials approaches, in combination with other technologies such as microfabrication and microfluidics, are well suited to assist studies of stem-cell biology through the creation of evolving systems that allow key variables to be systematically altered and their influence on stem-cell fate analysed. Thus, biomaterials technologies provide the exciting possibility of deconstructing and then reconstructing niches, allowing quantitative analysis of stem-cell behaviour in a manner not previously possible.

In this Review, we use specific examples to outline the various means by which biomaterials technologies have been, and could be, used to construct versatile model systems for stem-cell biology, as well as to develop carriers for stem cells and biomolecules, facilitating the *in vivo* use of stem cells in tissue engineering. We focus on hydrogels as one emerging and physiologically relevant class of biomaterial, although we acknowledge that other biomaterials have been, and will be, used in these applications. For a more comprehensive understanding of the diverse types and applications of biomaterials in stem-cell biology and bioengineering, we refer readers to several recent reviews[9–15]. We anticipate that insight will be gained from studies using biomaterials that allow the enhanced differentiation of embryonic stem cells and induced pluripotent stem cells into tissue-specific differentiated states and the propagation of adult stem cells without losing their stem-cell properties.

Designing 2D materials to control stem-cell fate *in vitro*

In vitro stem-cell research is carried out with cells cultured on flat substrates coated, for example, with collagen or laminin, on feeder-cell layers (that is, in co-culture experiments) or on or within hydrogels made from naturally derived ECM components (for example collagen or Matrigel). By far the most frequently used material for the culture of stem cells is rigid polystyrene tissue-culture plastic. Cells in plastic dishes are typically exposed to soluble factors in liquid media. These culture conditions are very different from the conditions experienced by cells in the body, where they are associated with anchored molecules presented in close proximity to surrounding cell surfaces and contained within an ECM that creates a relatively soft microenvironment. The constraints imposed on stem cells within the three-dimensional (3D) niche have effects that are still being explored and should not be ignored. With this goal in mind, two-dimensional (2D) biomaterial culture systems are highly advantageous as a simplified approach to deconstructing the niche and identifying and assessing the effects of individual niche components on stem-cell fate (Fig. 3).

[1]Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. [2]Baxter Laboratory in Stem Cell Biology, Department of Microbiology and Immunology, Institute for Stem Cells and Regenerative Medicine, Stanford University School of Medicine, Stanford, California 94305, USA.

© 2009 Macmillan Publishers Limited. All rights reserved

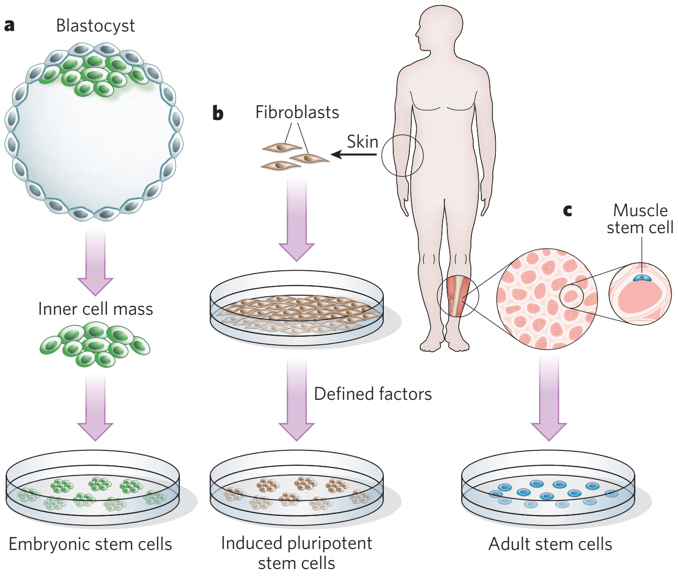


Figure 1 | Origins, promises and challenges of stem cells. a, Embryonic stem cells, which are derived from blastocysts (formed at an early stage of embryogenesis), provided the first human source of pluripotent cells that could be differentiated to generate any cell type. **b,** Induced pluripotent stem cells, which have all of the properties of embryonic stem cells, were first generated by introducing genes encoding four proteins into somatic cells, such as skin fibroblasts[90]. Embryonic stem cells and induced pluripotent stem cells have a seemingly unlimited self-renewal potential in culture, but the absence of methods to direct these cells into a single tissue-specific lineage robustly and reproducibly and to avoid the risk of tumour formation reliably have restricted their use in humans. Induced pluripotent stem cells overcome the problem of immune tolerance and the ethical issues faced by the use of embryonic stem cells and adult stem cells in patients, but current methods to reprogram somatic cells and to generate induced pluripotent stem cells are extremely slow and inefficient. **c,** Resident tissue-specific adult stem cells (for example muscle stem cells) lack the plasticity of embryonic stem cells and induced pluripotent stem cells but are not tumorigenic. They are primed for, and extremely efficient at, generating progeny that differentiate into specialized cell types. It is difficult to induce the self-renewal of adult stem cells in culture and to propagate the cells to yield clinically useful numbers *in vitro*, underscoring the importance of elucidating the role of the endogenous microenvironment in the regulation of stem-cell fate. A cross-sectional view of muscle fibres (red) surrounded by basement membrane (white) is shown, together with a muscle stem cell (blue); these stem cells reside on top of muscle fibres, beneath the basement membrane.

Probing biochemical stem-cell–ECM interactions in two dimensions

The identification of ECM molecules with biological relevance to stem-cell regulation is a critical step towards defining the regulatory influences of the stem-cell niche. Biomaterials approaches have been explored to define the identity, concentration and patterns of soluble or tethered ECM molecules singly (Fig. 3a) and in combination (Fig. 3b). Several groups have made ECM arrays by means of robotic spotting to screen for a molecule or combinations of molecules that induce fate changes[16–19]. For example, arrays consisting of 192 unique combinations of ECM and signalling molecules have been printed onto slides containing a thin coating of polydimethylsiloxane; and placental cadherin, epithelial cadherin, laminin and JAG1 (a ligand for the receptor NOTCH1) were each found to promote the conversion of mammary progenitor cells to myoepithelial or luminal epithelial fates[19].

Notably, not only is the rigidity of the tissue determined by the structure of the ECM (whether loosely or densely packed), but differences in density also result in different local concentrations of exposed ECM components, which in turn lead to differences in cell signalling and adhesion. In addition, the architecture of the ECM provides geometric cues to cells in the form of fibre diameter, length and crosslinking patterns, as well as surface irregularities ('nanotopography'). Two-dimensional approaches should greatly improve our understanding of the relevance of these key ECM properties to stem-cell biology[8].

Probing cell–cell interactions in two dimensions

The effects of cell–cell interactions are usually studied by culturing two cell types together; however, using such co-culture strategies makes it difficult to discern the role of particular molecules, be they soluble or tethered. In tissues, secreted growth factors and cytokines are mostly tethered to ECM components such as proteoglycans, whereas receptor ligands are presented to stem cells at the surface of nearby support cells. In both cases, molecule immobilization is proposed to have the critical role of increasing protein stability, promoting persistent signalling and inducing receptor clustering[20]. For example, covalent attachment of fibroblast growth factor 2 (FGF2) to a synthetic polymer stabilized the growth factor and increased its potency 100-fold relative to FGF2 in solution. In response to the tethered FGF2, embryonic stem cells exhibited increased proliferation and activation of ERK1 (also known as MAPK3), ERK2 (MAPK1), JNK (MAPK8) and c-Fos signalling[21]. Similarly, when epidermal growth factor (EGF) was covalently tethered to a biomaterial scaffold, it was more effective than its soluble counterpart in inducing the spread of mesenchymal stem cells and preventing Fas-ligand-induced death[22]. Finally, immobilized leukaemia inhibitory factor (LIF), but not soluble LIF, led to prolonged activation of LIF targets and maintenance of embryonic stem cells in an undifferentiated state with the capacity to generate chimaeric mice even after culture for more than 2 weeks[23].

The function of receptor ligands associated with cell membranes also is contingent on the mode of presentation. When tethered, DLL1 (a ligand for the receptor NOTCH1) resulted in an increase, relative to soluble DLL1, in the number of human cord-blood CD133$^+$ (PROM1$^+$) cells capable of reconstituting the circulation in irradiated mice[24]. Similarly, when tethered, rat JAG1 enhanced NOTCH1 signalling and increased the differentiation of rat oesophageal stem cells[25].

These examples demonstrate the importance of ligand presentation in stem-cell fate and function. Testing single candidate molecules is instructive, but to discover novel ligands and cytokines that have effects on stem cells, an unbiased high-throughput biomaterials approach is desirable. Such an approach, using spots of 576 different combinations of 25 acrylate-based polymers in arrays on the nanolitre-scale, found combinations that influenced embryonic stem-cell attachment, proliferation and differentiation[26]. Ideally, high-throughput approaches could be devised to incorporate the many other biophysical and biochemical parameters described above.

Probing biophysical stem-cell–matrix interactions in two dimensions

Ageing, injury and disease are often associated with increased deposition and altered organization of ECM components such as collagen, resulting in significant changes to the stiffness of the matrix, which most likely potentiate pathogenesis, for example in the case of Duchenne muscular dystrophy[27–29]. Natural and synthetic matrices can be produced to create cell-culture substrates with known elastic modulus (or stiffness) (Fig. 3c) and, unlike plastic substrates, they also provide diffusion of soluble molecules to the basal surface, as well as the apical surface, and can be used to test the relevance of homeostatic and disease-related matrix stiffness to stem-cell behaviour. Notably, soluble factors in culture media always act in conjunction with the tissue-culture matrix, and together they affect cell fate.

Table 1 | Current promises and limitations of stem-cell populations

Feature	Embryonic stem cells	Adult stem cells	Induced pluripotent stem cells
Artificial system	Yes	No	Yes
Pluripotent	Yes	No	Yes
Efficient differentiation	No	Yes	No
Expansion in culture	Yes	No	Yes
Rare cell type	No	Yes	Yes
Immune compatible	No	No	Yes
Teratoma risk	Yes	No	Yes

© 2009 Macmillan Publishers Limited. All rights reserved

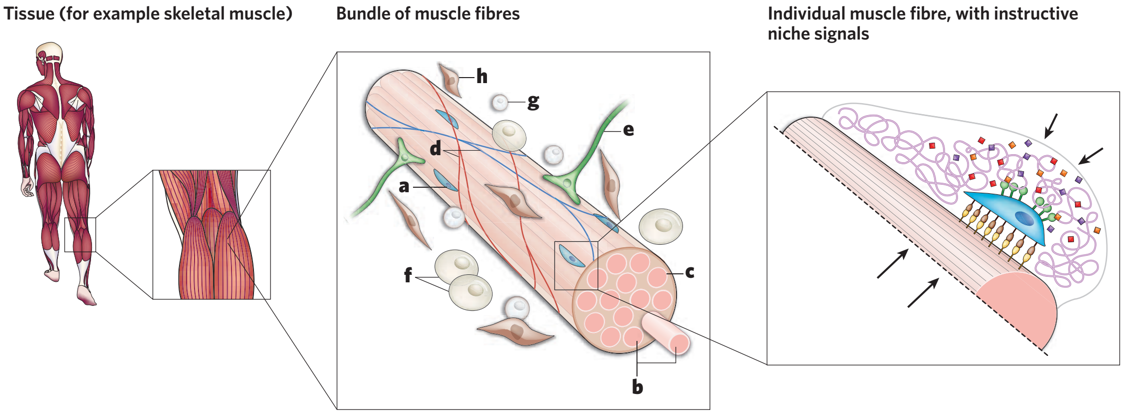

Tissue (for example skeletal muscle)
Bundle of muscle fibres
Individual muscle fibre, with instructive niche signals
a
b
c
d
e
f
g
h

Figure 2 | Biochemical and biophysical properties of stem-cell niches. Adult stem cells reside in tissue-specific microenvironments, called niches. Niches protect stem cells and regulate their functions. First described in *Drosophila melanogaster* and *Caenorhabditis elegans* ovary and testis, niches have now been characterized in many other tissues, including skeletal muscle (left panel). Muscle stem cells (**a**) reside on post-mitotic, multinucleated muscle fibres (**b**) and are ensheathed by a basement membrane (**c**) (central panel). The complexity of this stem-cell niche is increased by the presence of many other, non-muscle, cell types, including endothelial and blood cells in the vasculature (**d**), motor neurons (**e**), adipocytes (**f**), and circulating immune cells (**g**) and fibroblasts (**h**). Within the niche (right panel), spatially and temporally controlled biochemical mixtures of soluble and tethered chemokines, cytokines and growth factors (diamonds), as well as ECM molecules (purple) and ligands presented by muscle fibres (yellow), interact with transmembrane receptors displayed by muscle stem cells (brown and green) to regulate stem-cell fate. It is also becoming clear that the biophysical properties of the stem-cell microenvironment are crucial components of the niche; arrows indicate forces imposed on stem cells by the resistance of the ECM and surrounding tissue.

In a landmark study, human mesenchymal stem cells assumed morphological patterns and gene expression patterns consistent with differentiation into distinct tissue-specific cell types when exposed to polyacrylamide gels with a range of stiffnesses typical of brain, muscle and bone[30]. This study highlighted the potent influence of matrix mechanical properties on stem-cell fate and led to the exploration of further links between stem-cell behaviour and matrix elasticity. Since then, substrate stiffness has been shown to modulate the proliferation and differentiation of embryonic stem cells and certain types of adult stem cell. Specifically, adult neural stem cells cultured on a relatively stiff synthetic matrix gave rise primarily to glial cells, whereas on a softer matrix that more closely resembled the compliancy of *in vivo* brain tissue, neurons were the predominant cell type[31]. Furthermore, the rate of adult skeletal-muscle stem-cell proliferation increased as substrate stiffness increased[32].

A major challenge in studies of this type is separating the effects of matrix stiffness from those of ligand density. To eliminate this variable, 'tunable' gel systems in which matrix stiffness and ligand density can be independently controlled are especially advantageous. Using one such tunable, synthetic cell-culture system, human embryonic stem cells have been propagated and maintained in an undifferentiated state in the absence of a feeder layer[33]. We predict that once the profound effects of the physical properties of culture substrate on stem-cell fate are fully appreciated, culture platforms based on soft biomaterials are likely to largely replace those made of the standard, rigid, tissue-culture plastic.

Within the niche, cell shape is defined, in part, by the constraints imposed by the surrounding ECM on cells during development and in adulthood[34,35]. Although some of these effects are probably due to alterations in the adhesive interactions and crosstalk between the ECM and the cell as they work to define each other, there is ample evidence suggesting that physical control of cell shape alone can act as a potent regulator of cell signalling and fate determination[36] (Fig. 3d). One remarkable demonstration of the influence of cell shape on cell function used micropatterned ECM islands allowing precise and reproducible control of the size of the cell attachment area[37]. Single mesenchymal stem cells cultured on small islands adhered poorly, had a rounded morphology and acquired an adipogenic fate, whereas on larger islands they were adherent, spread out, exhibited increased focal adhesions and cytoskeletal reorganization, and acquired an osteogenic fate[38]. Furthermore, human embryonic stem cells cultured on spatially restricted islands yielded dense OCT4$^+$ (POU5F1$^+$) pluripotent colonies, whereas on large islands embryonic stem cells differentiated[39]. Such studies are just beginning to shed light on the profound impact that matrix architecture, during development and pathogenesis, has on cell-shape-induced changes to cytoskeletal organization and signalling, and subsequent stem-cell specification and function.

High-throughput single-cell analyses in 2D microenvironments

Traditional *in vitro* experiments are conducted on cell ensembles. In these studies, measurements entail averaging responses across an entire population. Consequently, behaviours such as apoptosis, changes in cell-cycle kinetics, changes in self-renewal, and differentiation may be missed. For stem-cell analyses, this poses a significant problem, as many stem-cell populations are heterogeneous. As a result, rare stem cells in a heterogeneous mixture may be missed, or analyses may be skewed by the behaviour of rapidly growing progenitor cells, because in many cases stem cells are non-dividing or grow significantly slower than do progenitors. Conventional cell-culture platforms are not readily applicable to the investigation of stem cells at the single-cell level. For example, standard multiwell plates such as 96-well plates would require large amounts of expensive culture-media components and do not offer sufficient throughput. This problem has been solved by the introduction of microwell array cultures for cell biology (Fig. 4). These modular platforms permit the analysis of a large number of single, spatially confined cells. They have recently been successfully applied in stem-cell biology, using both embryonic stem cells and adult stem cells (see, for example, refs 40–48).

Polymer-hydrogel networks such as those formed from polyethylene glycol (PEG) are useful in the production of microwell substrates, as they allow simultaneous and independent assessment of the effects of biophysical and biochemical properties on stem-cell fate at the clonal level. Currently available hydrogel-crosslinking chemistries and macromolecule architectures can generate a wide range of hydrogels with distinct and reproducible mechanical properties[49]. PEG is almost inert to protein adsorption, and proteins can be tethered to hydrogels by attaching a chemical moiety to proteins of interest and subsequently crosslinking it into the hydrogel network in a technique called microcontact printing[50] (Fig. 4a). By using a standard microfabrication technique with polydimethylsiloxane as a replica, it is possible to structure hydrogel arrays topographically to contain thousands of spatially segregated micropatterns, for example round microwells with proteins printed specifically at the bottom of each well[48].

Using a hydrogel-culture approach in conjunction with time-lapse microscopy, the behaviour of haematopoietic stem cells in response to a panel of soluble and tethered molecules was assessed. Division patterns

© 2009 Macmillan Publishers Limited. All rights reserved

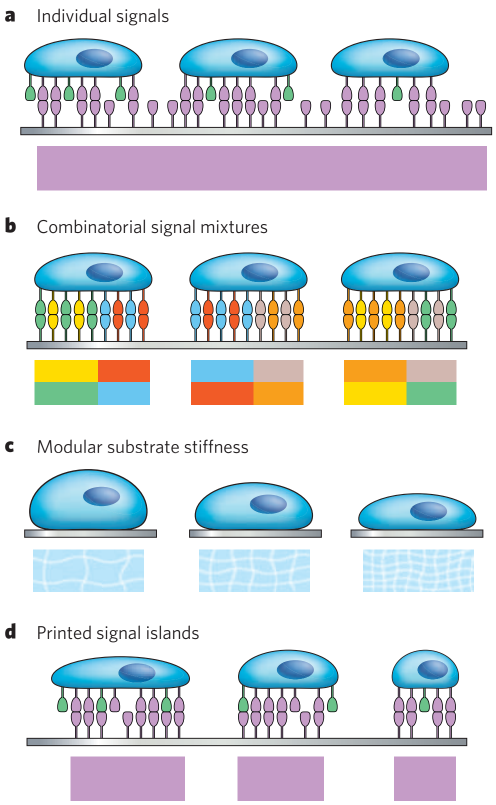

a Individual signals
b Combinatorial signal mixtures
c Modular substrate stiffness
d Printed signal islands

Figure 3 | Engineering 2D artificial stem-cell niches. The top part of each panel shows stems cells exposed to a specific, engineered 2D microenvironment (viewed from the side), and the bottom part shows a schematic of the microenvironmental features (viewed from above), represented as blocks of colour matching the signals that are present. The substrates (grey) encompass various materials, such as plastics, glass or hydrogels, except for in panel **c** (in which soft materials such as hydrogels are depicted). **a**, Individual signal molecules are displayed on the substrate. **b**, Combinatorial mixtures of signals that are generated, for example, by robotic protein spotting can be presented to stem cells. **c**, The desired substrate stiffness can be controlled by, for example, differential crosslinking of hydrogel networks. **d**, Microcontact printing of cell-adhesion or cell-regulatory proteins on inert surfaces allows control of protein spot size and, therefore, cell shape.

consistent with depletion (fast symmetrical division), asymmetrical self-renewal (asymmetrical cell division) and symmetrical self-renewal (symmetrical division) were observed and subsequently assayed *in vivo* for their ability to reconstitute the blood over a longer timescale. Remarkably, this study showed that exposure to single factors, such as WNT3A and neural cadherin, could induce self-renewal of haematopoietic stem cells *in vitro*[48]. Additionally, it provided strong support for the idea that *in vitro* stem-cell behaviour can be highly predictive of *in vivo* potential[46]. A similar approach can now be applied to any number of stem-cell types to identify previously unknown physical and chemical regulators and the relevant presentation of those molecules to elicit effects on stem-cell self-renewal and differentiation. The production of novel microwell arrays in which substrate stiffness, protein doses (such as in gradients) (Fig. 4b), and protein combinations (Fig. 4c) and their spatial arrangement (Fig. 4d) can all be controlled will be essential for the success of these studies.

In conclusion, deconstructing a complex 3D niche into 2D biomaterial model systems is a powerful and promising strategy for discovering new regulatory mechanisms governing stem-cell biology. The structural, biophysical and biochemical parameters of these model systems can be varied in myriad ways to identify and elucidate the effects of the components of putative stem-cell niches on stem-cell function. Given the precise control of nanometre-scale chemical and topographical features, as well as the possibility of computationally predicting fluid dynamics and transport conditions during cell culture, and the simplicity of collecting cells after culture, 2D platforms are poised to generate fresh insight into the biochemical and biophysical regulation of stem cells[51,52].

Designing 3D materials to control stem-cell fate *in vitro*

Whereas 2D approaches allow well-controlled analysis of the impact on stem cells of individual components of the niche, 3D approaches should allow reconstruction, and realization of the complexity, of the natural tissue (Fig. 5). In epithelial tissues (for example skin and gut), stem cells adhere to 2D, sheet-like basement membranes, but most stem-cell niches (for example those in bone marrow, brain and muscle) are 3D microenvironments composed of hydrated, crosslinked networks of ECM proteins and sugars. In three dimensions, stem cells can be exposed to a solid microenvironment that fully ensheathes them (Fig. 5a), in contrast to 2D platforms, in which cells are typically exposed to a solid, flat surface on the basal side and to liquid at the apical surface. However, although conceptually appealing, the construction of 3D artificial microenvironments is not simple[53]. There are chemical challenges in the production process, considerations of appropriate elasticity, and the need to overcome the physical constraints that impede migration or morphogenesis. First and foremost, in most cases, cell viability remains a problem; second, understanding the read-outs from such complex multicomponent systems is not straightforward. As a result, high-throughput analyses are currently not possible, and few of the many possible variables can be systematically explored. Nonetheless, progress is being made.

Several impediments to 3D culture must be overcome. First, to expose stem cells to an accurate 3D artificial environment, chemical approaches that allow the embedding of stem cells must be used. This is ideally performed *in situ* (that is, during the formation of the 3D material), which requires a mild and highly specific crosslinking chemistry so as not to compromise cell viability as a result of adverse side reactions. Several methods of forming synthetic or semi-synthetic hydrogel matrices under physiological conditions have been developed for this purpose and are reviewed in, for example, refs 54 and 55. Some of these approaches explore not only highly specific chemical or enzymatic reactions but also physical mechanisms of crosslinking, such as the molecular self-assembly of small-molecule building blocks (including peptides, peptide amphiphiles and oligonucleotides). Each of these approaches has been demonstrated to yield viable encapsulated cells after crosslinking; the strategies differ primarily in the hydrogel-network structures that are produced and in how cells respond to these different network structures (of which some are porous and others are dense meshworks).

Second, the biophysical characteristics of the 3D environment are important. Cells embedded in a 3D environment can suffer from a lack of gases and nutrients. This problem is overcome by using scaffolds made of solids such as polymers with interconnected porosity and by using hydrogel networks with microscopic meshes, as such structures readily allow the diffusion of macromolecules. Third, substrate elasticity and materials with mechanical properties closely approximating those of natural stem-cell niches are desirable[28], as described above. Last, physical constraints that impede cell proliferation, migration and morphogenesis should be avoided. To avoid the potential problems of having physical barriers in three dimensions, materials that have matrix porosity on the scale of cellular processes can be designed. For example, nanofibrillar hydrogels that contain microscopic pores large enough to facilitate cell growth have been developed[56]. An attractive alternative approach uses polymer gels that can be synthesized to contain chemically crosslinked substrates for proteases naturally secreted by cells, for example during cell invasion. This feature allows a dynamic interplay between the cells and their microenvironment such that the cells locally degrade and then 'remodel' the matrix. For example, PEG-based hydrogels have been rendered chemically degradable through hydrolytic breakdown of ester bonds[57] and have been developed with cleavage sites for cell-secreted matrix metalloproteinases or plasmin[4]. This cell-regulatable breakdown of the matrix allows cell migration and proliferation in a manner determined by the cells.

© 2009 Macmillan Publishers Limited. All rights reserved

Probing stem-cell–matrix interactions in three dimensions

A long-standing question in stem-cell biology and tissue engineering is that of how the numerous components of the stem-cell niche govern stem-cell fate in three dimensions. This question is difficult to address *in vivo* or using any existing 2D *in vitro* approaches. A 3D stem-cell niche is extremely complex (Fig. 2), and the number of its physical, chemical and mechanical effectors is too great to define in practice. Even if the specific nature of its components were known, testing them systematically would not be possible. Thus, developing new approaches aimed at high-throughput screening of combinations of 3D microenvironmental variables, in a manner analogous to 2D ECM protein microarrays or other cellular arrays described above, is a major goal[58–60].

The production of high-throughput microarrays of 3D matrices could be possible using robotic liquid-dispensing and printing approaches in combination with biomaterial-crosslinking chemistries. Combinatorial mixtures of liquid precursors of hydrogel networks can be deposited in minute volumes and at high density onto solid substrates[61]. In one of the first examples of this emerging strategy[62], 3D PEG-hydrogel arrays were produced to screen for the individual and combinatorial effects of gel degradability, cell-adhesion-ligand type and cell-adhesion-ligand density on the viability of human mesenchymal stem cells. Increased PEG-network degradability and greater cell-adhesion-ligand density were both found to increase the viability of the stem cells in a dose-dependent manner.

Measures of cell viability constitute a minimal first step. It is necessary to design more-sophisticated ways of measuring stem-cell proliferation, asymmetrical and symmetrical division, self-renewal and differentiation into selected lineages that can be assessed in three dimensions. One challenge in this endeavour will be to analyse cellular responses in three dimensions, for which one focal plane for microscopic read-out is not sufficient. Ultimately, it would be desirable to investigate the role of the 3D microenvironment in controlling stem-cell fate on a more comprehensive ('systems') level, integrating the complete set of relevant variables. Importantly, when promising candidate microenvironments are identified through such studies, selected materials need to be further evaluated using *in vivo* approaches, for example by transplantation of cell–matrix constructs into mice.

Probing cell–cell interactions in three dimensions

Important components of stem-cell niches are the cells that abut stem cells, which are sometimes referred to as support cells or niche cells. These can include vascular cells, neural cells, and stromal cells such as fibroblasts. They not only provide instructive secreted signalling cues but also send signals through transmembrane proteins or bound matrix proteins. Although this type of cellular crosstalk is conceptually appreciated as being highly significant to stem-cell behaviour (to quiescence, activation and proliferation), the study in three dimensions of which factors have a critical role and how they act together is a nascent field.

Nonetheless, progress is being made in technologies that would allow the investigation of such cell–cell signalling interactions in near-physiological 3D microenvironments (Fig. 5b). One approach is based on the electropatterning of mammalian cells within hydrogels[50]. Electropatterning localizes live cells (possibly of any type) within hydrogels, such as photopolymerized PEG gels, by using dielectrophoretic forces. Large numbers of multicellular clusters of precise size and shape have been formed in three dimensions on one focal plane. By modulating cell–cell interactions in 3D clusters of various sizes, this microscale tissue organization was, for example, shown to influence the biosynthesis of bovine articular chondrocytes, with larger clusters producing smaller amounts of sulphated glycosaminoglycan per cell.

Other work has combined gel patterning with microfluidic technology to analyse angiogenesis in 3D co-cultures[63]. Primary liver and vascular endothelial cells were cultured on each side wall of a collagen gel between two microfluidic channels. Morphogenesis of 3D hepatic cultures was found to depend on fluid flow across the nascent tissues. Vascular cells formed 3D capillary-like structures that extended across an intervening gel to the hepatocytes' tissue-like structures. This is a remarkable advance, as microvascular networks are considered to be important components of several stem-cell niches[6]. Thus, these approaches could prove useful in addressing fundamental questions in stem-cell biology.

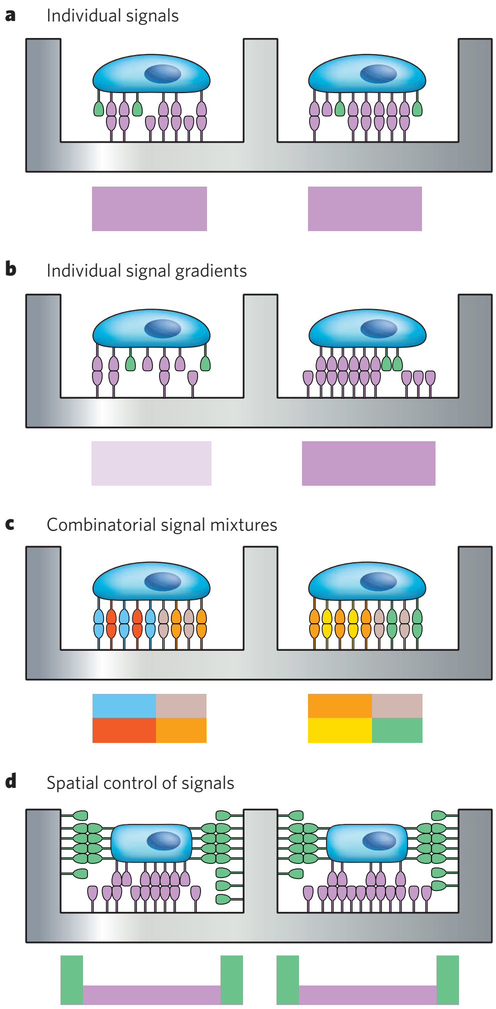

a Individual signals
b Individual signal gradients
c Combinatorial signal mixtures
d Spatial control of signals

Figure 4 | Engineering 'pseudo-3D' models of stem-cell niches. Microwell arrays allow the confinement of single stem cells and analysis of entire stem-cell populations at the individual cell level, overcoming the problem of heterogeneity of stem-cell populations. **a**, Microwell arrays can be readily engineered so that individual niche signals are presented at a certain concentration on the bottom of the well, by using manual microcontact printing. **b**, **c**, Robotic protein spotting on the microwell bottom should allow control of protein doses in each microwell, including the generation of protein gradients (**b**) or the production of combinatorial protein mixtures (**c**). **d**, Patterning approaches can be designed to allow the spatial arrangement of niche cues at the level of an individual, encapsulated stem cell. The top part of each panel shows stem cells exposed to a specific, engineered pseudo-3D microenvironment (viewed from the side), and the bottom part shows a schematic of the particular microenvironmental features (viewed from above (**a**–**c**) or from the side (**d**)).

3D biomolecule gradients in stem-cell biology

Morphogen gradients have long been known to regulate cell fate and tissue or organ development[64]. Biomolecule gradients are crucial regulatory components of dynamic tissue processes, not only during development but also during homeostasis and regeneration. Therefore, the creation of biomolecule gradients in 3D biomaterials systems has received increasing attention in stem-cell bioengineering (Fig. 5c). Such gradients could be shallow, such that a given cell experiences one concentration along its whole length, or steep, such that the cell experiences a different concentration at each end. Cells may migrate away from or

© 2009 Macmillan Publishers Limited. All rights reserved

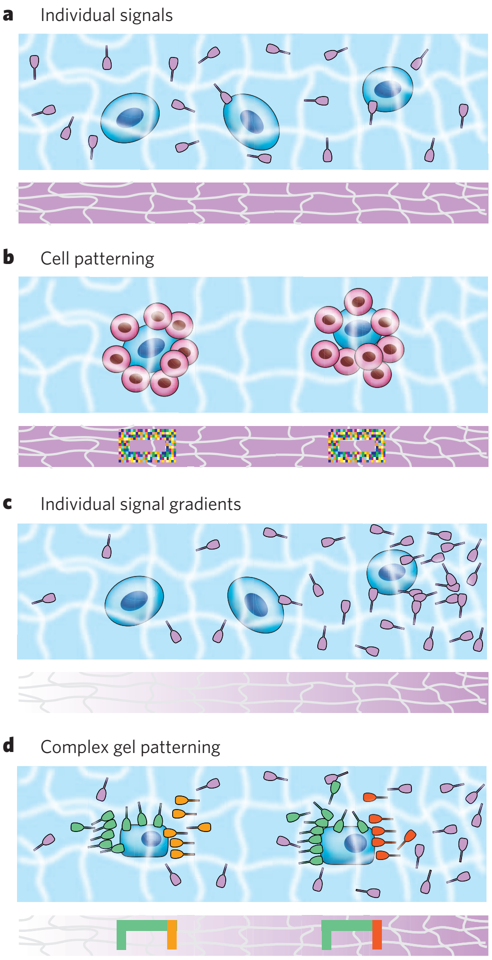


Figure 5 | Engineering 3D *in vitro* models of stem-cell niches. Mild and selective hydrogel-crosslinking chemistries are necessary for a true 3D embedding of stem cells in an artificial microenvironment that more closely mimics natural stem-cell niches. Polymer-hydrogel networks can be engineered with tailor-made biochemical and biophysical characteristics. **a**, Individual niche signals can be tethered to gel networks to probe their function in stem-cell behaviour. **b**, Three-dimensional micropatterning technologies such as electropatterning allow the arrangement of cells in 3D hydrogels in a spatially well-controlled manner. Using this technique, single stem cells could be patterned in three dimensions in contact with support cells (pink) that provide many regulatory niche cues. **c**, Niche cues could be displayed as large-scale gradients (which is currently only possible with non-tethered signals). **d**, Hydrogel networks can now be precisely micropatterned in three dimensions; for example, by light-controlled modification of biochemical gel characteristics (such as niche-signal availability) or biophysical gel characteristics (such as gel-crosslink density). The laser from a confocal microscope allows high spatial resolution, as well as dynamic control of 3D gel patterning. The top part of each panel shows cells exposed to a specific, engineered 3D microenvironment (viewed from the side), and the bottom part shows a schematic of the particular microenvironmental features (viewed from above (**a**–**c**) or from the side (**d**)).

towards a particular biomolecule concentration. Alternatively, when gradients are steep, cell polarity and asymmetry may be induced, just as in a stem-cell niche.

Arguably the most precise and robust way of generating a biomolecule gradient is through microfluidic technology[65], because microfluidics allows the well-controlled manipulation of very small amounts of fluid. Microfluidic gradient platforms have already been applied to stem-cell biology, albeit in two dimensions (see, for example, ref. 66). However, 3D gradient systems are rapidly being developed[67,68]. One example is a microfluidics-based approach whereby cells within alginate gels could be exposed to desirable soluble gradients in 3D microenvironments[67]. Applied to adult stem-cell culture, such intricate control over the biochemical microenvironment in three dimensions is an important step towards the *in vitro* recapitulation of stem-cell microenvironments that are more complex. The advantages of combining biomaterials engineering with microfluidics for stem-cell applications are clear[69]: this combination offers the potential for arrays of individually addressable cell-culture chambers[70,71] in which artificial microenvironments are exposed to spatially and temporally controlled biomolecule gradients (temporal control allowing delivery at any time during an experiment). Because proteins can be tethered to gel networks, it should be possible to combine tethering and soluble gradients of protein morphogens to mimic the exposure of cells to both ECM-bound protein gradients and soluble gradients, to recreate a stem-cell niche in three dimensions more accurately.

Mimicking the spatial 3D niche heterogeneity

Stem cells sense and respond to the spatial heterogeneity of 3D microenvironments. Many *in vivo* stem-cell microenvironments are 'polarized' structures, in that they expose individual stem cells to differential niche components. An example is the niche of the satellite cell (the canonical muscle stem cell), which is located between the muscle-fibre membrane and the surrounding basement membrane (Fig. 2). An ideal 3D *in vitro* model of a stem-cell niche would allow recapitulation of this type of complex architecture and manipulation at a desired time during an experiment, for instance to address the question of whether microenvironmental polarity dictates when a cell is quiescent and when it is activated.

Application of hydrogel engineering using photochemistry suggests that the construction of such complex microenvironments in three dimensions will be possible and will allow impressive precision and control over the dynamics[72–74] (Fig. 5d). For example, in photopolymerized PEG hydrogels, photolabile building blocks have been synthesized[74]: these can be cleaved by a controlled light beam to modulate biophysical and biochemical gel properties locally at a given time. Mesenchymal stem cells were shown to respond to locally induced network changes in stiffness and cell-adhesion properties; in a densely crosslinked gel, the decrease in crosslinking density obtained through cleavage of the backbone of the photolabile chain induced a significant morphological change in the encapsulated stem cells (initially round in shape, they became more spread out). Moreover, the controlled manipulation of the concentration of cell-adhesive peptide ligands in the PEG gel led to inducible changes in chondrocyte differentiation. Differentiation into chondrocytes increased when an RGD peptide, which binds to integrins, was removed using light at a later time during 3D cell culture.

Microfluidic technology could also be used to mimic to some extent the spatial heterogeneity of stem-cell microenvironments[75]. Several 3D matrices (such as type I collagen, Matrigel or fibrin) containing cells were micropatterned within a single microfluidic channel, stably interfacing each other. Cell culture was performed over several weeks and led to spatially restricted development of multicellular structures within designed patterns. These new methods will be of use in studying a great number of questions in stem-cell biology.

From artificial niches to 3D *in vitro* 'tissues'

The 3D approaches discussed above serve as powerful model systems to elucidate extrinsic stem-cell regulation, but they would not form an appropriate basis on which to reconstruct large-scale tissue models[76] using stem cells and biomaterials as building blocks, because they do not facilitate the modular and spatially well-controlled combination and positioning of these building blocks and they do not extend to scales of millimetres to centimetres. A technology known as bioprinting may be the method of choice in this endeavour, because theoretically it is feasible to combine layers of ECM and cell mixtures in modules of varying composition on a micrometre scale and in three dimensions. In bioprinting, custom-designed inkjet printers deposit, in a controlled layer-by-layer fashion, cells and biomaterials in almost picolitre-sized droplets at a rate of tens of thousands per second (see, for example,

© 2009 Macmillan Publishers Limited. All rights reserved

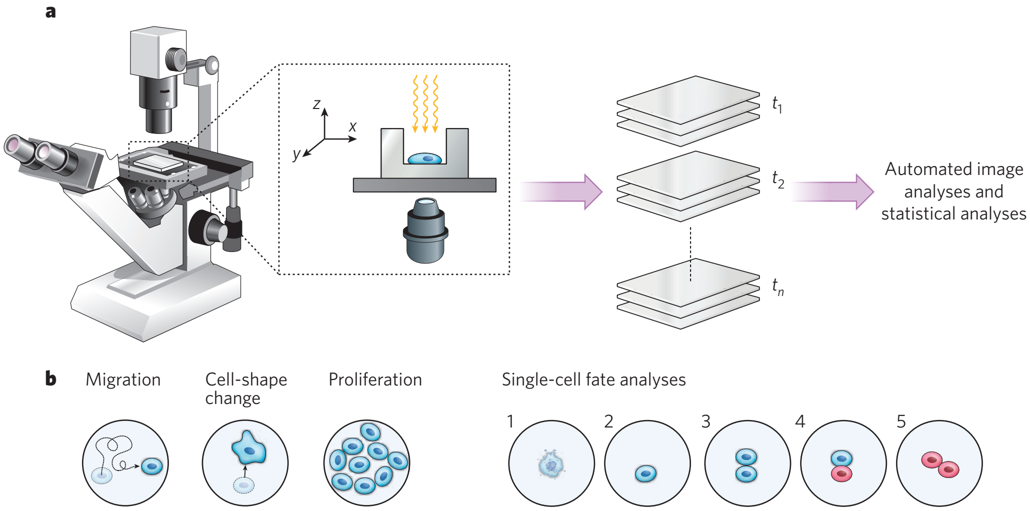

a
z
x
y
t1
t2
tn
Automated image analyses and statistical analyses
b
Migration
Cell-shape change
Proliferation
Single-cell fate analyses
1
2
3
4
5

Figure 6 | Quantitative investigations of *in vitro* stem-cell fates using live-cell microscopy. a, Time-lapse microscopy is a powerful way of probing the behaviour of live stem cells in artificial niches. Stem cells are imaged at various time points (t_1 to t_n) and locations to generate time-lapse movies, and automated image analysis and statistical analyses are used to quantify the dynamic cells' behaviour. **b,** A number of different read-outs, corresponding to different stem-cell functions, are available. Together with cell migration, changes in cell shape and changes in proliferation kinetics, the recording and automated analyses of changes in the fate of individual stem cells are crucial. Illustrated are cell death (1); quiescence (that is, non-cycling; 2); symmetrical self-renewal divisions (proliferation behaviour imposed in response to stress or trauma; 3); asymmetrical self-renewal divisions generating one daughter cell that retains stem-cell identity and one already partly differentiated (a behaviour thought to be dominant during homeostatic conditions; 4); and symmetrical depletion divisions, in which both daughter cells lose stem-cell function (the default behaviour of adult stem cells grown *in vitro*; 5).

ref. 77). On deposition on a substrate, these droplets can be polymerized to form a solid gel that could encapsulate stem cells or contain biomolecules with locally modular composition. Although the bioprinting field has arguably had little impact on stem-cell biology as yet, the results obtained so far with other cell types look promising. For example, viable 3D composites of embryonic neurons and astrocytes have been patterned in multilayered collagen[78]. Currently, bioprinting is cumbersome, mainly because a suitable 'bio-ink' (that is, a hydrogel system that can be rapidly crosslinked, with high spatial precision, and is simultaneously highly biologically active and permissive) is lacking. However, if this obstacle could be overcome, bioprinting could be a significant step towards achieving the long-standing goal of tissue engineers, namely the formation of functional tissues outside the human body.

Designing materials systems to control stem-cell fate *in vivo*

Biomaterials technologies also offer exciting opportunities to control the fate of stem cells *in vivo*, that is, at a site of tissue damage. Two main modes of application have been proposed: one in which biomaterials are used as carriers for introducing stem cells into damaged, diseased or aged tissue, and one in which biomaterials are used to augment endogenous stem-cell function. Here we briefly discuss these two approaches, the challenges they entail, and the promise they hold for future applications. For a more comprehensive review of such strategies, we refer readers to recent reviews[7,79].

Biomaterials-mediated *in vivo* delivery of stem cells and support cells

The transplantation of stem cells, or possibly any type of cell, for applications in regenerative medicine has serious limitations. First, survival and engraftment of transplanted stem cells is extremely poor (typically only a few per cent of all cells engraft); this is the main obstacle to the clinical translation of stem-cell biology. Second, in the absence of instructive cues in a disrupted biological environment characterized by abundant cell and tissue necrosis, such as in regenerating tissue, the fate of the engrafted cells may be poorly controlled. Biomaterials can be designed to act as carriers for the local delivery of stem cells, support cells or molecular niche cues. The biomaterials may markedly improve the impact of transplanted stem-cell populations. Many of the concepts described for *in vitro* use above could find useful application *in vivo*. For example, materials could be designed as multifunctional stem-cell microenvironments that affect tissue regeneration on multiple levels, including the following: delivering stem cells in a protective gel and enhancing viability; delivering support cells to increase the numbers and stimulate the function of endogenous stem cells; delivering diffusible cytokines to promote the mobilization of endogenous cells involved in repair, such as those that form blood vessels; displaying regulatory proteins to enhance survival and to stimulate self-renewal and expansion of the transplanted cells; and displaying regulatory proteins to stimulate tissue-specific differentiation for the purpose of large-scale tissue regeneration. We think that the spatial and temporal control of these features would enhance their utility in tissue regeneration, improving tissue function and overcoming the adverse effects of disease or ageing[80,81].

Biomaterials-controlled *in vivo* delivery of niche signals

Biomaterials concepts are also beneficial for the local and specific delivery of bioactive niche components. These components may be inhibitory or stimulatory molecules or drugs that might increase stem-cell numbers or function when delivered to the niche. This could be achieved by forming a scaffold that leads to timed drug (small chemical) or biomolecule delivery near a stem-cell niche or by targeted delivery of soluble microparticles or nanoparticles as carriers of such bioactive niche components[82]. Biofunctional polymer particles can now be engineered to be efficient in such applications. Specifically, they can be functionalized so that they bind to specific molecules on cells, are responsive to environmental signals such as proteases secreted by cells, or are delivered encapsulated in a manner that leads to temporally controlled release or cellular uptake[83,84].

The most challenging, but perhaps the ultimate, biomaterials goal is to create multicomponent, injectable materials designed to act as *de novo* niches *in vivo*. Heavily damaged, necrotic tissue may have lost microenvironments suitable for stem-cell occupancy, as is the case in aged or dystrophic muscles[80]. Artificial niches would need to incorporate appropriate 'homing' signals that could attract endogenous stem cells and localize them by means of known cell–cell or cell–matrix adhesive interactions. Then, once localized to these artificial niches, the cells would need to be exposed to tethered signals that control stem-cell function, in particular expansion by self-renewal division. Neighbouring vascular cells and neural cells would need access. Upon injury, the upregulation and release of proteases would enable the newly formed

© 2009 Macmillan Publishers Limited. All rights reserved

stem cells to escape the niche and contribute to differentiation and tissue regeneration. Cell transplantation has recently been used to show that the formation of a heterotopic haematopoietic microenvironment is possible[85]. Upon transplantation, MCAM (melanoma cell-adhesion molecule)-expressing subendothelial cells present in human bone-marrow stroma were shown to be capable of forming a miniature bone organ. In another example, macroporous polyester scaffolds pre-seeded with rat osteogenic cells were implanted into nude mice (which lack a thymus and therefore cannot mount an immune response to reject foreign, transplanted materials)[86]. This scaffold design led to the formation of an active haematopoietic marrow with stromal and haematopoietic compartments, of which the stromal compartment seemed to have attracted and retained endogenous haematopoietic precursor cells, thus acting as a functional artificial niche.

Future challenges

Both 2D and 3D biomaterials-based culture platforms have the potential to help researchers to identify novel biochemical and biophysical regulators of stem-cell fates (such as survival, quiescence, self-renewal and differentiation). Ultimately, these findings will translate into new biomolecule-based therapies to induce resident stem-cell function and promote the regeneration of aged, injured or diseased tissues *in vivo*[1].

A major hurdle for the advancement of most, if not all, of the described strategies lies not in the biomaterials field but rather in stem-cell biology. The identification of markers that specifically and robustly distinguish stem cells from their differentiated progeny (for example OCT4) has proved successful with embryonic stem cells but is particularly cumbersome with many adult stem-cell types such as haematopoietic stem cells, which currently require multiple positive and negative selective markers for robust identification. In addition to retrospective analyses by immunohistochemistry, prospective analyses would be a great advance here. In particular, there is a paucity of dynamic live-cell markers (for example stage-specific promoters driving the expression of fluorescent reporter genes with appropriate half-lives) that would allow gene expression changes to be monitored in real time in conjunction with morphological assessment.

Another current problem is the bottleneck in the analysis of the large data sets accumulated by exploring some of the biomaterials platforms described here. Although groups have presented computer-based algorithms to assay cell morphology and genealogical histories acquired by time-lapse microscopy[46,87,88], for the most part a large amount of manual correction is still required[48,89]. High-fidelity, fully automated analyses of cell behaviours (Fig. 6) (such as proliferation rate and division history, to generate genealogical histories; directed migration and velocity; and cell shape and size) could exponentially accelerate our understanding of stem-cell biology. However, although cells may express given markers and may have distinct proliferation behaviours, the only true test of *in vitro* data on stem-cell function is validation with an *in vivo* assay.

The rate at which biomaterials approaches are being applied to address questions in stem-cell biology ensures that new insight will be gained into the mechanistic regulation of stem-cell fate. However, although there is now a plethora of ingenious biomaterial platforms with which to analyse the influence of the biophysical and biochemical properties of stem-cell niches, these platforms have only just begun to be applied to directing stem-cell fate. Collaborative efforts between cell biologists and materials scientists are critical to answering the key biological questions and fostering interdisciplinary stem-cell research in directions of clinical relevance. ■

1. Blau, H. M., Sacco, A. & Gilbert, P. M. in *Essentials of Stem Cell Biology* 2nd edn (eds Lanza, R. *et al.*) 249–257 (Academic, in the press).
2. Blau, H. M., Sacco, A. & Gilbert, P. M. in *Encyclopedia of Stem Cell Research* (eds Svendsen, C. & Ebert, A.) (SAGE, in the press).
3. Daley, G. Q. & Scadden, D. T. Prospects for stem cell-based therapy. *Cell* **132,** 544–548 (2008).
4. Lutolf, M. P. & Hubbell, J. A. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. *Nature Biotechnol.* **23,** 47–55 (2005).
5. Scadden, D. T. The stem-cell niche as an entity of action. *Nature* **441,** 1075–1079 (2006).
6. Morrison, S. J. & Spradling, A. C. Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. *Cell* **132,** 598–611 (2008).
7. Discher, D. E., Mooney, D. J. & Zandstra, P. W. Growth factors, matrices, and forces combine and control stem cells. *Science* **324,** 1673–1677 (2009).
8. Guilak, F. *et al.* Control of stem cell fate by physical interactions with the extracellular matrix. *Cell Stem Cell* **5,** 17–26 (2009).
9. Chai, C. & Leong, K. W. Biomaterials approach to expand and direct differentiation of stem cells. *Mol. Ther.* **15,** 467–480 (2007).
10. Saha, K., Pollock, J. F., Schaffer, D. V. & Healy, K. E. Designing synthetic materials to control stem cell phenotype. *Curr. Opin. Chem. Biol.* **11,** 381–387 (2007).
11. Hwang, N. S., Varghese, S. & Elisseeff, J. Controlled differentiation of stem cells. *Adv. Drug Deliv. Rev.* **60,** 199–214 (2008).
12. Dawson, E., Mapili, G., Erickson, K., Taqvi, S. & Roy, K. Biomaterials for stem cell differentiation. *Adv. Drug Deliv. Rev.* **60,** 215–228 (2008).
13. Dellatore, S. M., Garcia, A. S. & Miller, W. M. Mimicking stem cell niches to increase stem cell expansion. *Curr. Opin. Biotechnol.* **19,** 534–540 (2008).
14. Little, L., Healy, K. E. & Schaffer, D. V. Engineering biomaterials for synthetic neural stem cell microenvironments. *Chem. Rev.* **108,** 1787–1796 (2008).
15. Burdick, J. A. & Vunjak-Novakovic, G. Engineered microenvironments for controlled stem cell differentiation. *Tissue Eng. A* **15,** 205–219 (2009).
16. Flaim, C. J., Chien, S. & Bhatia, S. N. An extracellular matrix microarray for probing cellular differentiation. *Nature Methods* **2,** 119–125 (2005).
17. Soen, Y., Mori, A., Palmer, T. D. & Brown, P. O. Exploring the regulation of human neural precursor cell differentiation using arrays of signaling microenvironments. *Mol. Syst. Biol.* **2,** 37 (2006).
This paper presents an approach to probing quantitatively the effects of molecular signals and signal combinations on stem-cell fate decisions.
18. Derda, R. *et al.* Defined substrates for human embryonic stem cell growth identified from surface arrays. *ACS Chem. Biol.* **2,** 347–355 (2007).
19. LaBarge, M. A. *et al.* Human mammary progenitor cell fate decisions are products of interactions with combinatorial microenvironments. *Integr. Biol.* **1,** 70–79 (2009).
20. Irvine, D. J., Hue, K. A., Mayes, A. M. & Griffith, L. G. Simulations of cell-surface integrin binding to nanoscale-clustered adhesion ligands. *Biophys. J.* **82,** 120–132 (2002).
21. Nur-E-Kamal, A. *et al.* Covalently attached FGF-2 to three-dimensional polyamide nanofibrillar surfaces demonstrates enhanced biological stability and activity. *Mol. Cell. Biochem.* **309,** 157–166 (2008).
22. Fan, V. H. *et al.* Tethered epidermal growth factor provides a survival advantage to mesenchymal stem cells. *Stem Cells* **25,** 1241–1251 (2007).
23. Alberti, K. *et al.* Functional immobilization of signaling proteins enables control of stem cell fate. *Nature Methods* **5,** 645–650 (2008).
This paper demonstrates the relevance of signalling-protein tethering to the fate of (embryonic) stem cells.
24. Suzuki, T. *et al.* Highly efficient *ex vivo* expansion of human hematopoietic stem cells using Delta1-Fc chimeric protein. *Stem Cells* **24,** 2456–2465 (2006).
25. Beckstead, B. L., Santosa, D. M. & Giachelli, C. M. Mimicking cell–cell interactions at the biomaterial–cell interface for control of stem cell differentiation. *J. Biomed. Mater. Res. A* **79,** 94–103 (2006).
26. Anderson, D. G., Levenberg, S. & Langer, R. Nanoliter-scale synthesis of arrayed biomaterials and application to human embryonic stem cells. *Nature Biotechnol.* **22,** 863–866 (2004).
27. Webster, C., Silberstein, L., Hays, A. P. & Blau, H. M. Fast muscle fibers are preferentially affected in Duchenne muscular dystrophy. *Cell* **52,** 503–513 (1988).
28. Discher, D. E., Janmey, P. & Wang, Y. L. Tissue cells feel and respond to the stiffness of their substrate. *Science* **310,** 1139–1143 (2005).
29. Paszek, M. J. *et al.* Tensional homeostasis and the malignant phenotype. *Cancer Cell* **8,** 241–254 (2005).
30. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126,** 677–689 (2006).
This paper demonstrates the important role of matrix stiffness in the fate of (mesenchymal) stem cells.
31. Saha, K. *et al.* Substrate modulus directs neural stem cell behavior. *Biophys. J.* **95,** 4426–4438 (2008).
32. Boonen, K. J., Rosaria-Chak, K. Y., Baaijens, F. P., van der Schaft, D. W. & Post, M. J. Essential environmental cues from the satellite cell niche: optimizing proliferation and differentiation. *Am. J. Physiol. Cell Physiol.* **296,** C1338–C1345 (2009).
33. Li, Y. J., Chung, E. H., Rodriguez, R. T., Firpo, M. T. & Healy, K. E. Hydrogels as artificial matrices for human embryonic stem cell self-renewal. *J. Biomed. Mater. Res. A* **79,** 1–5 (2006).
34. Folkman, J. & Moscona, A. Role of cell shape in growth control. *Nature* **273,** 345–349 (1978).
35. Chen, C. S., Mrksich, M., Huang, S., Whitesides, G. M. & Ingber, D. E. Geometric control of cell life and death. *Science* **276,** 1425–1428 (1997).
36. Wozniak, M. A. & Chen, C. S. Mechanotransduction in development: a growing role for contractility. *Nature Rev. Mol. Cell Biol.* **10,** 34–43 (2009).
37. Chen, C. S., Alonso, J. L., Ostuni, E., Whitesides, G. M. & Ingber, D. E. Cell shape provides global control of focal adhesion assembly. *Biochem. Biophys. Res. Commun.* **307,** 355–361 (2003).
38. McBeath, R., Pirone, D. M., Nelson, C. M., Bhadriraju, K. & Chen, C. S. Cell shape, cytoskeletal tension, and RhoA regulate stem cell lineage commitment. *Dev. Cell* **6,** 483–495 (2004).
This paper highlights the role of cell-shape control in regulating the fate of (mesenchymal) stem cells.
39. Peerani, R. *et al.* Niche-mediated control of human embryonic stem cell self-renewal and differentiation. *EMBO J.* **26,** 4744–4755 (2007).
40. Chin, V. I. *et al.* Microfabricated platform for studying stem cell fates. *Biotechnol. Bioeng.* **88,** 399–415 (2004).
41. Mohr, J. C., de Pablo, J. J. & Palecek, S. P. 3-D microwell culture of human embryonic stem cells. *Biomaterials* **27,** 6032–6042 (2006).
42. Khademhosseini, A. *et al.* Co-culture of human embryonic stem cells with murine embryonic fibroblasts on microwell-patterned substrates. *Biomaterials* **27,** 5968–5977 (2006).

© 2009 Macmillan Publishers Limited. All rights reserved

43. Karp, J. M. *et al.* Controlling size, shape and homogeneity of embryoid bodies using poly(ethylene glycol) microwells. *Lab Chip* **7,** 786–794 (2007).
44. Moeller, H. C., Mian, M. K., Shrivastava, S., Chung, B. G. & Khademhosseini, A. A microwell array system for stem cell culture. *Biomaterials* **29,** 752–763 (2008).
45. Ungrin, M. D., Joshi, C., Nica, A., Bauwens, C. & Zandstra, P. W. Reproducible, ultra high-throughput formation of multicellular organization from single cell suspension-derived human embryonic stem cell aggregates. *PLoS ONE* **3,** e1565 (2008).
46. Dykstra, B. *et al.* High-resolution video monitoring of hematopoietic stem cells cultured in single-cell arrays identifies new features of self-renewal. *Proc. Natl Acad. Sci. USA* **103,** 8185–8190 (2006).
47. Cordey, M., Limacher, M., Kobel, S., Taylor, V. & Lutolf, M. P. Enhancing the reliability and throughput of neurosphere culture on hydrogel microwell arrays. *Stem Cells* **26,** 2586–2594 (2008).
48. Lutolf, M. P., Doyonnas, R., Havenstrite, K., Koleckar, K. & Blau, H. M. Perturbation of single hematopoietic stem cell fates in artificial niches. *Integr. Biol.* **1,** 59–69 (2009).
This paper presents a combination of *in vitro* and *in vivo* methods to deconstruct a stem-cell niche and probe the effects of its individual key components on the fate of single (haematopoietic) stem cells.
49. Jia, X. & Kiick, K. L. Hybrid multicomponent hydrogels for tissue engineering. *Macromol. Biosci.* **9,** 140–156 (2009).
50. Albrecht, D. R., Underhill, G. H., Wassermann, T. B., Sah, R. L. & Bhatia, S. N. Probing the role of multicellular organization in three-dimensional microenvironments. *Nature Methods* **3,** 369–375 (2006).
This paper presents an interesting approach to the micropatterning of cells in 3D hydrogel microenvironments.
51. Lutolf, M. P. & Blau, H. M. in *Advances in Tissue Engineering* (ed. Polak, J.) Ch. 9 (World Scientific, in the press).
52. Lutolf, M. P. & Blau, H. M. in *Mater. Res. Soc. Symp. Proc.* Vol. 1140 (eds Prasad Shastri, V., Lendlein, A., Liu, L., Mikos, A. & Mitragotri, S.) 1140-HH07-07 (Materials Research Society, 2009).
53. Lutolf, M. P. Artificial ECM: expanding the cell biology toolbox in 3D. *Integr. Biol.* **1,** 235–241 (2009).
54. Hennink, W. E. & van Nostrum, C. F. Novel crosslinking methods to design hydrogels. *Adv. Drug Deliv. Rev.* **54,** 13–36 (2002).
55. Kopecek, J. & Yang, J. Y. Hydrogels as smart biomaterials. *Polym. Int.* **56,** 1078–1098 (2007).
56. Silva, G. A. *et al.* Selective differentiation of neural progenitor cells by high-epitope density nanofibers. *Science* **303,** 1352–1355 (2004).
57. Lin, C. C. & Anseth, K. S. PEG hydrogels for the controlled release of biomolecules in regenerative medicine. *Pharm. Res.* **26,** 631–643 (2009).
58. Underhill, G. H. & Bhatia, S. N. High-throughput analysis of signals regulating stem cell fate and function. *Curr. Opin. Chem. Biol.* **11,** 357–366 (2007).
59. Gidrol, X. *et al.* 2D and 3D cell microarrays in pharmacology. *Curr. Opin. Pharmacol.* **9,** 664–668 (2009).
60. Fernandes, T. G., Diogo, M. M., Clark, D. S., Dordick, J. S. & Cabral, J. M. S. High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research. *Trends Biotechnol.* **27,** 342–349 (2009).
61. Lee, M. Y. *et al.* Three-dimensional cellular microarray for high-throughput toxicology assays. *Proc. Natl Acad. Sci. USA* **105,** 59–63 (2008).
62. Jongpaiboonkit, L., King, W. J. & Murphy, W. L. Screening for 3D environments that support human mesenchymal stem cell viability using hydrogel arrays. *Tissue Eng. A* **15,** 343–353 (2009).
63. Sudo, R. *et al.* Transport-mediated angiogenesis in 3D epithelial coculture. *FASEB J.* **23,** 2155–2164 (2009).
64. Tam, P. P. L. & Loebel, D. A. F. Gene function in mouse embryogenesis: get set for gastrulation. *Nature Rev. Genet.* **8,** 368–381 (2007).
65. Whitesides, G. M. The origins and the future of microfluidics. *Nature* **442,** 368–373 (2006).
66. Chung, B. G. *et al.* Human neural stem cell growth and differentiation in a gradient-generating microfluidic device. *Lab Chip* **5,** 401–406 (2005).
67. Choi, N. W. *et al.* Microfluidic scaffolds for tissue engineering. *Nature Mater.* **6,** 908–915 (2007).
This paper is a good example of how microfluidic technology can be used to generate well-controlled protein gradients in 3D cell matrices.
68. Peret, B. J. & Murphy, W. L. Controllable soluble protein concentration gradients in hydrogel networks. *Adv. Funct. Mater.* **18,** 3410–3417 (2008).
69. van Noort, D. *et al.* Stem cells in microfluidics. *Biotechnol. Prog.* **25,** 52–60 (2009).
70. Gomez-Sjoberg, R., Leyrat, A. A., Pirone, D. M., Chen, C. S. & Quake, S. R. Versatile, fully automated, microfluidic cell culture system. *Anal. Chem.* **79,** 8557–8563 (2007).
71. Lii, J. *et al.* Real-time microfluidic system for studying mammalian cells in 3D microenvironments. *Anal. Chem.* **80,** 3640–3647 (2008).
72. Hahn, M. S., Miller, J. S. & West, J. L. Three-dimensional biochemical and biomechanical patterning of hydrogels for guiding cell behavior. *Adv. Mater.* **18,** 2679–2684 (2006).
73. Wosnick, J. H. & Shoichet, M. S. Three-dimensional chemical patterning of transparent hydrogels. *Chem. Mater.* **20,** 55–60 (2008).
74. Kloxin, A. M., Kasko, A. M., Salinas, C. N. & Anseth, K. S. Photodegradable hydrogels for dynamic tuning of physical and chemical properties. *Science* **324,** 59–63 (2009).
This paper presents a powerful method of influencing stem-cell fate by locally manipulating the biochemical and biophysical properties of a 3D hydrogel matrix.
75. Gillette, B. M. *et al.* *In situ* collagen assembly for integrating microfabricated three-dimensional cell-seeded matrices. *Nature Mater.* **7,** 636–640 (2008).
76. Khetani, S. R. & Bhatia, S. N. Engineering tissues for *in vitro* applications. *Curr. Opin. Biotechnol.* **17,** 524–531 (2006).
77. Mironov, V., Kasyanov, V., Drake, C. & Markwald, R. R. Organ printing: promises and challenges. *Regen. Med.* **3,** 93–103 (2008).
78. Lee, W. *et al.* Three-dimensional bioprinting of rat embryonic neural cells. *Neuroreport* **20,** 798–803 (2009).
79. Mooney, D. J. & Vandenburgh, H. Cell delivery mechanisms for tissue repair. *Cell Stem Cell* **2,** 205–213 (2008).
80. Conboy, I. M. *et al.* Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature* **433,** 760–764 (2005).
81. Adams, G. B. *et al.* Therapeutic targeting of a stem cell niche. *Nature Biotechnol.* **25,** 238–243 (2007).
82. Zhang, L. *et al.* Nanoparticles in medicine: therapeutic applications and developments. *Clin. Pharmacol. Ther.* **83,** 761–769 (2008).
83. Rothenfluh, D. A., Bermudez, H., O'Neil, C. P. & Hubbell, J. A. Biofunctional polymer nanoparticles for intra-articular targeting and retention in cartilage. *Nature Mater.* **7,** 248–254 (2008).
84. Gu, F. *et al.* Precise engineering of targeted nanoparticles by using self-assembled biointegrated block copolymers. *Proc. Natl Acad. Sci. USA* **105,** 2586–2591 (2008).
85. Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131,** 324–336 (2007).
86. Gomi, K., Kanazashi, M., Lickorish, D., Arai, T. & Davies, J. E. Bone marrow genesis after subcutaneous delivery of rat osteogenic cell-seeded biodegradable scaffolds into nude mice. *J. Biomed. Mater. Res. A* **71A,** 602–607 (2004).
87. Eilken, H. M., Nishikawa, S.-I. & Schroeder, T. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* **457,** 896–900 (2009).
88. Glauche, I., Lorenz, R., Hasenclever, D. & Roeder, I. A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell Prolif.* **42,** 248–263 (2009).
89. Ravin, R. *et al.* Potency and fate specification in CNS stem cell populations *in vitro*. *Cell Stem Cell* **3,** 670–680 (2008).
90. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).

Acknowledgements We acknowledge a US National Institutes of Health (NIH) postdoctoral training grant (CA09151) (P.M.G.); a Leukemia and Lymphoma Society Fellowship (5367-07) and the European Young Investigator award (M.P.L.); and NIH grants (AG009521, AG020961 and HL096113), a Juvenile Diabetes Research Foundation Grant (34-2008-623), a Muscular Dystrophy Association grant (4320), a California Institute for Regenerative Medicine (CIRM) Tools and Technologies grant (RT1-01001), a Stanford Bio-X award (IIP3-34), and the Baxter Foundation (H.M.B.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to M.P.L. (matthias.lutolf@epfl.ch) or H.M.B. (hblau@stanford.edu).

© 2009 Macmillan Publishers Limited. All rights reserved

NATURE|Vol 462|26 November 2009|doi:10.1038/nature08603

Biomaterial systems for mechanosensing and actuation

Peter Fratzl[1] & Friedrich G. Barth[2]

Living organisms use composite materials for various functions, such as mechanical support, protection, motility and the sensing of signals. Although the individual components of these materials may have poor mechanical qualities, they form composites of polymers and minerals with a remarkable variety of functional properties. Researchers are now using these natural systems as models for artificial mechanosensors and actuators, through studying both natural structures and their interactions with the environment. In addition to inspiring the design of new materials, analysis of natural structures on this basis can provide insight into evolutionary constraints on structure–function relationships in living organisms and the variety of structural solutions that emerged from these constraints.

Nature has always been a source of inspiration for technical developments, but only in recent years have materials scientists started to consider the complex hierarchical structure of natural materials as a model for the development of new types of high-performance engineering materials[1,2]. It is by no means obvious how the lessons learned from biological materials can be applied to the design of new engineering materials. The reasons for this difficulty are some striking differences in the design strategies that are common in engineering and the ways in which natural materials are constructed[2,3]. First, the choice of elements is by far greater for the engineer. Elements such as iron, chromium and nickel are only trace elements in biological tissues and are not used in metallic form, in contrast to their use in different types of steel, for example. A second difference is in the way that materials are actually made. Whereas the engineer takes a 'top-down' approach, selecting a material to fabricate a part according to an established goal and an exact design plan, natural structures develop in the opposite way (that is, 'bottom up'). Both the material and the whole organism (a plant or an animal) grow according to the principles of biologically controlled self-assembly. This provides control over the structure of the material at all times and levels of hierarchy, and it is the key to the outstanding success with which composites are used as structural materials in nature. Therefore, many crucial features of biological materials are worth understanding[3]. These include their hierarchical structure[4], their remarkable fracture resistance[5,6], their multifunctional[7] or adaptive[8–10] properties, and their self-healing capacity[11].

To derive useful biologically inspired strategies in materials science, it is not sufficient simply to observe naturally occurring structures. Organisms have to cope with a multitude of environmental constraints that typically differ greatly from the boundary conditions to be met by materials developed for technological application. In an animal or a plant, a material evolved to serve a mechanical function also needs to fulfil many other criteria. For example, it may have to grow under temperature, pressure and pH conditions indispensable for the existence of life. Also, the material's constituents have to be available in the habitat of the species in question. Furthermore, the material may have to serve not only structural functions but also functions needed for camouflage, signalling, defence against bacteria and parasites, and so on. The constraints on the manufacture of an engineering material are drastically different. They include consumer acceptance, compatibility with other technical systems, and the time and cost of manufacturing. Whereas the engineer generally knows the constraints and selects a material appropriately, biomimetic materials science requires the study of a pre-existing natural material that represents the solution to an unknown, multifaceted problem, which makes the transfer of principles to materials engineering more difficult. The relationship between the function of the biological material and its structure and composition has to be fully established before any principle useful in materials science can be extracted. As a result, there is no biomimetic materials research without proper biological research, including a thorough analysis of what a material is made for under the conditions of the organism's species-specific behaviour and ecological situation. A consequence of this is that although the study of materials in organisms may inspire radically new materials designs, it does not lead to rapid solutions in materials engineering.

The combination of materials science and biology also contributes significantly to the biological understanding of organisms by helping to establish structure–function relationships: developing mathematical or technical models of biological systems helps to clarify the function of their components. It may even allow quantitative predictions about the evolutionary role and relative importance of certain parameters in the development of particular functions enforced by natural selection[12].

Here we present the sophisticated mechanosensory systems of spiders and actuation systems in plants as examples that illustrate the potential of research combining engineering with biology. We discuss vibration, tactile and airflow sensors in spiders; the snapping system in the Venus flytrap; a hydration-driven motor in wheat awns; and a system for controlled bending in trees. In all these cases, important functional characteristics are based on material properties closely matching biological needs, and impressive material and structural solutions, sometimes of deceptive simplicity, have been discovered.

Sensors and actuators

Living organisms depend on real-time and stored information about their internal and external worlds, in particular — in animals — for the purpose of moving around to secure the energy necessary for their metabolism and reproduction. As a consequence, there is a rich diversity

[1]Max Planck Institute of Colloids and Interfaces, Department of Biomaterials, Research Campus Golm, 14424 Potsdam, Germany. [2]University of Vienna, Department of Neurobiology and Cognition Research, Althanstrasse 14, 1090 Vienna, Austria.

© 2009 Macmillan Publishers Limited. All rights reserved

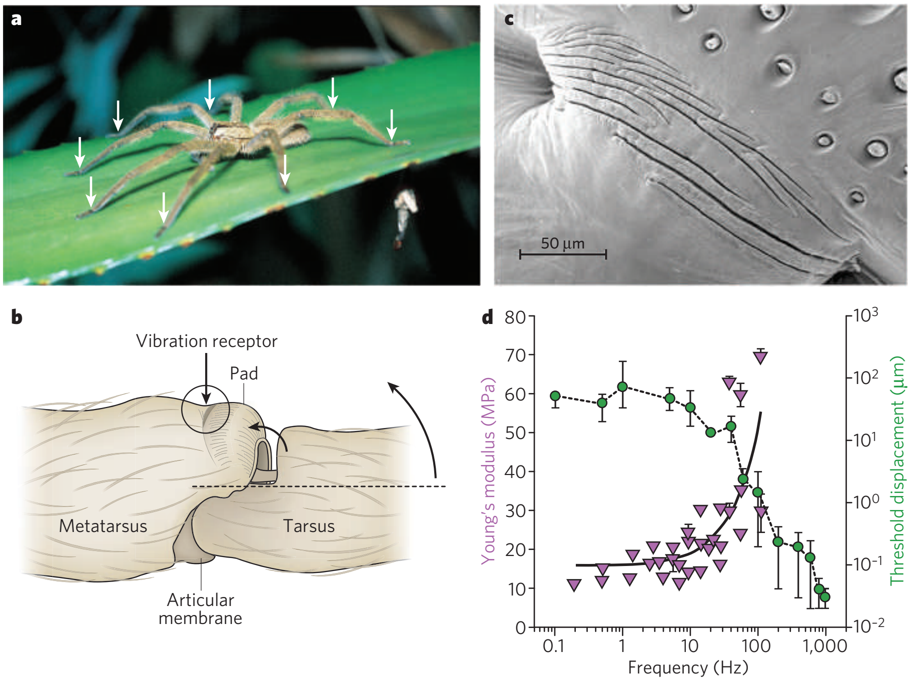


Figure 1 | Vibration-sensitive slit organ. **a**, *Cupiennius salei*, with arrows pointing to the location of the vibration sensors, on the legs[14]. **b**, The vibration sensor dorsal on the metatarsus is stimulated by compression following the upwards movement of the tarsus, indicated by the two curved arrows. **c**, Scanning electron micrograph of the vibration detector (dorsal view, area depicted in circle in **b**). **d**, Young's modulus of the pad material (purple) as a function of vibration frequency, compared with the vibration sensor's physiological threshold curve (green). Error, 1 s.d. (Panel **a** reproduced, with permission, from ref. 14. Panel **b**–**d** reproduced, with permission, from ref. 19.)

of biosensors and bio-actuators, in organisms ranging in complexity from the level of bacteria to that of humans. In many ways, the processes of mechanical sensing are the reverse of those by which actuators generate mechanical forces. Both the interpretation of mechanical signals such as tactile input and the generation of complex movement patterns require sophisticated data processing. It is becoming increasingly clear that an enormous amount of sensory filtering takes place in the periphery (that is, outside the central nervous system), in particular in less complex animals, such as insects and spiders, which often have only tiny brains. This implies that to a large degree the central nervous system is relieved of the task of recognizing the biologically relevant stimulus patterns in a chronically noisy environment. To understand this, it is essential to identify sensory organs as matched filters reflecting particular stimulus patterns typical of, and relevant to, a particular species. Similarly, the clever design of actuators often allows complex movements without complex signal processing. Although in this way the flexibility of the response possible with complex signal processing may be lost, the advantage gained may be a much greater response speed, sometimes — as in the case of certain plants — relying entirely on the environment to drive the required actuation.

Lessons from spiders in detecting mechanical stimuli

Many spiders live in a world of substrate vibrations. Vibrations of the spider web signal the presence of entangled insect prey, and self-generated vibrations are actively used as courtship signals to communicate with a prospective mate. As a result, spiders are as sensitive to vibration as the most sensitive species in the animal kingdom, such as the cockroach and the scorpion (human vibration sensitivity being modest in comparison)[13,14]. Their most important vibration detector is based on a slit system embedded in the exoskeleton. The slit system locally enhances the mechanical compliance of the exoskeleton and thus allows deformation of the slit by tiny forces, even though the spider's exoskeleton is made of a material with a stiffness close to that of bone[15,16]. By contrast, mechanosensitive hair-like sensilla protrude from the exoskeleton's surface, functioning as sensors by way of the hair's deflection either by direct contact forces or by the frictional forces of airflow.

Lyriform slit systems

The most intensively studied case of vibratory courtship in a spider is that of *Cupiennius salei*[17], which is a large Central American wandering spider that lives on plants such as bromeliads instead of in a web. It is nocturnal and leaves its retreat on the plant in the dark to prey or to court elsewhere on the plant (Fig. 1a). On encountering the pheromone-laden safety thread of a female, the male starts scratching the plant with his pedipalps (short appendages close to the mouth) and oscillating his opisthosoma (equivalent to the abdomen in insects), thereby introducing vibrations into the plant through his eight legs. These vibrations are dominated by low-frequency components of around 80–100 Hz and travel as bending waves through the plant to the female receiver. The female may be several metres away (on a large plant) and still receive the message and respond to it with her own vibrations, which are of even lower frequency (~30 Hz). She thereby maintains contact with the male and guides him to her location. The physical properties of both the sender's signal and the signal-transmitting plant structures are well matched to the receiver's vibration sensor. The use of low frequencies is important because their attenuation is much less pronounced (in the range of 0.3 dB cm^{-1}) than that of frequencies higher than a few hundred hertz.

As in all sensory systems, a key problem is recognizing the relevant signal among the ever-present background noise. One way to overcome this problem is to be insensitive to frequencies typical of such background noise (which in the given case has peaks usually well below 10 Hz (ref. 18)). The spider vibration sensor is a high-pass filter: it has a low threshold sensitivity at up to about 10–40 Hz, and its sensitivity increases by three to four orders of magnitude at higher frequencies (about 100 fold for every 10-fold increase in frequency), reaching 10 nm to 1 nm at a stimulation frequency of 1 kHz (that is, movement by these distances can be sensed). In terms of substrate deflection, the spider is particularly sensitive to the frequencies contained in courtship and prey signals but not to those of vibrations of abiotic origin (for example those due to wind).

This explains the biological significance of the high-pass characteristic of the sensitivity curve (Fig. 1d). Its origin lies in the key vibration sensor of spiders, a compound, or lyriform, slit sense organ (Fig. 1c). This type of sensor is embedded in the exoskeleton and measures minute cuticular strains as low as a few microepsilons (10^{-6}) as they are caused by haemolymph pressure, muscular activity and, in the case being discussed, substrate vibrations. By compression of the slits, nervous impulses are set off in sensory cells associated with them. As seen in Fig. 1b, such compression results from an upward movement of the tarsus, the most distal leg segment, caused by substrate vibration. However, the tarsus does not transmit the stimulus directly to the vibration sensor; rather, it pushes against a cuticular pad that is located between the two and is well placed to filter the stimulus mechanically. This pad is largely responsible for the vibration sensor's high-pass properties.

© 2009 Macmillan Publishers Limited. All rights reserved

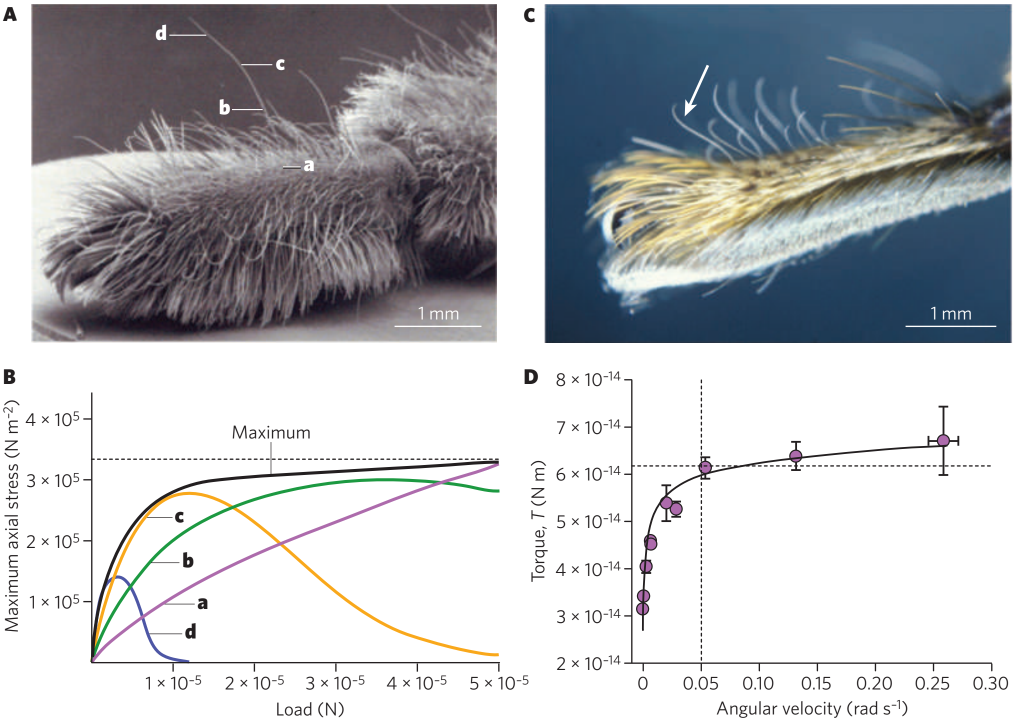


Figure 2 | Sensory hairs: tactile and airflow sensors. A, Final segment of a spider leg, with points on the tactile hair[26] indicated by letters, which refer to the curves in **B. B,** Dependence on increasing stimulus force (load) of maximum axial stresses at different positions along the hair[26]. **C,** Airflow sensor (arrow) on the final segment of the spider leg[14]. **D,** Torque measured when deflecting the hair, as a function of angular velocity. The dashed lines show the transition towards asymptotic behaviour. Errors, 1 s.d. (Panels **A** and **B**, reproduced, with permission, from ref. 26. Panel **C**, reproduced, with permission, from ref. 14. Panel **D** reproduced, with permission, from ref. 35.)

Atomic force microscopy and surface force spectroscopy applied to the pad in live spiders reveal the cuticular material to have a Young's modulus, *E*, of about 15 MPa at low frequencies. As the frequency is increased above about 30 Hz, however, the values rapidly increase, to about 70 MPa at 112 Hz (ref. 19) (Fig. 1d). This strong frequency dependence of the elastic modulus indicates the viscoelastic nature of the pad material and that its energy absorption is maximized at low frequencies as a result of the time-dependent relaxation of the macromolecular material in the vicinity of the glass transition[20]. Importantly, the glass-transition temperature of the pad material is estimated to be 25 ± 2 °C (ref. 19), which is near the normal temperature in the spider's habitat. The compliance of the pad material is greatest at frequencies below 30 Hz; it therefore absorbs energy well at low frequencies, filtering out environmental noise. This energy dissipation, however, decreases drastically as the Young's modulus increases at the higher frequencies typical of the vibrations relevant to the spider. The biological significance of the cuticular pad is suggested by a comparison of the frequency dependence of its elastic response with that of the vibration sensor's physiological threshold sensitivity (Fig. 1d).

Given that *Cupiennius* is a nocturnal animal, the pad should be stiffer and a better stimulus transmitter at lower, night-time, temperatures than at higher, daytime, temperatures; that is, the organ should be more sensitive when being so makes biological sense. We have electrophysiological evidence that this is the case. The Q_{10} temperature coefficient (the rate of change following a temperature increase by 10 °C) of the threshold vibration magnitude that elicits a nervous response varies between −1.3 and −5.5 in the tested temperature range between 14 °C and 32 °C, depending on stimulus frequency and the slit examined[21].

Mechanosensitive hair sensilla

Four hundred million years of evolution have brought about spider senses that impress through their perfect functional match with the specifics of biological needs. As is the case for the strain-sensitive vibration detector, for mechanosensitive hair sensilla much of this 'engineering' resides in structures of stimulus uptake and transformation. Again, the material is ideally suited to being fine-tuned to meet specific mechanical demands, owing to the wide spectrum of mechanical properties that the arthropod cuticle may assume. Cuticle is a composite material with fibre-reinforced laminations. It consists of microfibres (high-molecular-weight chitin polymers resembling plant cellulose) embedded in a protein matrix. The remarkable mechanical adaptability of cuticle is a result of differences in the degree of crosslinking of the matrix and is enhanced by the modification of various structural properties such as the proportions of fibres and matrix, the water content and the orientation of the fibres (for spider cuticle[22,23]).

To reveal the potential of the bauplan (structural characteristics) of a mechanosensitive hair sensillum and the potential in the variability of its material properties, we compare a tactile hair stimulated by contact forces with a wind-sensitive hair responding to the slightest movement of the surrounding air.

Typically, the tactile hair considered here (length, ~2.5 mm) (Fig. 2A) is stimulated from above when the spider is moving around at night on plants and in the small spaces characteristic of its dwelling plants. Contact forces deflect the hair shaft, which rotates around an axis close to its base. Dendrites of three sensory cells terminate near the inner end of the hair shaft. Their action potentials signal the stimulus properties (in particular the occurrence and velocity of hair deflection) to the central nervous system[24].

Such tactile hairs hit surfaces thousands of times at speeds of at least 11 cm s^{-1} (ref. 25) during a spider's adult lifetime, and it is vital that they be structured so as not to fracture under these impacts. Their design teaches us how to combine protection against overload with high sensitivity to small deflections[26]. Owing to the elastic restoring forces (the spring stiffness, *S*, is of the order of 10^{-8} N m rad^{-1}) at its suspension (that is, at the structures coupling the hair shaft to the exoskeleton), which the stimulating forces have to overcome, the hair shaft is not only deflected, as a rigid rod would be, but also bent. When the hair is pushed down and bends, the point of load introduction shifts towards the hair base. As a consequence, the effective lever arm and the stimulating moment decrease as the loading force increases. The bending moments therefore reach only about 20% of those expected for a rigid hair and saturate at about 4 × 10^{-9} N m. Additionally, at its base the hair is never deflected by more than ~12°. Similarly, the bending moment increases much more slowly with large loading forces than with small ones; this offers protection against breaking, as well as an extension of the mechanical working range and higher mechanical sensitivity for small deflections than for large.

Finite-element analysis has revealed additional 'engineering tricks'[26]. Under bending, axial surface stresses are the major stress component in the hair. Importantly, the largest stress values do not exceed ~3 × 10^5 N m^{-2},

© 2009 Macmillan Publishers Limited. All rights reserved

although during a loading cycle the loads introduced at the different contact points differ greatly and the section of maximum longitudinal stress moves along the hair (Fig. 2B, from a to d). As a result, critical stress values are avoided, and the hair shaft is a structure of uniform maximum strength. This is achieved through the change in the second moment of area, *J*, along the hair's length as seen from longitudinal sections of the shaft and their variation along its length[27]. The spider not only protects its tactile hair from breaking but also economizes on the hair's material and weight. According to finite-element analysis[24,27], the benefits of these 'engineering tricks' are only gained if the model value of the Young's modulus of the cuticular material making up the hair shaft is close to the actual material value (~18 GPa). In conclusion, these hairs are 'well-designed' light-weight structures whose key properties are the change in the hair diameter along its length, the ability to bend away from the stimulus, and the adjustment of the suspension's stiffness and the hair shaft's Young's modulus[27].

A hair sensitive to the movement of the surrounding medium (air) and deflected by the slightest frictional forces of the medium particles has to be different from a tactile hair. Such sensory hairs are called trichobothria in spiders (Fig. 2C) and filiform hairs in insects. Their most remarkable property is their absolute sensitivity, which ranks them among the most sensitive biological sensors known. Work in the range of only 10^{-20} J is needed to drive the hair over one oscillation cycle and to elicit an action potential. This work is in the range of $k_B T$ (k_B, Boltzmann constant; T, ambient temperature), which for comparison can be thought of as a fraction of the energy contained in a single quantum of green light[28–30]. These exquisitely fine hairs, which in *Cupiennius* are ~10 μm in diameter at the base and 0.1–1.4 mm long, work close to the fundamental limits imposed by thermal noise[28–30].

From a materials point of view, the most significant difference between these hairs and the tactile hairs is the high flexibility of their suspension in the exoskeleton. Specifically, the value of the spring stiffness, *S*, is smaller by about four orders of magnitude (that is, of the order of 10^{-12} N m rad^{-1}) than that of the tactile hair. Extremely low values were calculated for the damping constant, *R*, as well (of the order of 10^{-15} N m s rad^{-1})[12,31]. As a consequence, the trichobothrium does not bend when driven by the frictional forces of the air. This 'perfected' interaction between the air and the hair is interpreted largely using fluid mechanics and has been studied intensively in many ways, including mathematical modelling[12,14].

From a biological perspective, the correlation between sensor properties and their behavioural significance is again of particular interest. *Cupiennius* uses information extracted from air movements by its many trichobothria (about 90 on each of its legs) when catching prey. When alerted by the airflow generated by a flying insect, the spider jumps into the air to catch it, implying that it not only detects its prey but also can determine its position. As in the case of substrate vibrations, the airflow signal has to be distinguished from noise. Whereas background air movement during the spider's nocturnal activity is dominated by very low frequencies (<10 Hz) and low flow velocities (<0.1 m s^{-1}) with little fluctuation (<15%), an effective prey stimulus fluctuates highly in flow velocity, with root-mean-squared values from ~25% to >50%, velocities of up to 1 m s^{-1} and a frequency range that is much broader and extends higher than 100 Hz (refs 14, 32–34).

Recently, the mechanical properties of the hair suspension were measured directly by using surface force spectroscopy and applying directly calibrated forces in the range of nanonewtons. According to these measurements, combined with viscoelastic modelling[35], the torque resisting hair motion, *T*, and its time rate of change, are highly dependent on hair deflection velocity (Fig. 2D). From the perspective of viscoelastic materials, this is not surprising. However, it is unexpected from a biological point of view: the torque needed to deflect the hair at low angular velocities (of the order of 10^{-4} rad s^{-1}), ~3×10^{-14} N m, is only about half of that needed at higher angular velocities (of the order of 10^{-1} rad s^{-1}), ~6×10^{-14} N m. The oscillatory nature of the deflections of a trichobothrium under natural stimulus conditions is well supported by the mechanical properties of the hair suspension. The evidence for this is as follows[35].

First, angular velocities of hair motion due to natural stimulation span a broad range with peak values of up to 150 rad s^{-1} and few values below 0.05 rad s^{-1}. In the biologically relevant range of velocities, the suspension behaves like a three-parameter Kelvin solid[35]. Because the velocities tend to be high, the viscous element will resist deformation, and the entire torque will be carried by the elastic elements and will change linearly with the hair's angular deflection, as is observed. The main spring element in the hair suspension is thought to reside in the membrane connecting the hair shaft to the exoskeleton. It is likely to contain resilin, which is an elastomeric protein that is known for its efficiency in elastically storing energy and whose elastic modulus (Young's modulus) increases with frequency[36].

Second, the viscoelastic behaviour of the hair suspension at low angular velocities, when the dashpot (a damper that resists motion by way of viscous friction) deforms and leads to a lower torque being necessary to attain a certain hair deflection, facilitates the start of hair motion from rest. This underlines the phasic nature of the system and supports its mechanical response to the highly fluctuating prey signals with frequent changes in velocity (Fig. 2D). The sources of the damping properties of the hair suspension have not yet been identified. A likely candidate is the receptor lymph surrounding the inner lever arm of the hair shaft and the structures coupling the sensory cell dendrites to it. Displacement of the lymph in such a confined space is expected to be highly viscous. From these arguments, we conclude that the suspension of the trichobothria is well adapted to detect the highly turbulent prey signals preferentially. These arguments also complement physiological data demonstrating the strictly phasic nature of the nervous response of the sensory cells, as well as central nervous system interneurons[37], and they complement insights into the air–hair interaction derived from fluid mechanics[12].

Considering the sensitivity, selectivity, ruggedness and miniaturization common in biological sensors, it is not surprising that the production of synthetic sensors based on biological principles has long been a goal. However, only recently — with the advent of new materials, including soft elastomeric materials, and new microfabrication techniques — has designing bioinspired mechanosensors become a realistic goal[38,39]. Furthermore, our increasingly deep understanding of the relevant biological principles and advances in the mathematical modelling of sensor physics have advanced bioinspired sensor technology to a point where it seems justified to expect a wealth of innovation in the near future. There have been efforts by several laboratories to apply principles underlying hair-like sensilla to the design of electromechanical medium-motion sensors. Much of the motivation for this derives from the bulkiness of the available technical sensors and the difficulty in making spatially highly resolved measurements of fluids both close to a surface and simultaneously at many points, as for instance occurs in the sensor arrays of spiders in air and the sensors of the fish lateral-line system in water[40,41].

Two examples illustrate these engineering efforts[12]. The first is inspired by arthropod filiform hairs — in particular those of the cricket — and works in air[42,43]. High-aspect-ratio hairs made of an epoxy-based polymer (SU-8) and up to 1 mm long are supported by membranes made of Si_xN_y whose stiffness can be varied, changing the mechanical sensitivity to the hair's deflection. The actual transduction of the hair shaft's deflection in response to an electrical signal is accomplished capacitatively by using electrodes on the membrane's inner surface that move against the facing substrate. Recently, a similar sensor was given directionality by the arrangement of four capacitors at the hair base, reminiscent of the four sensory cells attached to the base of a spider trichobothrium[12,43]. The second example is mainly inspired by the fish lateral-line system and works in water[44–46]. In this case, a post up to ~1 mm in length and made either of gold with electroplated permalloy or of SU-8 is attached vertically to a horizontally oriented silicon cantilever beam. Transduction is by piezoresistive strain gauges at the cantilever's base, which is bent by the fluid forces acting on the post.

Pointing to the equivalence of sensing and actuating, microcantilevers based on microelectromechanical systems technology and serving sensory purposes have also been developed to serve as microactuators. These hair-like actuators are, for instance, driven electrostatically[47] or magnetically[48] and used to move and mix extremely small volumes of fluid. Polyimide bimorph microactuators consisting of two layers with different thermal

© 2009 Macmillan Publishers Limited. All rights reserved

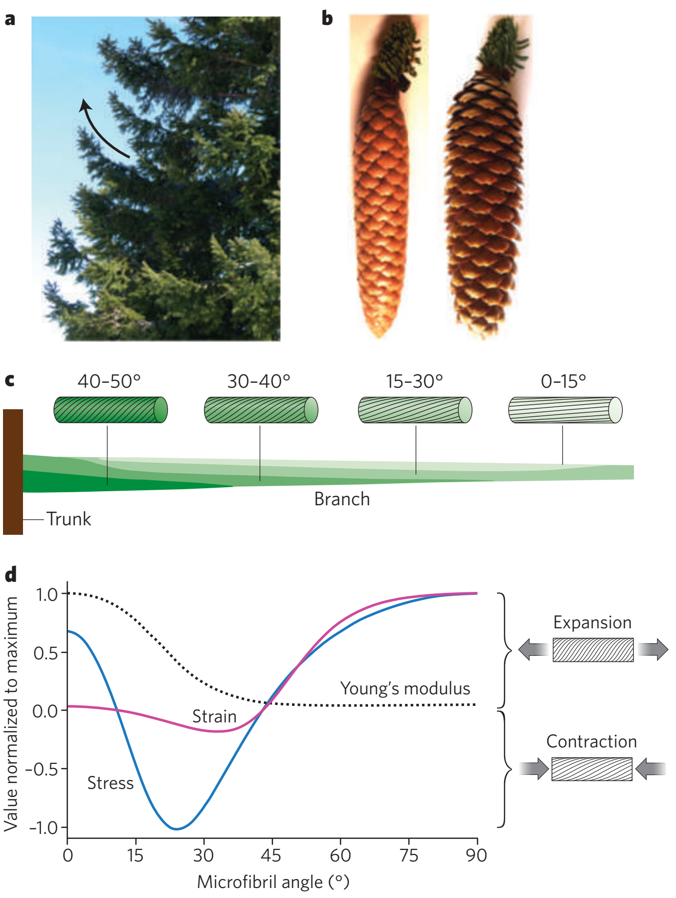


Figure 3 | Common actuators in spruce trees based on cell-wall swelling. **a**, Spruce branches typically curve upwards (arrow). **b**, Depending on the ambient humidity, spruce cones change shape, opening when they dry (right) and closing when rewetted (left). **c**, Distribution of cellulose microfibril angles in a spruce branch (based on data from ref. 63). Tube-like cells have a thick cell wall containing cellulose fibrils embedded in a matrix of hemicelluloses and lignin. Within the main part of the cell wall (the S2 layer), the cellulose fibrils have a spiral arrangement around the central lumen of the cell. The spiral angle with respect to the cell axis, called the microfibril angle, varies throughout the branch as indicated by the colour distribution. **d**, Schematic illustration of the effect of the microfibril angle on the swelling behaviour of cylindrical cells. The blue and purple curves show the variation of the axial stress (when the cell is not allowed to change length) and the axial strain (when no constraint is put on the cell), respectively. The black dashed curve shows the Young's modulus of the cells. For microfibril angles larger than 45°, the cells expand in the axial direction; whereas, for smaller microfibril angles, they contract. This implies that, on swelling, cells on the lower side of the branch expand in the axial direction, whereas cells on the upper side contract. We note that the largest contractile stress and contractile strain occur at different microfibril angles (minima of the blue and purple curves, respectively), implying that the optimal configurations for generating contractile stresses and movement are different. The stiffness of the cylinder in the axial direction also depends to a large extent on the microfibril angle[52,64,65].

expansion coefficients and driven by resistive heaters have been designed for the same purpose[49].

Lessons from plant tissue in complex actuation by swelling

Other ideas for actuator systems are inspired by plants that are able to move as a result of water absorption[50]. Plants, most of whose movements are slow, are not as highly dynamic as animals. Nonetheless, they have developed a variety of motion systems, mainly based on hygromorphic principles[51]. Some of these systems have been described and compared in a recent review[52]. Generally, water-based movements and water transport are essential functions in plants. The systems associated with these functions are potentially interesting for biomimetic research. Recently, a synthetic water transport system was devised in which the transpiration principle of trees was translated into a technological device[53]. The underlying principle is that water is lost by evaporation from the leaves, which creates a negative pressure inside the plant's water-conducting tissues and results in the water being pushed upwards[54]. This means that the energy needed to raise the water in a tree is directly provided by the sun evaporating water at the leaves' surfaces. The key problem, which was not resolved until recently, is to prevent the disruption of the water column under the large tension necessary in a microfluidic water-conducting vessel in a high tree. This is an example of a system in which detailed elucidation of the physical principles underlying the biological process led to the construction of a synthetic device with the same properties[53]. Similar developments may be expected once the physical principles of plant actuators are fully understood.

One of the problems is that different types of plant movement, such as growth, seed dispersal and the catching of prey, occur at different speeds. This has been analysed, and plant movements have been classified[55] according to temporal and spatial scales. Water swelling and shrinking are generally responsible for slow and small-scale movements, whereas elastic instabilities cause fast and large-scale movements. Another way of discussing these differences is to consider the manner in which energy is stored before being transformed into the kinetic energy associated with movement. In movements in which metabolic processes are directly involved, chemical energy is typically used to fuel the actuation. This implies some delay in the generation of movement, because any chemical reaction requires a certain time to reach completion. This is true for plant movements induced by turgor pressure, in which an osmotic pressure is built up by active cell processes. Molecular motors in animal muscle cells also need chemical energy for contraction, limiting the contraction rate of muscles and therefore the speed of animal movement[56,57]. For extremely fast movements in plants, such as the propelling of seeds, energy has to be stored in a different, more immediately accessible, way. An interesting method is the storage of energy in elastic form, such as in a spring. This is also known in animals, for example in powering the jumping of a flea[58]. Processes resulting in slow movement, for example water absorption and desorption in a hygroscopic tissue, such as cellulose-based cell walls, may occur passively as a consequence of changing air humidity. Examples are the opening of pine cones[50,59] or the movement of wheat seeds[60].

In this section, we discuss two examples of plant movements recently described. In the first example, the swelling of the plant cell wall induces static stress, as in softwood branches (Fig. 3a, c, d), or slow movements, as in pine and spruce cones (Fig. 3b) and various seeds[60–62]. In the second example, an elastic instability is used to trigger a fast movement in the Venus flytrap, which catches insects mainly to increase its nitrogen supply (Fig. 4).

Movement by cell-wall swelling

Many actuation systems in plants have the common feature that the movement is generated by a differential swelling of different parts of the tissue, similar to the function of a bimetal strip measuring temperature. A passive movement of this kind was first described as a feature of the opening of pine cones[50,59], which happens when they dry in the air, similarly to the spruce cone shown in Fig. 3b. The basis for the differential swelling of different parts of the tissue is the intricate structure of the plant cell wall[4,52]. Not unlike the fibrous structure of arthropod cuticle, the major part of this wall is composed of cellulose fibrils that are just a few nanometres in diameter and are embedded in a hygroscopic matrix containing hemicelluloses and lignin. The cellulose fibrils wind in spirals around the central cell lumen. Their angle relative to the cell's long axis is called the microfibril angle (MFA). Figure 3c shows schematics of the distribution of cells with different MFAs in the branch of a spruce[63]. Cells with cellulose fibrils oriented almost in parallel to their long axes are located on the upper side of the branch, whereas on the lower side the MFA is close to 40°. Recent studies[64,65] have shown that in this system swelling generates internal stresses that bend the branch upwards as shown in Fig. 3a.

In this case, the physical origin of the anisotropic distribution of stresses is in simple geometric constraints. A section of cell wall that is

© 2009 Macmillan Publishers Limited. All rights reserved

swelling will extend less in the direction of the fibrils. As the fibrils are tilted relative to the cell's long axis, this constraint leads to the cell having more complex behaviour: it will either expand or contract depending on the MFA[64]. The results of simple model calculations[64] are summarized in Fig. 3d. At large MFAs, the cell expands longitudinally on swelling, as the cellulose fibres resisting strain along their axes are mostly arranged circumferentially. For MFAs less than 45°, however, the constraint imposed by the fibrils prevents the cell from expanding in the longitudinal direction (and the cell then swells predominantly laterally). This behaviour again reverses at very small MFAs. As a consequence, a simple geometric parameter, namely the cellulose MFA, regulates the local expansion behaviour. By combining cells with small MFAs on one side of an organ and cells with large MFAs on the other (Fig. 3c), the resultant differential expansion leads to a bending of the organ. In hardwoods, such as poplar, additional actuating mechanisms have evolved. A cellulose layer fills the lumen within the cylindrical cells and creates an internal pressure on swelling. This helps to generate tensile stresses on the upper side of the branch and makes the bending even more efficient[52,66].

Differential expansion is also the driving mechanism for the movement of wild wheat seeds[60,61]. The awns attached to these seeds bend with changing humidity by using a mechanism based on cell-wall swelling. Under daily-changing humidity conditions, the awns perform a swimming movement, propelling the seeds lying on the ground. The awns are covered with oriented silica spicules that generate direction-dependent friction and ensure that each awn moves in only one direction. We note that the requirements for large strains (as for movements of the wheat awn) and for large stresses (required in the tree branch) lead to different optimum values of the MFA (Fig. 3d). This is reminiscent of the lever-arm principle, according to which the same work can be converted into large forces and small displacements or vice versa, but force and displacement cannot both be large.

The wheat awn is an example in which the plant is also able to use solar energy (and the changes in air humidity it induces between morning, when dew increases humidity, and midday, when the air is drier) directly to propel the seeds. No active metabolism is involved in this movement. All the relevant cells in the awns are dead. Microstructured surfaces based on soft or stiff nanopillars embedded in hydrogels were recently developed as actuation or gripping devices that switch between states as the humidity varies[10,67]. In another attempt to mimic hygromorphic plant behaviour, plastic–paper bilayers have been used to construct model devices that open and close like flowers with changing humidity[50].

The snapping of the Venus flytrap

As already mentioned, fast plant movements are not possible with water swelling alone owing to the relatively long reaction and diffusion times involved. The closure of the Venus flytrap has been studied, and its physical principle has been modelled[68] (Fig. 4). This closure is a rapid movement (60% of the total displacement occurs in 0.1 s (ref. 68)) based on fluid flow between the inner and outer faces of the leaf and a consequent change in its natural curvature (Fig. 4). This process is an active one and is relatively slow. For suitable leaf dimensions, however, an intermediate situation may be reached in which the system is mechanically bistable and the leaf snaps shut through the release of elastic bending energy[68] (Fig. 4). This snapping movement is fast enough to catch the prey.

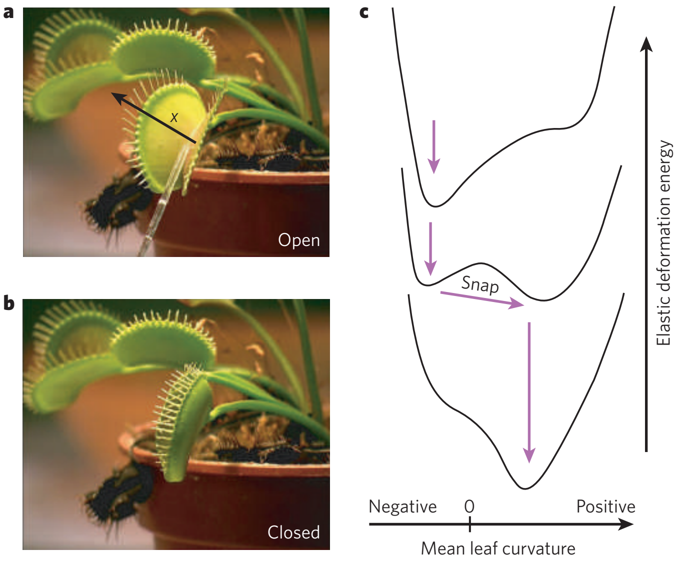


Figure 4 | Closure mechanism of the Venus flytrap. a, b, A Venus flytrap in its open (**a**) and closed (**b**) states. (Panels reproduced, with permission, from ref. 68.) **c,** After the prey touches the inner surface of the open leaf of a Venus flytrap, water flows between the inner and outer faces of the leaf, changing its curvature in the *x* direction (**a**). In addition, the elastic-energy profile changes from one for which the energy is minimum for the open leaf, which has negative curvature (top curve), to one for which the minimum corresponds to the closed state, which has positive curvature (bottom curve). For suitable values of leaf thickness and stiffness, an intermediate bistable state (centre curve) may occur, in which the leaf snaps rapidly from the open state to the closed state[68].

Outlook

The examples from spiders and plants presented here show the clever application of simple mechanical tricks in sensing and actuating systems of creatures generally considered much less complex than mammals. These organisms have either only a small nervous system (spiders) or none at all (plants). Therefore, signal processing by a central nervous system must have a comparatively reduced role. Despite this, both the exploration of the mechanical environment and the execution of mechanical actions are very efficient. The filtering of information from sensors and its transmission to actuators depend to a considerable degree on hardware consisting of versatile fibre-reinforced and laminated materials systems based on cellulose or chitin. Such sensory and actuation systems may be less dynamic than our human senses and muscles but have the advantage of greater autonomy. Some of the plant actuation systems described here function even without the support of a living organism, by taking advantage of changing air humidity cycles. Systems of this type may be extremely valuable for use in small robots or autonomously moving devices. The exploration of natural models has just begun, and only a few attempts have been made to create biomimetic devices; these have been based on hair sensors[38,39,42–46] or on gel-based fibre composite actuators[10]. In addition, simple model calculations[64] indicate that the orientation of fibres in these gels allows the type of movement to vary in arbitrarily complex ways. Fibre orientation also determines whether a given mechanical energy is transformed into free movement (that is, large deformation at small stresses) or into large internal stresses (at small or nearly zero deformation).

Another remarkable feature of natural sensory and actuator systems is that they are based on material systems with exceptional variability. This variability is controlled by several parameters, two important ones being differences in the degree of crosslinking of the matrix and the variety of possible fibre arrangements[69] of the chitin and cellulose fibres (in arthropods and plants, respectively) within the hierarchical structure of these composites[4]. As a consequence, it is barely possible to distinguish between material and structure in these biological systems, a lack of distinction that is another feature from which we can learn. By studying the rich variety of sensor and actuator systems in animals and plants from a materials point of view, both in more depth and also in a comparative way, we may discover new construction principles and material combinations with which to develop new types of microsensor and microactuator with a broad range of applications. Currently, efforts are under way to translate the microscopic principles of cell motility arising from protein-based systems such as flagella and molecular motors[70,71] into technical systems with many applications, from lab-on-a-chip devices to microrobotics and nanorobotics[72,73]. Biological solutions for mechanosensing and actuating on the basis of chitin and cellulose architectures will be an ideal complement to this work. ■

© 2009 Macmillan Publishers Limited. All rights reserved

1. Ortiz, C. & Boyce, M. C. Bioinspired structural materials. *Science* **319,** 1053–1054 (2008).
2. Aizenberg, J. & Fratzl, P. (eds) *Adv. Mater.* **21** (*Biological and Biomimetic Materials* special issue) (2009).
3. Fratzl, P. Biomimetic materials research: what can we really learn from nature's structural materials? *J. R. Soc. Interface* **4,** 637–642 (2007).
4. Fratzl, P. & Weinkamer, R. Nature's hierarchical materials. *Prog. Mater. Sci.* **52,** 1263–1334 (2007).
5. Munch, E. *et al.* Tough, bio-inspired hybrid materials. *Science* **322,** 1516–1520 (2008).
6. Bonderer, L. J., Studart, A. R. & Gauckler, L. J. Bioinspired design and assembly of platelet reinforced polymer films. *Science* **319,** 1069–1073 (2008).
7. Messersmith, P. B. Multitasking in tissues and materials. *Science* **319,** 1767–1768 (2008).
8. Vaia, R. & Baur, J. Adaptive composites. *Science* **319,** 420–421 (2008).
9. Capadona, J. R., Shanmuganathan, K., Tyler, D. J., Rowan, S. J. & Weder, C. Stimuli-responsive polymer nanocomposites inspired by the sea cucumber dermis. *Science* **319,** 1370–1374 (2008).
10. Sidorenko, A., Krupenkin, T., Taylor, A., Fratzl, P. & Aizenberg, J. Reversible switching of hydrogel-actuated nanostructures into complex micropatterns. *Science* **315,** 487–490 (2007).
    This paper describes an artificial system with actuation by a hydrogel reinforced with stiff elements.
11. van der Zwaag, S. (ed.) *Self Healing Materials. An Alternative Approach to 20 Centuries of Materials Science* (Springer, 2007).
12. Humphrey, J. A. C. & Barth, F. G. in *Advances in Insect Physiology* Vol. 34 (eds Casas, J. & Simpson, S. J.) 1–80 (Elsevier, 2008).
    This is an in-depth treatment of the biomechanics and physical–mathematical modelling of the sensing of medium motion by arthropod filiform hairs.
13. Barth, F. G. in *Springer Handbook of Auditory Research* Vol. 10 (eds Hoy, R. R., Popper, A. N. & Fay, R. R.) 228–278 (Springer, 1998).
14. Barth, F. G. *A Spider's World: Senses and Behavior* (Springer, 2002).
15. Hößl, B., Böhm, H. J., Rammerstorfer, F. G. & Barth, F. G. Finite element modeling of arachnid slit sensilla — I. The mechanical significance of different slit arrays. *J. Comp. Physiol. A* **193,** 445–459 (2007).
    This paper demonstrates the value of computational mechanics in an effort to understand the strange arrangements of strain-sensitive slits and their mechanical interaction in the spider exoskeleton.
16. Hößl, B., Böhm, H. J., Rammerstorfer, F. G., Mullan, R. & Barth, F. G. Studying the deformation of arachnid slit sensilla by a fracture mechanical approach. *J. Biomech.* **39,** 1761–1768 (2006).
17. Barth, F. G. in *Orientation and Communication in Arthropods* (ed. Lehrer, M.) 247–272 (Birkhäuser, 1997).
18. Barth, F. G., Bleckmann, H., Bohnenberger, J. & Seyfarth, E.-A. Spiders of the genus *Cupiennius* Simon 1891 (Araneae, Ctenidae). *Oecologia* **77,** 194–201 (1988).
19. McConney, M. E., Schaber, C. F., Julian, M. D., Barth, F. G. & Tsukruk, V. V. Viscoelastic nanoscale properties of cuticle contribute to the high-pass properties of spider vibration receptor (*Cupiennius salei* Keys). *J. R. Soc. Interface* **4,** 1135–1143 (2007).
    This paper describes a striking example of the role of non-nervous stimulus-conducting structures as mechanical filters and their match with biological needs.
20. Sperling, L. H. *Polymeric Multicomponent Materials: An Introduction* (Wiley, 1997).
21. Vogel, E. & Barth, F. G. *Vibrationsempfindlichkeit bei Cupiennius salei: Zum Einfluss efferenter nervöser Kontrolle und der Temperatur.* Master's thesis, Univ. Vienna (2009).
22. Neville, A. C. *Biology of the Arthropod Cuticle* (Springer, 1975).
23. Barth, F. G. Laminated composite material in biology. Microfiber reinforcement of an arthropod cuticle. *Z. Zellforsch. Mikrosk. Anat.* **144,** 409–433 (1973).
24. Albert, J. T., Friedrich, O. C., Dechant, H.-E. & Barth, F. G. Arthropod touch reception: spider hair sensilla as rapid touch detectors. *J. Comp. Physiol. A* **187,** 303–312 (2001).
25. Friedrich, O. C. *Zum Berührungssinn von Spinnen.* PhD thesis, Univ. Vienna (2001).
26. Dechant, H.-E., Rammerstorfer, F. G. & Barth, F. G. Arthropod touch reception: stimulus transformation and finite element model of spider tactile hairs. *J. Comp. Physiol. A* **187,** 313–322 (2001).
27. Dechant, H. E. *Mechanical Properties and Finite Element Simulation of Spider Tactile Hairs.* PhD thesis, Vienna Technical Univ. (2001).
28. Humphrey, J. A. C., Barth, F. G., Reed, M. & Spak, A. in *Sensors & Sensing in Biology & Engineering* (eds Barth, F. G., Humphrey, J. A. C. & Secomb, T. W.) 129–144 (Springer, 2003).
29. Shimozawa, T., Murakami, J. & Kumagai, T. in *Sensors & Sensing in Biology & Engineering* (eds Barth, F. G., Humphrey, J. A. C. & Secomb, T. W.) 145–158 (Springer, 2003).
30. Thurm, U. in *Biophysik* (eds Hoppe, W., Lohmann, W., Markl, H. & Ziegler, H.) 691–696 (Springer, 1982).
31. Barth, F. G., Wastl, U., Humphrey, J. A. C. & Devarakonda, R. Dynamics of arthropod filiform hairs. II. Mechanical properties of spider trichobothria (*Cupiennius salei* Keys). *Phil. Trans. R. Soc. Lond. B* **340,** 445–461 (1993).
32. Barth, F. G. & Höller, A. Dynamics of arthropod filiform hairs. V. The response of spider trichobothria to natural stimuli. *Phil. Trans. R. Soc. Lond. B* **354,** 183–192 (1999).
33. Barth, F. G., Humphrey, J. A. C., Wastl, U., Halbritter, J. & Brittinger, W. Dynamics of arthropod filiform hairs. III. Flow patterns related to air movement detection in a spider (*Cupiennius salei* Keys). *Phil. Trans. R. Soc. Lond. B* **347,** 397–412 (1995).
34. Klopsch, C., Barth, F. G. & Humphrey, J. A. C. in *Proc. 5th Int. Symp. Turbulence and Shear Flow Phenomena* 1023–1028 (Technical Univ. Munich, 2007).
35. McConney, M. E. *et al.* Surface force spectroscopic point load measurements and viscoelastic modelling of the micromechanical properties of air flow sensitive hairs of a spider (*Cupiennius salei*). *J. R. Soc. Interface* **6,** 681–694 (2009).
36. Gosline, J. *et al.* Elastic proteins: biological roles and mechanical properties. *Phil. Trans. R. Soc. Lond. B* **357,** 121–132 (2002).
37. Friedel, T. & Barth, F. G. Wind-sensitive interneurones in the spider CNS (*Cupiennius salei*): directional information processing of sensory inputs from trichobothria on the walking legs. *J. Comp. Physiol. A* **180,** 223–233 (1997).
38. Johnson, E. A. C., Bonser, R. H. C. & Jeronimidis, G. Recent advances in biomimetic sensing technologies. *Phil. Trans. R. Soc. A* **367,** 1559–1569 (2009).
39. McConney, M. E., Anderson, K. D., Brott, L. L., Naik, R. R. & Tsukruk, V. V. Bioinspired material approaches to sensing. *Adv. Funct. Mater.* **19,** 2527–2544 (2009).
40. Beckwith, T. G., Marangoni, R. D. & Lienhard, J. H. *Mechanical Measurements* (Addison-Wesley, 1993).
41. Bleckmann, H. in *Sensory Systems Neuroscience* (eds Hara, T. & Zielinski, B.) 411–444 (Academic, 2006).
42. Dijkstra, M. *et al.* Artificial sensory hairs based on the flow sensitive receptor hairs of crickets. *J. Micromech. Microeng.* **15,** S132–S138 (2005).
43. Krijnen, G. J. M. *et al.* MEMS based hair flow-sensors as model systems for acoustic perception studies. *Nanotechnology* **17,** S84–S89 (2006).
44. Fan, Z. F. *et al.* Design and fabrication of artificial lateral line flow sensors. *J. Micromech. Microeng.* **12,** 655–661 (2002).
45. Barbier, C., Humphrey, J. A. C. & Paulus, J. in *2007 Proc. ASME Int. Mech. Eng. Congress and Exposition* 1–6 (ASME, 2007).
46. Chen, N. *et al.* Design and characterization of artificial haircell sensor for flow sensing with ultrahigh velocity and angular sensitivity. *J. Microelectromech. Syst.* **16,** 999–1014 (2007).
47. den Toonder, J. *et al.* Artificial cilia for active micro-fluidic mixing. *Lab Chip* **8,** 533–541 (2008).
48. Evans, B. A. *et al.* Magnetically actuated nanorod arrays as biomimetic cilia. *Nano Lett.* **7,** 1428–1434 (2007).
49. Suh, J. W. *et al.* CMOS integrated ciliary actuator array as a general-purpose micromanipulation tool for small objects. *J. Microelectromech. Syst.* **8,** 483–496 (1999).
50. Reyssat, E. & Mahadevan, L. Hygromorphs: from pine cones to biomimetic bilayers. *J. R. Soc. Interface* **6,** 951–957 (2009).
    This paper explores possible ways of generating hygromorphic actuators based on pine cone movement.
51. Haupt, W. *Bewegungsphysiologie der Pflanzen* (Thieme, 1977).
52. Burgert, I. & Fratzl, P. Actuation systems in plants as prototypes for bio-inspired devices. *Phil. Trans. R. Soc. A* **367,** 1541–1557 (2009).
53. Wheeler, T. D. & Stroock, A. D. The transpiration of water at negative pressures in a synthetic tree. *Nature* **455,** 208–212 (2008).
54. Scholander, P. F., Hammel, H. T., Bradstreet, E. D. & Hemmingsen, E. A. Sap pressure in vascular plants: negative hydrostatic pressure can be measured in plants. *Science* **148,** 339–346 (1965).
55. Skotheim, J. M. & Mahadevan, L. Physical limits and design principles for plant and fungal movements. *Science* **308,** 1308–1310 (2005).
56. Gülch, R. W. Force–velocity relations in human skeletal muscle. *Int. J. Sports Med.* **15** (suppl. 1), 2–10 (1994).
57. Hill, A. V. The mechanics of active muscle. *Proc. R. Soc. Lond. B* **141,** 104–117 (1953).
58. Pennycuick, C. J. *Newton Rules Biology: A Physical Approach to Biological Problems* 30–39 (Oxford Univ. Press, 1992).
59. Dawson, C., Vincent, J. F. V. & Rocca, A. M. How pine cones open. *Nature* **390,** 668 (1997).
60. Elbaum, R., Zaltzman, L., Burgert, I. & Fratzl, P. The role of wheat awns in the seed dispersal unit. *Science* **316,** 884–886 (2007).
61. Elbaum, R., Gorb, S. & Fratzl, P. Structures in the cell wall that enable hygroscopic movement of wheat awns. *J. Struct. Biol.* **164,** 101–107 (2008).
62. Kulić, I. M., Mani, M., Mohrbach, H., Thaokar, R. & Mahadevan, L. Botanical ratchets. *Proc. R. Soc. B* **276,** 2243–2247 (2009).
63. Färber, J., Lichtenegger, H. C., Reiterer, A., Stanzl-Tschegg, S. & Fratzl, P. Cellulose microfibril angles in a spruce branch and mechanical implications. *J. Mater. Sci.* **36,** 5087–5092 (2001).
64. Fratzl, P., Elbaum, R. & Burgert, I. Cellulose fibrils direct plant organ movements. *Faraday Discuss.* **139,** 275–282 (2008).
    This paper gives a theoretical description of the biomimetic concept for an actuation system based on fibre-reinforced hydrogel systems.
65. Burgert, I., Eder, M., Gierlinger, N. & Fratzl, P. Tensile and compressive stresses in tracheids are induced by swelling based on geometrical constraints of the wood cell. *Planta* **226,** 981–987 (2007).
66. Goswami, L. *et al.* Stress generation in the tension wood of poplar is based on the lateral swelling power of the G-layer. *Plant J.* **56,** 531–538 (2008).
67. Pokroy, B., Kang, S. H., Mahadevan, L. & Aizenberg, J. Self-organization of a mesoscale bristle into ordered, hierarchical helical assemblies. *Science* **323,** 237–240 (2009).
68. Forterre, Y., Skotheim, J. M., Dumais, J. & Mahadevan, L. How the Venus flytrap snaps. *Nature* **433,** 421–425 (2005).
69. Seidel, R. *et al.* Mapping fibre orientation in complex-shaped biological systems with micrometre resolution by scanning X-ray microdiffraction. *Micron* **39,** 198–205 (2008).
70. Schwille, P. & Diez, S. Synthetic biology of minimal systems. *Crit. Rev. Biochem. Mol. Biol.* **44,** 223–242 (2009).
71. Li, D. B. *et al.* Molecular, supramolecular, and macromolecular motors and artificial muscles. *Mater. Res. Soc. Bull.* **34,** 671–681 (2009).
72. Whitesides, G. M. & Lipomi, D. J. Soft nanotechnology: 'structure' vs. 'function'. *Faraday Discuss.* **143,** 373–384 (2009).
73. Dong, L. X., Subramanian, A. & Nelson, B. J. Carbon nanotubes for nanorobotics. *Nano Today* **2,** 12–21 (2007).

Acknowledgements The work of F.G.B. was generously supported by the Austrian Science Fund and the BioSenSE programme of the US Defense Advanced Research Projects Agency. The productive collaboration with engineers, in particular J. A. C. Humphrey, F. G. Rammerstorfer and V. V. Tsukruk and their research groups, is gratefully acknowledged. P.F. was supported by the Alexander von Humboldt Foundation in the framework of the Max Planck Research Prize. He is grateful for discussions with, and advice from, co-workers and colleagues, in particular I. Burgert, R. Elbaum, O. Paris, J. Dunlop and L. Mahadevan.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to P.F. (fratzl@mpikg.mpg.de) or F.G.B. (friedrich.g.barth@univie.ac.at).

© 2009 Macmillan Publishers Limited. All rights reserved

Materials engineering for immunomodulation

Jeffrey A. Hubbell[1,2], Susan N. Thomas[1] & Melody A. Swartz[1,2]

The engineering of materials that can modulate the immune system is an emerging field that is developing alongside immunology. For therapeutic ends such as vaccine development, materials are now being engineered to deliver antigens through specific intracellular pathways, allowing better control of the way in which antigens are presented to one of the key types of immune cell, T cells. Materials are also being designed as adjuvants, to mimic specific 'danger' signals in order to manipulate the resultant cytokine environment, which influences how antigens are interpreted by T cells. In addition to offering the potential for medical advances, immunomodulatory materials can form well-defined model systems, helping to provide new insight into basic immunobiology.

The term 'immunobioengineering' is used to describe efforts by immunologists and engineers to design materials, delivery vehicles and molecules both to manipulate and to better understand the immune system. Examples are the engineering of material surfaces to induce or prevent complement activation, the engineering of adjuvants to activate the immune system, the engineering of antigen or adjuvant carriers for subunit vaccine delivery, and the engineering of microenvironments to determine the interaction kinetics of mature dendritic cells and naive T cells. These advances not only will contribute to prophylactic vaccine strategies for infectious diseases but also are likely to affect immunotherapeutics, particularly for cancer, and new approaches to prevent or treat allergies and autoimmune diseases. The field is rapidly evolving along with advances in our understanding of immunology and is also contributing to our knowledge of basic immunology.

In this Review, we describe the current state of immunobioengineering as it intersects with the field of materials science, and we give a perspective on its current and future directions. We focus on materials for immunomodulation, particularly with respect to dendritic-cell modulation. We begin by providing a brief introduction to the targets for delivery: the types of cell that materials are being designed to target, the tissues in which those cells reside, the intracellular compartments within those cells, and the influence that delivery to those particular compartments has on immunological outcome. We then discuss the biomolecular payloads used to activate immune cells and the design of the materials used to deliver those 'danger' signals along with, in the context of vaccination, antigens. Finally, we highlight materials approaches that are being developed to explore basic immunological function, especially how dendritic cells and T cells interact.

Tissue and cellular targets

As the general goals of immunobioengineering are to probe and manipulate the immune system, we start with a general discussion of tissue, cellular and subcellular targets — what we want to target and why — to guide the design principles discussed below.

The immune cells most frequently targeted include B cells, macrophages and dendritic cells, which are all effective antigen-presenting cells (APCs)[1]. Dendritic cells are typically considered the most specialized, because they present antigen to their cognate naive T-cell partners and instruct them what do with it (for example induce anergy, tolerance or immunity)[2]. They are the main focus of this Review, although recent evidence points to basophils being an important APC involved in T helper 2 (T_H2)-cell immunity[3].

In the most simplistic, classic view, dendritic cells remain in an immature state while sampling antigens in their environment to present to T cells for the maintenance of self-tolerance (that is, they present antigen without costimulatory molecules); this includes in the lymph node, where functionally immature dendritic cells sample antigens drained with lymph from the periphery[4]. When they encounter pathogenic or endogenous danger signals (described below) or adjuvants (as engineered danger signals) while taking up antigen, they begin to mature and express the chemokine receptor CCR7, which allows them to migrate into the nearest draining lymphatic vessels and then to the lymph node[5]. There, they present processed antigenic peptides with maturation-induced costimulatory molecules to T cells to initiate an adaptive, or antigen-specific, immune response.

Thus, in this simplistic view, antigen presented by immature dendritic cells (in the presence of transforming growth factor-β1 (TGF-β1) and interleukin-10 (IL-10)) maintains tolerogenic responses (Fig. 1a), whereas that presented by mature dendritic cells in the presence of immunogenic cytokines can lead to immunogenic T cells. In reality, there are exceptions: mature dendritic cells can induce tolerogenic responses, and there are 'partially mature' dendritic cells, whose function is poorly understood[1].

Moreover, there are many subsets of dendritic cells, including the following: plasmacytoid dendritic cells, which secrete interferon-α (IFN-α); myeloid dendritic cells, which can secrete large amounts of IL-12; follicular dendritic cells, which do not express major histocompatibility complex (MHC) class II molecules; lymphoid dendritic cells, which can secrete large amounts of IFN-γ; $CD8^-$ or $CD8^+$ dendritic cells, of which the former do not carry the CD8 antigen and the latter do; and several tissue-specific subtypes such as Langerhans cells. Although still poorly understood, these subtypes can be specialized for different functions, such as antigen 'cross-presentation' by $CD8^+DEC205^+$ dendritic cells (DEC205 also known as LY75) residing in the splenic T-cell zone, or MHC class II presentation by $CD8^-$ dendritic cells residing in the red pulp[6]. Much of what we know about dendritic-cell behaviour comes from *in vitro* studies using dendritic cells derived from peripheral blood monocytes, haematopoietic progenitor cells or bone marrow; these cells are typically differentiated into immature dendritic cells by using IL-4

[1]Institute of Bioengineering, [2]Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

© 2009 Macmillan Publishers Limited. All rights reserved

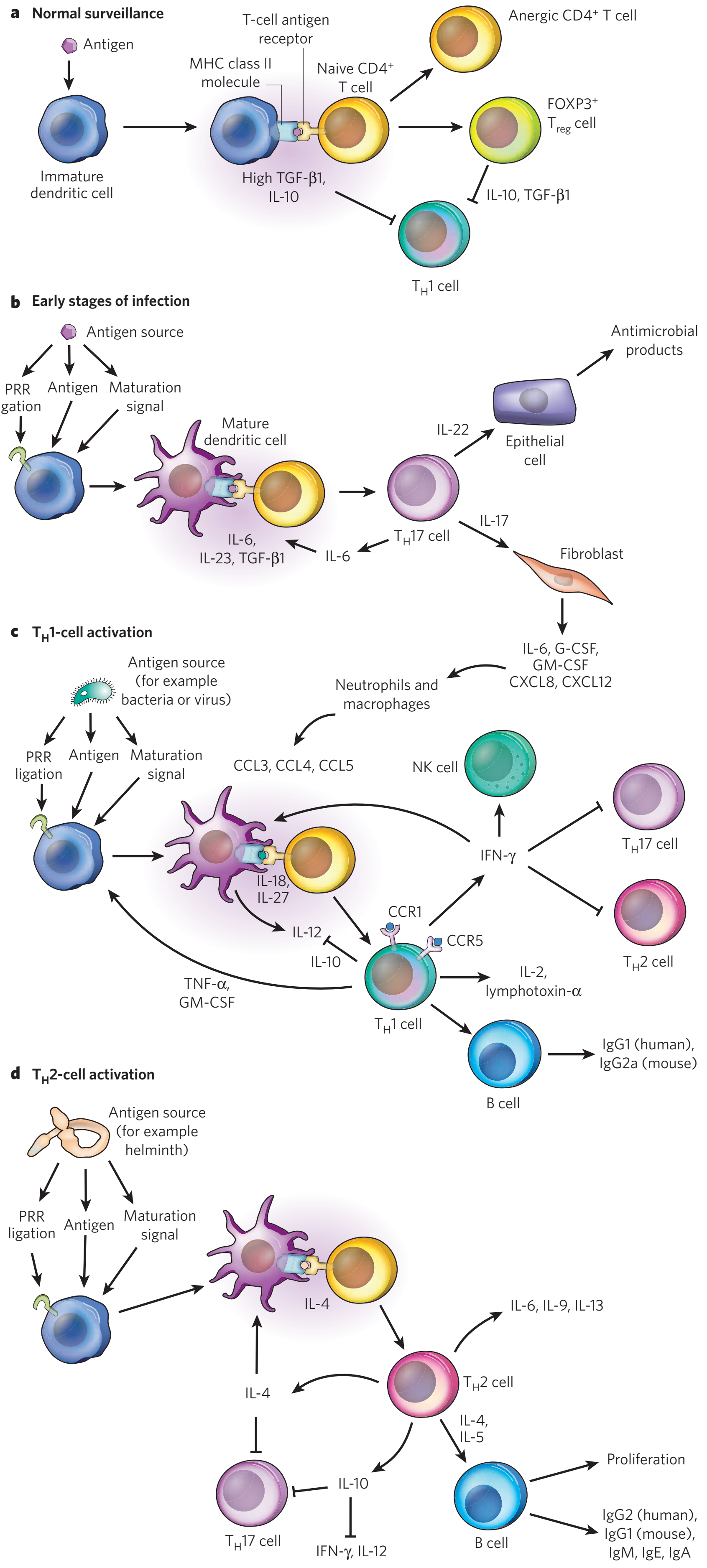


Figure 1 | Design principles for activating antigen-specific CD4$^+$ T cells. To activate naive CD4$^+$ (helper) T cells, the T-cell antigen receptor needs to recognize antigen that has been loaded onto MHC class II molecules and presented by dendritic cells. The response of the T cells depends not only on this receptor–ligand interaction but also on costimulatory molecules being presented by the dendritic cells and on the cytokine environment during activation. **a**, In normal surveillance mode, dendritic cells in their immature state constantly pick up antigen and present it, without costimulatory molecules, to T cells. This leads to T-cell anergy (that is, the T cell cannot receive further signals) and/or activation of inducible (FOXP3$^+$) regulatory T (T$_{reg}$) cells when the cytokines TGF-β1 and IL-10 are present. T$_{reg}$ cells themselves secrete TGF-β1 and IL-10, which inhibit T_H1 cells. In fact, all types of T cell can secrete IL-10. **b**, By contrast, when dendritic cells that have been activated by pattern-recognition receptor (PRR) ligation and maturation signals from microbial products present antigen to T cells together with costimulatory molecules, they can drive a T_H1-, T_H2- or T_H17-cell response. For example, at the early stages of bacterial or viral infection, antigen presentation occurs in a microenvironment that contains IL-6 and IL-23 (but no IL-4), stimulating naive CD4$^+$ T cells to differentiate into T_H17 cells. These cells secrete IL-6, IL-22 (which induces epithelial cells to produce antimicrobial peptides) and IL-17 (which activates local fibroblasts). The fibroblasts, in turn, attract neutrophils and macrophages through the secretion of cytokines such as GM-CSF and IL-6 and chemokines such as CXCL8 and CXCL12. **c**, At later stages of infection, these activated neutrophils and macrophages, in turn, secrete the chemokines CCL3, CCL4 and CCL5, which attract additional T cells and promote the differentiation of CD4$^+$ T cells into T_H1 cells. A T_H1-cell response occurs when dendritic cells mature in the presence of cytokines such as TNF-α and GM-CSF and when PRRs such as Toll-like receptors (TLRs) are activated by microbial products, leading to IL-12 secretion. IL-12 is a key cytokine for T_H1-cell activation, which can be enhanced by IFN-γ. T_H1 cells secrete IFN-γ, which (in addition to promoting further T_H1-cell activation) activates natural killer (NK) cells and inhibits the activation of T_H2 and T_H17 cells. T_H1 cells also produce TNF-α, as well as IL-2 and lymphotoxin-α, all of which can drive B cells to differentiate into opsonizing-antibody-producing plasma cells (which predominantly produce IgG). Under certain conditions, IL-10 is secreted by T_H1 cells as an inhibitory feedback regulator. **d**, T_H2-cell responses are elicited when dendritic cells present antigen in the presence of IL-4, which is secreted by T_H2 cells themselves and inhibits T_H17 cells. T_H2 cells also secrete IL-5, which together with IL-4 stimulates B-cell proliferation and antibody production (especially antibody of the classes IgM, IgA and IgE). Other cytokines secreted by T_H2 cells include IL-6, IL-9, IL-13 and IL-10, the last of which inhibits IFN-γ production by T_H1 cells and IL-12 production by dendritic cells. Therefore, when engineering immune responses, it is important to consider the cytokine microenvironment, in addition to how the antigen will be presented by dendritic cells, both of which can be modulated by PRR signalling and uptake mechanisms.

© 2009 Macmillan Publishers Limited. All rights reserved

Table 1 | Influence of the cytokine microenvironment on immune responses

Desired response	Cytokine environment of dendritic-cell activation	Cytokines produced by activated dendritic cells	Cytokines produced by activated T cells	Effect on other cells	Natural inducers
T_{reg} cell	High TGF-β1 and IL-10; low IL-6 and IL-12	TGF-β1	IL-10	Suppresses $CD4^+$ and $CD8^+$ T-cell proliferation	Self antigens
T_H17 cell	IL-23, high TGF-β1 and high IL-6	TNF-α	IL-17, IL-22 and IL-6	Activates fibroblasts to produce IL-6, G-CSF, GM-CSF, CXCL8 and CXCL12, which attract neutrophils and macrophages, creating a T_H1-cytokine microenvironment	Bacterial and viral infections
T_H1 cell	IL-12, IL-18 and IL-27	IL-12	IFN-γ, IL-2 and lymphotoxin-α	Blocks T_H17- and T_H2-cell development; activates B cells to produce IgG1 (in humans)	Bacterial and viral infections
T_H2 cell	IL-4 and IL-6	IL-1	IL-4, IL-5, IL-6, IL-9, IL-10 and IL-13	Induces B-cell proliferation and antibody class switching; promotes T_{reg}-cell development; activates macrophages (through the alternative pathway)	Helminth infections

and granulocyte–macrophage colony-stimulating factor (GM-CSF, also known as CSF2) and matured by using lipopolysaccharide (LPS) or tumour-necrosis factor-α (TNF-α). Therefore, care must be taken in translating *in vitro* data from generic 'dendritic cells' to the *in vivo* situation, with an appreciation for tissue-specific dendritic-cell subsets, and it should be noted that there are many differences between such subsets in rodents and those in humans.

The response that a dendritic cell elicits depends on many factors, including the state of maturation of the cell, how the antigen was taken up and processed by the cell, and even the tissue in which the cell was activated. Antigens presented in the context of MHC class I molecules are recognized only by $CD8^+$ T cells, whereas those bound to MHC class II molecules are recognized by $CD4^+$ T cells. Dendritic cells, classically $CD8^+$ dendritic cells, can present antigen in the context of both classes of MHC molecule. Some of these complex interactions with $CD4^+$ T cells are illustrated in Fig. 1 and described in Table 1. Importantly, the cytokine environment in which both dendritic-cell activation and communication between dendritic cells and $CD4^+$ T cells occurs can control the response and should be considered when choosing tissue targets. These cytokines can be secreted by dendritic cells on activation or inactivation, by the activated $CD4^+$ T cells themselves, by neutrophils and macrophages recruited to the inflammatory site and, finally, by stromal cells that can become activated during inflammation. Because the cytokine profile present during contact with the dendritic cell can determine $CD4^+$ T-cell fate, engineering approaches that manipulate or make use of the cytokine environment are important design considerations. For example, a biomaterials vaccine platform in which IL-10 expression was knocked down using short interfering RNA has been explored, showing strongly positive effects; from Fig. 1 and Table 1, it is evident that this was done to block the inhibitory role that IL-10 has in T_H1-cell and T_H2-cell development[7].

APCs traffic through almost every tissue in the body. Generally speaking, epithelial tissues have strong immunosurveillance activity, as they form barriers to the outside world through which most pathogenic entry occurs. In peripheral tissues such as the skin, Langerhans cells, other dendritic cells and macrophages detect danger signal and antigens, signal other immune cells to the site by means of chemokine secretion and migrate to the nearest draining lymph node to activate an immune response. Most traditional vaccines delivered in the skin or muscle target such cells, including those vaccines formulated in alum (the standard adjuvant, which consists of particulate aluminium phosphates; described further below), an adjuvant that targets peripheral APCs through its 'depot effect'. The skin, lungs, gut and lymph nodes are common target tissues for both natural immunomodulatory agents and prophylactic or therapeutic ones, and different immune responses can be achieved in different target tissues[8].

The lymph node is an emerging target tissue of interest. Dendritic cells and B cells reside in the lymph nodes, and T cells traffic through the paracortical region, the architecture of which is optimized for rapid cell trafficking to help naive T cells to make contact with thousands of dendritic cells to find their APC partners[9]. Immature dendritic cells patrol peripheral tissues and then migrate to the lymph nodes after they take up antigen, but the lymph nodes also contain many immature dendritic cells that constantly sample antigens carried through the lymph nodes by lymph drained from the peripheral tissues; the latter population of dendritic cells may function mainly for the purpose of maintaining tolerance[4]. However, when these immature lymph-node dendritic cells were exposed to antigen together with a maturation stimulus, they were able to activate T cells, suggesting that lymph-node dendritic cells could be potential targets for immunomodulatory agents[4] (approaches for targeting the lymph node are described below). Also, it has been shown that dendritic-cell presentation of peptide-antigen-loaded MHC class II molecules in the draining lymph node of a tissue following subcutaneous antigen delivery came in two discrete stages: first, by the lymph-node dendritic cells, which acquired antigen in the lymph node; and, second, by the dendritic cells that had migrated there from the injection site[10].

Further in support of the lymph node as a target, the humoral response is apparently initiated by lymph-node B cells, which take up antigen directly in the lymph-node follicles, rather than by migrating B cells or by dendritic cells, indicating that lymph-node targeting may be advantageous for protective antibody-generating vaccines[11]. This approach has not been widely investigated, and it remains to be determined how the T-cell response differs when initiated by peripherally activated dendritic cells and when initiated by lymph-node dendritic cells, as well as how the chemokine balance shifts in the lymph node and how this ultimately affects the long-term response. This is particularly important in light of emerging evidence demonstrating the importance of lymph-node T-cell homing in tolerogenesis[12,13]. Thus, although lymph-node targeting for immunomodulation has interesting prospects, much more research is needed.

The mucosal epithelia are a major site of immunosurveillance, responding to most immune challenges encountered by an organism. The mucosae of the airways, the digestive tract and the vagina, among others, contain special lymphoid tissues, referred to as the mucosa-associated lymphoid tissues. In these tissues, antigens are collected by microfold (M) cells and transferred to dendritic cells, which in collaboration with T cells induce a specialized humoral response of secretory IgA in the mucosal secretions, as well as a cellular response[14]. Much of the protective response of the mucosa derives from secretory IgA. Systemic vaccination, for example by intramuscular injection, does not typically induce strong mucosal immunity, whereas vaccination of the mucosa does, although in a region-specific manner[14]. Given the importance of inducing mucosal immunity and protection against the vast number of pathogens that enter through these routes, the different mucosae are important tissue targets, and the M cells and dendritic cells in the mucosa-associated lymphoid tissues are important cellular targets[14].

In addition to understanding the tissue and cellular targets for immunotherapeutics, it is important to understand the subcellular targets. APCs process antigen from the cytosol or following endocytosis for display in MHC class I molecules and/or MHC class II molecules, respectively[15]. Intracellular proteins (for example those produced by viral pathogens) are degraded by the proteasome, released into the cytosol and subsequently translocated into the endoplasmic reticulum by the transporter associated with antigen processing (TAP) to be loaded into the peptide-binding site of MHC class I molecules[1]. By contrast, proteins internalized from the

© 2009 Macmillan Publishers Limited. All rights reserved

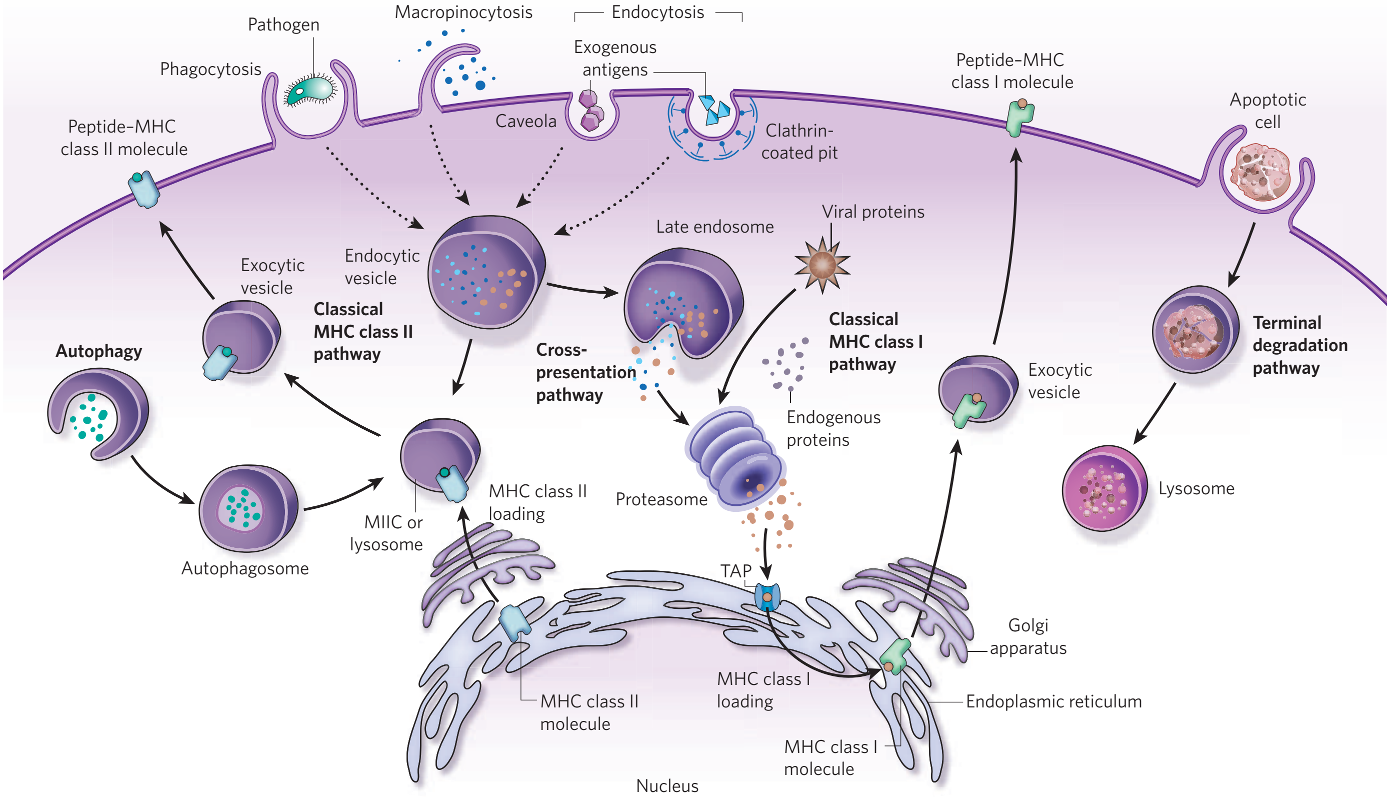


Figure 2 | Design principles for targeting the intracellular pathways of dendritic cells to modulate antigen presentation. A simplified view of antigen presentation by dendritic cells. Left, exogenous particles, proteins or pathogens can be taken into the cell through various pathways, including phagocytosis (for particles >1 μm), macropinocytosis (<1 μm), and endocytosis from caveolae (~60 nm) or clathrin-coated pits (~120 nm). Exogenous antigens are then processed in endocytic vesicles (phagosomes, endosomes, lysosomes and/or endolysosomes; dashed arrows represent multiple vesicular steps). Processed antigen (peptide) is subsequently loaded onto MHC class II molecules (which have been assembled in the endoplasmic reticulum, transported through the Golgi apparatus and targeted to endocytic compartments) in a lysosome or MHC class II compartment (MIIC). The peptide–MHC class II complexes then move through exocytic vesicles to the cell surface, where antigen presentation occurs. MHC class II loading of endogenous antigen provided by autophagy can also occur, particularly when the cell is under stress. Right, antigen can be loaded onto MHC class I molecules through two main pathways. In the classical pathway, endogenous or viral proteins in the cytosol are processed through the proteasome, transported into the endoplasmic reticulum through the molecule TAP (transporter associated with antigen processing), loaded onto MHC class I molecules, and then transported through the Golgi apparatus and exocytic vesicles to the cell surface for presentation. In addition, exogenous antigens that have been phagocytosed, macropinocytosed or endocytosed can be cross-presented on MHC class I molecules by some subsets of dendritic cell. In this pathway, antigen either may be loaded in endocytic compartments (not shown) or may escape endosomes and arrive in the cytosol, where it is processed through the proteasome as usual, loaded onto MHC class I molecules and transported to the surface. Finally, terminal degradation pathways can occur (for example when apoptotic cells are internalized). See refs 1 and 96–98 for details about antigen processing.

extracellular environment (for example from phagocytosed bacteria) are digested in endolysosomes, and peptides derived from these are loaded onto MHC class II molecules, which are then transported to the plasma membrane[1]. Some subsets of dendritic cells, classically CD8$^+$ dendritic cells, can also efficiently present exogenously obtained peptides on MHC class I molecules, a process known as cross-presentation, making dendritic cells unique among the APCs (macrophages can also cross-present antigen but with much lower efficiency[1]). As mentioned above, CD8$^+$ T cells are primed by interactions between the T-cell antigen receptors (TCRs) of CD8$^+$ T cells and MHC class I molecules, and CD4$^+$ T cells are primed by the binding of their TCRs to MHC class II molecules. Therefore, the choice of the route of antigen delivery, and thus which class of MHC molecule presents the antigen, directly (but not solely, as the nature of dendritic-cell activation by danger signals is also important) controls the type of antigen-specific T-cell response that is obtained. These presentation pathways, in their simplest and most classic form, are illustrated in Fig. 2.

Materials as tools to modulate immune-cell function

As mentioned, APCs — in particular dendritic cells — are responsible for integrating a myriad of external biomolecular stimuli to produce an adaptive immune response. Antigen presentation leads to immunogenicity only when costimulatory molecules are presented with the MHC-class-I-associated or MHC-class-II-associated peptide antigen. Thus, to interpret what to do with the antigen, dendritic cells use receptor systems, such as pattern-recognition receptors (PRRs), that can sense either pathogen-derived or endogenous danger signals[16,17]. A pathogen may be recognized by several PRRs either simultaneously or sequentially, activating distinct or shared signalling pathways. How the dendritic cell responds, and therefore the quality of the induced adaptive immunity, is determined by the danger signals to which the dendritic cell is exposed[17].

PRRs largely recognize pathogen-derived biomolecules referred to as pathogen-associated molecular patterns (PAMPs), which are evolutionarily distant non-self molecules such as LPS and viral double-stranded RNA[18]. Many endogenous molecules can also trigger PRR activation; such molecules, known as danger-associated molecular patterns (DAMPs), are typically associated with tissue damage or distress. A select group of ligands for a few representative PRRs, along with the immunological responses they induce, are described in Table 2.

Recognition of PAMPs and DAMPs by PRRs occurs both through extracellular activation cascades such as the complement system and through intracellular signalling pathways that can be initiated at the dendritic-cell surface, in the endosome after phagocytosis or in the cytosol[16]. PRR expression patterns vary significantly between subcellular

© 2009 Macmillan Publishers Limited. All rights reserved

Table 2 | Examples of PRRs, their PAMP ligands and associated immune responses

PRR	Ligand	Associated immune response	Synthetic analogue	Biomaterials engineering approaches	Induced immune response
TLRs					
TLR3	Double-stranded RNA	Induces type I IFNs and pro-inflammatory cytokines (TNF-α, IL-6 and IL-12)[16,85]	Poly(I:C)[55]	pH-sensitive biodegradable polyketals co-encapsulating ion-paired protein antigen and poly(I:C) into ~1–3-μm-diameter particles[55]	Increased the number of IFN-γ-producing antigen-specific CD8$^+$ T cells; elicited TNF-α and IL-2 production by CD8$^+$ T cells *in vitro*[55]
				Biodegradable poly(D,L-lactide-co-glycolide) microspheres co-encapsulating protein antigen with poly(I:C)[79]	Increased the number of IFN-γ-producing antigen-specific CD8$^+$ T cells *in vivo*[79]
TLR4	LPS	Induces IL-6, IL-12 and TNF-α; upregulates costimulatory molecule and type I IFN production by dendritic cells[16,86]	Tetra-acyl lipid A[80]	Poly(lactic-co-glycolic acid) nanoparticles co-encapsulating tumour-associated protein antigen and tetra-acyl lipid A[80]	Increased the number of IFN-γ-producing antigen-specific CD8$^+$ T cells; induced production of pro-inflammatory cytokines (TNF-α, IL-12, IFN-γ, IL-2 and IL-6) at the tumour site[80]
TLR5	Flagellin	Induces strong IgM and IgG responses[87]	Recombinant flagellin domains[88]	Poly(methyl vinyl ether-co-maleic anhydride) nanoparticles coated with flagella-enriched extract[89]	Not determined
TLR7	Single-stranded RNA	Induces IFN-α, IL-12 and TNF-α; recruits dendritic cells and cytotoxic T cells; increases T-cell activation by APCs[90]	Imiquimod[91]	Imiquimod administered immediately after delivery of plasmid DNA coated onto 2-μm-diameter gold particles[91]	Increased the number of mature dendritic cells in draining lymph nodes; enhanced antigen-specific CD4$^+$ and CD8$^+$ T-cell responses, biased towards a predominance of T$_H$1 cells[91]
TLR9	Unmethylated bacterial DNA	Activates immune cells and cytokine production for strong T$_H$1-type responses[92]	CpG oligonucleotides[79,92]	Biodegradable poly(D,L-lactide-co-glycolide) microspheres co-encapsulating protein antigen with CpG oligonucleotides[79]	Increased the number of IFN-γ-producing antigen-specific CD8$^+$ T cells and the level of cytolysis; improved protection against vaccinia virus infection compared with separately administered antigen and adjuvant[79]
NALPs					
NALP3	Particulate matter such as asbestos, silica or alum	Induces pro-inflammatory cytokine (IL-1β)[93]	Polymeric microparticles[64]	Biodegradable poly(D,L-lactide-co-glycolide) and polystyrene microparticles administered in conjunction with protein antigen[64]	Increased the secretion of IL-1β by dendritic cells; induced higher antibody titres; elicited IL-6 production by T cells; recruited and activated CD11b$^+$Gr1$^-$ cells *in vivo*[64]
Complement-associated					
C3	Carbohydrates and bacterial proteins	Induces pathogen clearance by opsonization of pathogens[94]	Polymeric materials containing nucleophiles[28]	Complement-activating, nucleophile-containing polymeric nanoparticles[28]	Induced antigen-specific monoclonal antibody; induced IFN-γ-producing antigen-specific CD8$^+$ T cells[28]
			Recombinant C3d[72,95]	Recombinant, trimeric C3d protein conjugated to protein antigen through an avidin bridge[18]	Induced a more robust and protective response to antigen when administered with IFA than when protein antigen administered in IFA or with alum[18]
				Recombinant, trimeric C3d–antigen fusion DNA vaccine[95]	Generated higher-binding, early-appearing and neutralizing antibody responses; increased the number of IFN-γ-producing antigen-specific CD8$^+$ T cells[95]

The responses shown are related to engineering approaches to triggering these pathways or types of immunity using biomaterials. The table is not comprehensive but highlights a few recent biomaterials-based strategies for immunotherapeutic applications. C3, complement component 3; IFA, incomplete Freund's adjuvant; NALP3, NACHT domain-, leucine-rich repeat- and PYD-containing protein 3; poly(I:C), polyinosinic acid/polycytidylic acid.

compartments, immune-cell types and subtypes, and tissues[16,19]. One of the major PRR classes is the Toll-like receptor (TLR) family, members of which recognize a large number of pathogen-derived ligands and a smaller number of endogenously derived ligands. Of the TLRs, some — such as TLR4 and TLR2, which recognize LPS and lipoteichoic acid respectively — may be expressed on the plasma membrane. By contrast, others — such as TLR7, TLR8 and TLR9, which recognize bacterial RNA or DNA[16] — may be present in the endosomal compartment. Many other intracellular and membrane-expressed PRRs are involved in the recognition of viral nucleic acids or bacterial and fungal carbohydrates by dendritic cells; these include cytosolic NOD-like receptors (such as NALP3), which activate the dendritic-cell inflammasome in response to bacterial and endogenous danger signals[16]. Again, the dendritic cell integrates these signals to determine whether to mature, how to process and present the antigen, and which cytokines to produce.

For these reasons, one important task in immunobioengineering is to develop delivery strategies by which to present antigen — along with PAMPs to ligate particular PRRs — so as to induce a desirable T$_H$1-type or T$_H$2-type adaptive immune response. We contemplate the targeting and penetration of barriers as objectives for materials design: the barrier of the antigen being able to find, or being found by, the APCs that reside in the tissues performing surveillance for signals of infection, the barrier of the tissue interstitium after injection of antigen into connective tissue such as skin, the barrier of the mucosa after antigen is sprayed into the nasal sinus or inhaled into the lungs, the barrier of entry to the endolysosomal compartment of the cell after endocytosis, and the barrier of entry to the cytosol presented by the endosomal membrane. Materials with different design principles and characteristics are being developed to accomplish these delivery tasks.

Materials for enhancing antigen uptake by APCs

Materials design considerations for enhancing uptake by APCs have focused on recognition and recruitment. With regard to recognition, some subclasses of dendritic cell possess an endocytic receptor, DEC205 (ref. 20), which has been successfully used to enhance dendritic-cell uptake, for example with an antigen or a biomaterial particle conjugated to anti-DEC205 antibodies[21,22]. Recruitment involves chemoattracting other APCs to the delivery site, and strategies include the use, variously,

© 2009 Macmillan Publishers Limited. All rights reserved

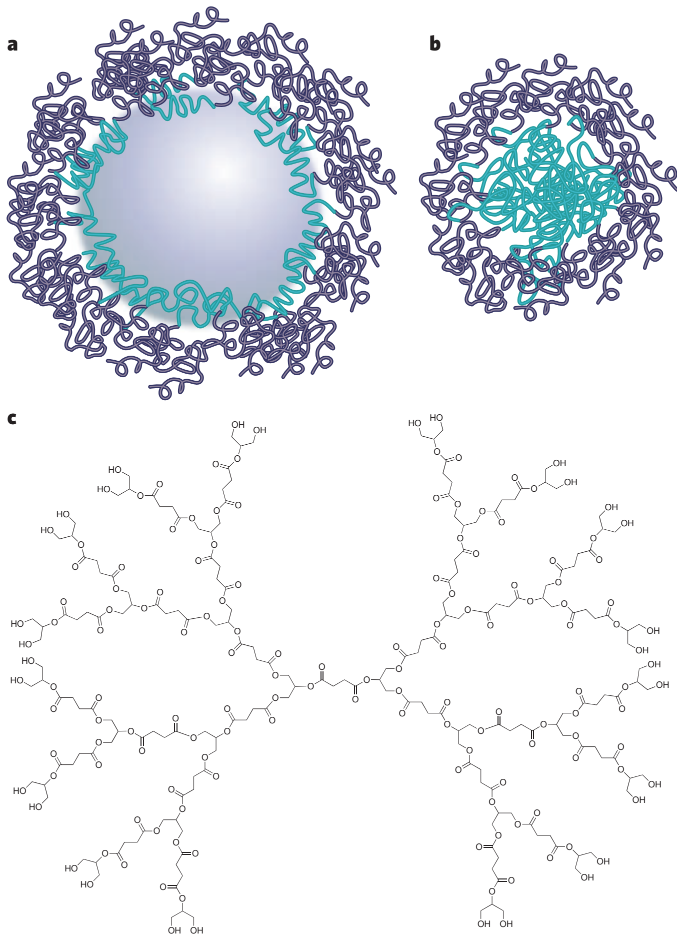


Figure 3 | Means of forming polymeric nanoparticles. **a**, Emulsion polymerization is carried out in a continuous phase, with an emulsifier, here shown as an ABA block co-polymer (A in purple, and B in green), surrounding monomer droplets (centre). The surfactant PEG-*bl*-polypropylene glycol-*bl*-PEG is a convenient emulsifier, in that the terminal hydroxyls on the polymer may be used for antigen grafting (not shown), for example as in ref. 28. **b**, Self-assembly of amphiphilic block co-polymers can yield very small nanoparticles (as depicted). Diameters of less than 15 nm have been achieved[33] using PEG-*bl*-polypropylene sulphide (PEG-*bl*-PPS). **c**, Even smaller structures for antigen display can be formed by synthesizing branched polymers, such as the dendrimer depicted. For further examples of dendrimers, see ref. 37. In all panels, antigen and danger signals may be attached to the nanoparticle surfaces. Note that the structure in **c** is much smaller than those in **a** and **b**.

of degradable polyester particles to create gradients of dendritic-cell chemoattractants[23], injectable hydrogels[24], and degradable scaffolds that simultaneously release cytokines, PAMPs and antigen[25]. As an extreme example, cell transplantation approaches involve isolating dendritic cells from a subject, exposing them to antigen *in vitro* (referred to as 'loading'), stimulating them and then re-implanting them[26].

Materials for penetrating tissue barriers

Whereas the section above describes approaches to delivering the dendritic cell to a material by targeting, recruitment or transplantation, approaches are also being explored to take the material directly to the endogenous dendritic cells in the draining lymph node after injection into the tissue interstitium, with particle size being the primary material control parameter. In most tissues, there exists a slow interstitial flow from the blood capillaries to the lymphatic capillaries, of the order of 0.1–1 μm s^{-1} (ref. 27). It is through this interstitial flow that macromolecules are swept into the lymphatics, and that immature lymph-node dendritic cells (and follicular B cells, as described earlier[11]) can sample self molecules and pathogen-derived molecules present in the tissues. This approach to targeting APCs is highly robust, as long as the immunotherapeutic particles are neither too big (resulting in entrapment in the interstitium) nor too small (resulting in absorption into the blood). The dependence of targeting on size has been probed with biomaterial nanoparticles sterically stabilized with polyethylene glycol (PEG) brushes: particles of ~20–25 nm in diameter were very efficiently delivered to lymph nodes; particles of 45–50 nm in diameter were somewhat less efficiently delivered (64%, plus the 20–25-nm nanoparticles); and particles of 100 nm in diameter were poorly delivered (8%, plus the smaller particles)[28]. It is interesting to compare these sizes with those of viruses: the smallest viruses (for example the single-stranded DNA parvovirus or the RNA picornaviruses such as poliovirus) are in the 20–30-nm range, but most are substantially larger (for example adenoviruses, 70–80 nm; retroviruses, 100–200 nm; and poxvirus, 100 × 200 × 300 nm^3)[29]. As described below, it is possible to use materials chemistry to access sizes as low as, and even smaller than, those of biological viruses.

Three general material-fabrication schemes are useful for forming polymer nanoparticles in the 20–30-nm range: emulsion polymerization, self-assembly and branched-polymer synthesis (other techniques are available for inorganic nanoparticles). These schemes are illustrated in Fig. 3. In emulsion polymerization[30], surfactant micelles are formed in an aqueous environment, and hydrophobic monomer is loaded within the micelle and polymerized. The result is a surfactant-stabilized polymer nanoparticle, the size distribution of which can be controlled by the ratio of surfactant to monomer. The particles contain a hydrophobic core within a hydrophilic corona (from the surfactant), to the surface of which hydrophilic molecules such as antigens and PAMPs may be conjugated[31]. Useful surfactants include block co-polymers, such as PEG-*bl*-polypropylene glycol-*bl*-PEG, also known as Pluronics[28].

Self-assembly is a powerful method for forming very small particles. In principle, the size distribution of particles formed by self-assembly can be very narrow, owing to the potential to approach equilibrium. Typically, amphiphilic block co-polymers are dissolved in a water-miscible organic medium that is a solvent for both block compositions, and this solution is subsequently dropped into water, which is a solvent for one block but not the other, forcing micellization with the hydrophobic block at the core of the micelle. Polymer micelles are intrinsically unstable structures that can disassemble after the infinite dilution that follows injection into the body, driving considerations of how to engineer an optimal dissociation rate.

The hydrophobicity of the core-forming block is an important consideration; for example, the critical micelle concentration of block co-polymers containing polypropylene glycol (such as Pluronics) is not as favourable (low) as that of analogous block co-polymers containing polypropylene sulphide (PPS), in which the oxygen atoms in the polymer backbone of polypropylene glycol have been replaced with sulphur atoms[32]. Using this materials chemistry, it is possible to access the subviral size range; for example, PEG_{44}-*bl*-PPS_{10} forms spherical micelles that have 7-nm cores and 14-nm total diameters[33] and that demonstrate slow dissociation. Alternatively, very high-molecular-weight polypropylene glycol (the hydrophobic block) in Pluronics micelles can serve as a stabilizing influence[34].

In addition to hydrophobicity, and the molecular size of the hydrophobe in a self-assembling block co-polymer, the melting temperature (T_m) and the glass-transition temperature (T_g) are important. On the one hand, low-T_g polymers have the advantage of being readily formed at normal production temperatures, thus approaching equilibrium micelle size and shape; on the other hand, higher-T_g hydrophobic block compositions (T_g >37 °C) or crystalline hydrophobic polymers (T_m >37 °C) self-assemble at supraphysiological temperatures for very stable use at 37 °C, below T_g or T_m. Micelle, and also polymersome[35] (see below), processability can thus be engineered by manipulating any of the material parameters mentioned above to exploit the equilibrium nature of the materials, as well as by slowing the dissociation rate to suit practical use.

© 2009 Macmillan Publishers Limited. All rights reserved

Biologically derived molecules may also readily self-assemble, as can be observed in virus-like particles, which are now in clinical use. Virus-like particles are biotechnologically produced, self-assembled structures of viral capsid proteins and can approach the 50-nm diameter range[36]. Heterologous production of the viral capsid proteins ensures their self-assembly without a viral genome, and self-assembled protein particles therefore possess the great advantage of biological functionality and intrinsic immunoreactivity, without the potential for infectivity. As these materials are biologically derived, however, they are more expensive and can be stored less stably than synthetic materials. Nevertheless, they are clinically highly effective and serve as an excellent model to be mimicked by biomaterials scientists.

Branched polymers, especially dendrimers, have been developed as very small nanoparticles to display antigen and danger signals[37]. The advantages of branched-polymer strategies include the stability ensured by the covalent bonds within the polymer and a very narrow size distribution due to the nature of their synthesis. This synthesis is performed stepwise from a multifunctional initial scaffold, by adding difunctional monomer in serial couple–deprotect steps. Thus, the number of end groups on the dendrimer doubles with each serial step. As well-characterized display systems of well-defined and controllable size, these materials have great potential as vaccine platforms[38].

Other materials design considerations for polymer particles intended for eventual therapeutic use relate to materials stability and elimination. Although a degradation mechanism is necessary, ensuring stability during storage is paramount. It is very difficult to dry and then resuspend nanoparticles at their original size distribution; moreover, gentle drying processes such as lyophilization are expensive, which may hinder the use of a vaccine form in global applications. Therefore, storage in water, preferably without refrigeration, is an important design goal. Promising materials chemistries have been explored to engineer nanoparticles that degrade either by oxidation or reduction (requiring that the nanoparticles be stored in a controlled atmosphere, which is inexpensive) or by pH-triggered hydrolysis (requiring storage at a stable pH). Also, simple dissociation to form unimolecular final products of a molecular weight low enough for renal clearance (less than ~10,000 g mol^{-1}) is a feasible approach (requiring storage above the critical micelle concentration)[39]. These and other chemistries will be introduced in more detail below.

Materials for penetrating mucosal barriers

Mucosal surfaces present a target for vaccination that is both appealing and challenging. Most pathogens invade the body through a mucosal route, be it the nasal cavity, the airways, the gut, the vagina or the rectum; therefore, establishment of a secretory IgA immune response in these tissues would be especially beneficial for providing protection. Delivery of vaccines to these surfaces is substantially complicated by the mucosal layer that otherwise protects them, at least in part, against pathogen entry. The mucus consists of a physically crosslinked, viscoelastic hydrogel, with mesh sizes of the order of 10–100 nm (ref. 40). Barrier penetration is largely restricted for particles that are greater in diameter than a few hundred nanometres[40,41], although particles that are ~50 nm in diameter can diffuse in mucus almost as freely as they do in water[40].

Particle surface properties have a major role in particle penetration of mucus. It has been observed that even very large particles, 200–500 nm in diameter, can penetrate mucus when appropriately grafted with surface PEG chains[42]. This effect depends strongly on the molecular weight of the grafted PEG, with grafts comprising shorter chains (2,000 g mol^{-1}) penetrating well but those comprising longer chains (10,000 g mol^{-1}) penetrating several orders of magnitude more slowly. The field of mucoadhesion has been explored from a polymer science perspective: surface-tethered polymer chains interpenetrate, and entangle within, the mucin polymer network, leading to adhesion associated with entanglement and disentanglement[43], and longer chains interpenetrate more effectively than shorter ones[44]. These concepts are illustrated in Fig. 4. Here, mucoadhesion is not beneficial, as the vaccine particles become entrapped in the mucosal barrier; therefore, PEG chains long enough to prevent adsorption, but not long enough to lead to entanglement, are desired.

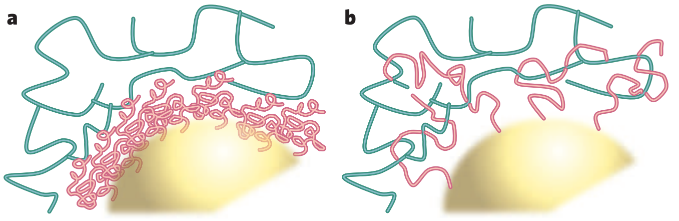


Figure 4 | Steric stabilization versus entanglement in mucus. Nanoparticles need to gain access to the mucosal epithelia for antigen delivery or transfection, so they must be able to penetrate the mucous layer. The grafting of polymers such as PEG (pink) to nanoparticles (yellow) has been explored as a way of blocking the adsorption of particles to components of the mucus (green). Studies into mucosal bioadhesion have examined various physical regimes of polymers. Shorter, denser graft layers tend to sterically stabilize the nanoparticle surface (**a**). By contrast, longer, sparser grafts allow interpenetration of the two polymers (the grafted chains and the mucous network) (**b**), leading to adhesion to the mucus[43,44,99] and unfavourable nanoparticle penetration[42,100].

Gene vectors, which are useful in DNA vaccination (that is, with the antigen being expressed from the delivered DNA), present a particular challenge in mucosal penetration, in that most nanoparticle complexes containing DNA are formed with cationic carriers, such as cationic lipids; these cationic charges can dramatically limit particle transit through the negatively charged mucous layer[45]. Cationic lipid mixtures have been developed with PEG-grafted lipid components to enhance particle penetration[46], although the polymer chains may interfere with later processes of endosomal destabilization and gene uptake. To address this, vectors with PEG chains that are removed by cellular processes (see the next section) during endosomal processing are being investigated[47].

Materials for intracellular targeting

As described above, the detection of both antigens and danger signals is complex and takes place in different compartments of the cell. For example, antigens for presentation by MHC class I pathways must be available within the cytosol, whereas those for presentation by MHC class II molecules must be present within the endolysosomal compartment (Fig. 2). With regard to danger signals for APC activation, even considering just the TLR family, receptors for some ligands (such as hydrophobic bacterial cell-wall components, which bind to TLR4, or bacterial flagellae components, which bind to TLR5) are present on the plasma membrane, whereas receptors for others (single-stranded RNA, which binds to TLR7, or unmethylated bacterial CpG DNA, which binds to TLR9) are present and active within the endolysosome. The spatial details of antigen and danger-signal delivery are therefore important. Several polymers that accomplish endolysosomal delivery are described below and are shown in Fig. 5.

A number of chemical reactions have been engineered to release particle payload within the endolysosomal compartments. As the antigen in the endosomal vesicles is processed and the vesicles mature towards lysosomal fusion, the intravesicular pH is lowered, first to a pH of ~6 in the endosome and then to a pH of ~5 in the lysosome, relative to the extracellular pH, 7.4. Additionally, the reduction–oxidation state of these compartments changes: the endosome is rendered reductive, whereas the lysosomal compartment is substantially oxidative, compared with the mildly oxidative extracellular environment. For these reasons, both pH-sensitive and reduction–oxidation-sensitive materials are being studied.

Oxidation at low pH represents the final stages of endolysosomal processing, with exposure in the lysosome to a number of reactive oxygen species. Oxidation-sensitive dissociation of self-assembling block co-polymers has been engineered, by designing block co-polymers (mentioned above) containing a hydrophobic PPS block; on exposure to oxidative conditions, the block is converted to hydrophilic polypropylene sulphoxide and, ultimately, to the more hydrophilic polypropylene sulphone[48]. When block co-polymer architectures are selected

© 2009 Macmillan Publishers Limited. All rights reserved

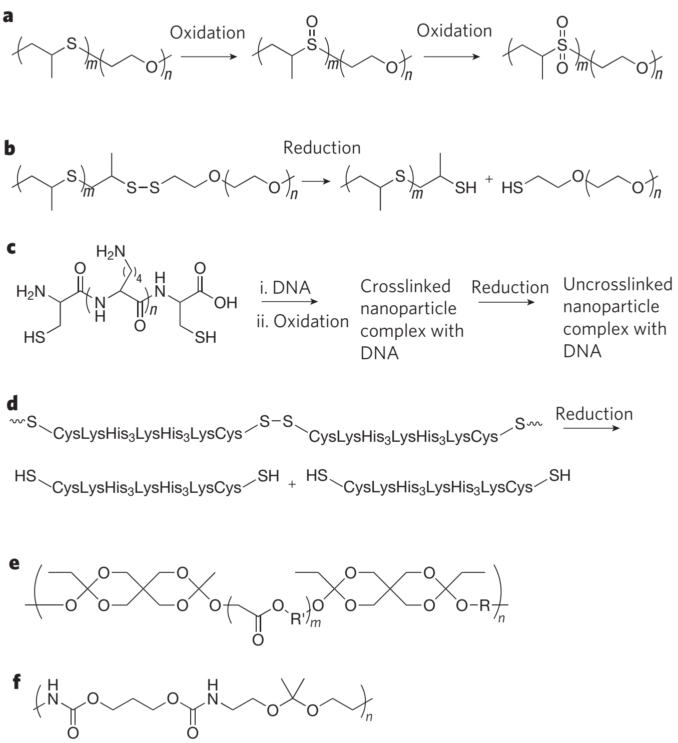

a
Oxidation
Oxidation
b
Reduction
c
i. DNA
ii. Oxidation
Crosslinked nanoparticle complex with DNA
Reduction
Uncrosslinked nanoparticle complex with DNA
d
CysLysHis3LysHis3LysCys
CysLysHis3LysHis3LysCys
Reduction
HS CysLysHis3LysHis3LysCys SH + HS CysLysHis3LysHis3LysCys SH
e
f

Figure 5 | Examples of polymers used for endolysosomal delivery. **a**–**d**, Polymers sensitive to oxidation within the lysosome (**a**) or reduction within the endosome (**b**–**d**). **a**, Oxidation of a hydrophobic sulphide, ultimately to a hydrophilic sulphone, leads to dissociation of self-assembled vesicles of the macroamphiphiles, which after oxidation are only hydrophilic and no longer amphiphilic[48]. **b**, Reduction of a disulphide link between the hydrophobic and hydrophilic blocks of a vesicle-forming macroamphiphile leads to vesicle rupture[35]. **c**, Reduction of an AB multiblock polymer leads to dissociation of a complex with DNA in the endosome[50]. **d**, Reduction of an analogous peptide-crosslinked DNA particle leads to DNA release[51]. **e**–**f**, Polymers sensitive to hydrolysis during acidification of the endolysosome, utilizing orthoesters (**e**) and ketals (**f**). R and R′ are usually alkyl groups, to adjust the hydrophobicity of the material.

with approximately equal block volumes, so as to form vesicles known as polymersomes, on oxidation the polymersomes transform into worm-like micelles, then into spherical micelles and finally into soluble polymer, as the ratio of the effective volumes of the hydrophobic block and the hydrophilic regions decreases eventually to zero[49], releasing the contents of the polymersome.

Earlier in the processes of endolysosomal processing, endocytosed nanoparticles encounter a reductive environment. Self-assembling block co-polymers with architecture PEG-SS-PPS, that is with a reducible disulphide connection between the hydrophobic and hydrophilic blocks, have been shown to destabilize within 15 min of endocytosis in a macrophage-like cell line (a model of APCs), releasing the contents of the polymersomes within the early endosome[35]. Linear multiblock co-polymers have also been developed; they consist of DNA-binding peptides that are flanked on each side by a cysteine residue and polymerized by oxidation *in vitro*[50]. The resultant linear, high-molecular-weight polymer binds DNA and condenses it into a nanoparticle, referred to as a polyplex, but reduction within the endosome leads to multiple chain scission to form products that bind only weakly to DNA, thereby releasing the material. If endosomal disruption is also accomplished (see below), efficient transfection with antigen-encoding DNA can be achieved[50]. Moreover, if nucleotide-based PAMPs, such as CpG oligonucleotides, are included, efficient delivery to the endosomal receptors can be achieved. Low-molecular-weight disulphide peptides have also been used to good effect as reduction-sensitive crosslinkers for the endosomal release of polyplexes[51].

The pH gradient experienced during endolysosomal processing has also been used for endosomal release. Whereas simple degradable polyesters such as polylactic acid, polyglycolic acid and their co-polymers are degraded by acid-catalysed hydrolysis, the hydrolysis rates at even lysosomal pH are so slow that they are not very useful for endosomal delivery. For this reason, more acid-sensitive, hydrolytically sensitive links, such as orthoesters[52] and ketals[53], have been sought. For example, particles crosslinked with ketal moieties have been developed for biomolecular delivery[53], including vaccination[54], through the targeting of endosomal release. Endosomal release of CpG oligonucleotides has also been shown to be feasible using these materials chemistries[55]. Degradation rates at lysosomal pH can be ~20-fold faster than at extracellular pH[55], providing the desired trigger.

Whereas the endosomal compartment is the interesting target for MHC class II loading, MHC class I presentation requires the antigen payload to be present in the cytosol. Thus, disruption of the endosomal membrane barrier to access the cytosol is an important target. Endosomal disruption is also necessary for DNA vaccination, in which plasmid DNA must be expressed to produce the antigen. Materials schemes that have been investigated for these purposes are illustrated in Fig. 6.

Both reductive and pH triggers have been used to control endosomal escape. With regard to reduction, the polymer fragments produced by reduction of the PEG-SS-PPS block co-polymers mentioned above apparently possessed sufficient surface activity to disrupt the endosomal membrane, making the payload available within the endosome within 15 min but within the cytosol within 2 h (ref. 35). Triggers sensitive to pH are more commonly sought. One approach involves the engineering of polymers that express pH-dependent surface activity and thus membrane-disruptive activity, for example polypropyl acrylic acid; these polymers display no membrane-disruptive activity at extracellular pH, but at pH 6–6.5 are strongly membranolytic[56]. These polymers have demonstrated strong potential for antigen delivery, processing, and presentation by MHC class I molecules[56]. Oligocations and polycations can also destabilize the cell membrane; cationic cell-penetrating peptides based on oligoarginine have been incorporated in pH-sensitive ketal-crosslinked nanoparticles — the membranolytic moieties becoming available as the pH is lowered and hydrolysis proceeds — with beneficial effect for intracellular accumulation[57].

Polycations can have an additional favourable effect on membrane destabilization, owing to an osmotic imbalance that ensues during their protonation. Polyethylene imine possesses a favourable pK_a value for protonation within the endosome, the associated osmotic effect being referred to as the ‘proton-sponge’ effect[58]. The use of such polymers is, however, hampered by cytotoxicity associated with polyethylene imine’s contact with the cell’s membranes. To retain the proton-sponge effect yet eliminate cytotoxicity, core–shell nanoparticles have been studied using sequential emulsion polymerization, first of a secondary-amine-containing monomer of appropriate pK_a (diethylaminoethyl methacrylate), to form a core, and then of a second monomer, to form a corona (or shell)[59]. Regional separation of the functions of the particle led to favourable cytosolic release, through endosomal disruption by means of the proton-sponge effect, with markedly lower cytotoxicity than caused by free polyethylene imine[60].

Materials interaction with, and penetration of, cell membranes to access the cytosol is complex, involving hydrophobic as well as electrostatic interactions. One new class of particle has demonstrated intriguing potential, based on an intrinsic membrane-penetrating ability derived from its spatial distributions of hydrophobicity and charge. Each particle consists of a spatially heterogeneous organic adsorbed layer formed atop a 6-nm gold nanoparticle to create <1-nm striations of hydrophobicity and negative charge[61]. These particles, but not those with the same overall charge density but lacking the striations, were able to pass directly through the plasma membrane. This concept, if it can be used as a general design principle for other particle implementations, will be very powerful.

The cytosol represents a particularly attractive target for antigen-encoding DNA, as well as for protein antigen. If materials approaches to cytosolic targeting can allow highly efficient DNA delivery concurrently with prolonged PRR activation, dendritic cells could become a very attractive target for DNA vaccination. Given the logistical advantages

© 2009 Macmillan Publishers Limited. All rights reserved

of delivery of antigen-encoding DNA (speed of production, cost and stability) relative to those of delivery of protein antigen, the attraction of such approaches for counteracting pathogens that vary seasonally (such as influenza virus) and pathogens that mainly affect the developing world is considerable.

Materials to trigger immune-specific functions

Now that we have introduced the systems by which to deliver antigens and danger signals to specific cellular and subcellular compartments, we consider how the materials affect dendritic cells either directly or indirectly, for example through mediation by protein–material interactions or through a delivered biomolecular payload (although much of this, with the exception of the antigen, has been addressed above). Because PRRs are fundamental to the initiation of immunity, owing to their mediation of the recognition of PAMPs and DAMPs, biomaterials are being implemented to deliver natural or synthetic immunomodulatory agents to these receptors. However, the material itself may also be intrinsically biologically active, by virtue of its particulate character or as a result of protein interactions at the biomaterial surface.

The role of particle size and shape

As discussed above, particle size can control the biological transport and hence the bioavailability of a material to a remarkable degree, for example through the tissue interstitium or across a mucosal barrier. However, material size characteristics may also be an important determinant of a material's immunological activity. For example, the most common adjuvants in clinical use are insoluble aluminium phosphates (alum), which form aggregates with the protein antigen and create an antigen depot. These particles, which might be considered chemically inert, enhance cellular IL-1β secretion[62] by means of NALP3 (NACHT domain-, leucine-rich repeat-, and PYD-containing protein 3, also known as NLRP3; ref. 63) to induce antibody-mediated protective immunity; therefore, the particles are themselves intrinsically recognized as a sign of danger. Because material particles can be both effective and economical, understanding and engineering this effect is of great interest. Recent work suggests that this is possible, with evidence that polymeric microparticles endocytosed by dendritic cells trigger the inflammasome by means of NALP3 and, in concert with endogenous signals, induce both humoral and cellular immunity[64]. *In vitro*, inflammasome-activation-associated IL-1β secretion by dendritic cells in response to particle treatment (in addition to LPS stimulation) was size dependent and maximal at particle diameters between 400 and 1,000 nm (ref. 64). Given the <100-nm size limit for interstitial transport to access the lymph nodes, where large number of dendritic cells reside, the engineering of nanoscale particles to activate the inflammasome efficiently warrants further study.

Particle shape may also be important; for example, for micrometre-scale particles the ability of macrophages to phagocytose a particle depends more on its shape than on its size[65]. This phenomenon is determined by actin mechanics at the points of particle contact. These results raise the intriguing possibility that shape may have such a role, either in uptake or in APC activation, at shorter length scales as well.

The role of particle hydrophobicity

Within the diverse family of TLR4-ligand and TLR2-ligand PAMPs (including LPS, lipopeptide and peptidoglycan) and DAMPs (including hyaluronan fragments, heat-shock proteins and fibronectin), an underlying biochemical thread may be the presence of hydrophobic domains, suggesting hydrophobicity to be a universal sign of danger sensed by TLRs[18]. Exploiting this connection may be of great interest as a potentially economical strategy by which to ligate hydrophobicity-sensing PRRs and possibly avoid costly conjugation schemes for recombinant protein ligands or synthetic PRR ligands. The supramolecular organization of biomacromolecules that contain hydrophobic domains and have the capacity to hide or expose these hydrophobic patches seems to determine their ability to activate immunity[18].

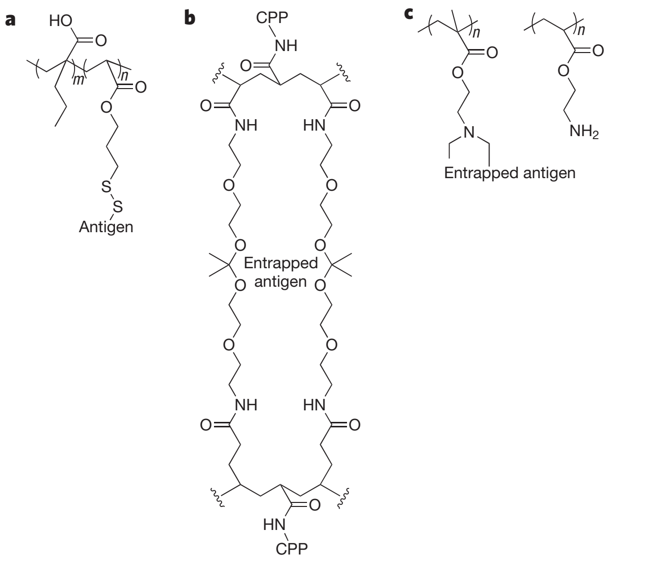


Figure 6 | Examples of polymers used for cytosolic delivery. a, Polypropyl acrylic acid disrupts membranes in a pH-dependent manner, with membranolytic activity at about the pH of early endosomes (~6.5). A block co-polymer is shown, of polypropyl acrylic acid and a monomer linked through a disulphide bond to the antigen, which can be released in the reductive environment of the endosome[56]. **b,** Crosslinked particles that are hydrolytically sensitive at endosomal pH, releasing a cell-penetrating peptide (CPP, which consists of polyarginine[57]), are formed by inverse emulsion polymerization of acrylamide, a CPP-grafted acrylamide and a ketal-containing bisacrylamide crosslinker, yielding the polymer shown here. Hydrolysis both degrades the crosslinks in the polymer and releases the CPP, destabilizing the endosome. **c,** The proton-sponge effect[58] for endosomal disruption, and thus cytosolic delivery, has been implemented in core–shell nanoparticles by first polymerizing particles of diethylaminoethyl methacrylate and then sequentially polymerizing aminoethyl methacrylate, both with crosslinking[59]. One monomer forms the core (left), and one forms the corona, or shell (right).

Given that microparticle formulations can be formed by nanoscale self-assembly of hydrophobic microdomains, approaches could be designed to control the degree of nanoparticulate hydrophobic-domain exposure, to exploit such a mechanism. However, the task of preserving a hydrophobic material surface once exposed is difficult: in the presence of protein in the biological environment, protein adsorption can rapidly obscure hydrophobic interfaces. From the perspective of inducing immunity, adsorption may have some advantages; the roles of immunoregulatory molecules (such as scavenger receptors and complement component C1q) in controlling normal hydrophobic molecule transport and clearance rather than in initiating an immune response[18], as well as in molecular recognition specificity in TLR4 ligation rather than in TLR2 ligation (as these ligands induce T_H1-type and T_H2-type responses, respectively), remain to be elucidated to determine materials design guidelines for such a strategy. Thus, materials hydrophobicity may itself be a danger signal or may be interpreted by dendritic cells through the intermediating layer of adsorbed plasma proteins.

Complement activation

Whereas hydrophobicity in materials modulates relatively nonspecific protein adsorptive interactions, certain materials features can mimic features of pathogen surfaces to activate innate immune pathways. One such recognition cascade is complement, the alternative pathway of which recognizes certain primary hydroxyls[66] and other surface nucleophiles[67] to react at the site of a strained thioester in C3, forming material-bound activation product C3b[68]. Although much biomaterials research seeks to avoid such interactions, immunobioengineering can exploit complement activation, particularly in light of the variety of ways in which complement can affect innate and adaptive immunity[69,70]. Notably, the C3 activation products C3d and C3b have been shown to be molecular adjuvants capable of inducing strong antigen-specific

© 2009 Macmillan Publishers Limited. All rights reserved

humoral immunity[71–73], and materials have been developed to exploit C3 activation for adaptive immunity[28].

Studies suggest that surface biochemistry such as sulphation may control the activation and deposition of complement species on cellular or material surfaces by controlling the adsorption of complement factor H (CFH) and CFD[74,75]. Generating complement-opsonized particulates *in situ* by modulating material surface chemistry may therefore represent an inexpensive and powerful strategy to harness the molecular adjuvant properties of C3b and C3d. Interestingly, C1q, which binds to immunoglobulin on antigen ligation, also binds to hydrophobic molecules or aggregates such as LPS and liposomes[18]. Hence, the incorporation of hydrophobic domains could activate complement through the classical pathway to use the immunomodulatory properties of C2 and C4.

Functionalization and encapsulation

We have seen that the biological context in which the targeted PRR is encountered by the material (for example at the cell surface or in the endosome) must be considered when designing materials delivering bioavailable molecules for PRR ligation. In addition, the benefits of antigen display relative to encapsulation and of danger-signal co-encapsulation relative to co-delivery must be considered. For polymers that degrade too slowly for antigen encapsulation schemes, such as polylactic-co-glycolic acid, adsorption of antigen onto biomaterial particles may be more beneficial[76]. Within these systems, co-encapsulation of PAMPs, such as CpG oligonucleotides or TLR4 ligands, is much more beneficial than co-administration, providing support for prolonged stimulation of dendritic cells after antigen collection[77–80]. Indeed, when dendritic cells encounter both self antigen and pathogen-derived antigen, they use the coexistence of antigen and TLR ligand within the same endosome to distinguish between the two and enhance presentation of the pathogen-derived antigen on MHC class II molecules[81]. Bringing the antigen together with the PAMP, so as to model the situation found in the pathogen, therefore seems to be a sound design principle.

Materials as models for basic immunobiology

When immune cells interact with pathogens, or with each other, they do so in a complex display of a number of molecular mediators of antigen uptake, processing, presentation and activation. The challenge of teasing out molecular mechanisms from this process in its full complexity can be daunting. Materials science in immunobioengineering allows the development of tools with which to assess molecular and cellular hypotheses. Whereas the sections above highlight research goals with translational ends, here we briefly touch on some that are more basic and mechanistic.

One opportunity for biomaterials research in immunobioengineering is the creation of synthetic pathogens for the study of dendritic-cell activation and downstream interactions with T cells that trigger adaptive immunity. Because dendritic cells have evolved to recognize such a diverse array of PAMPs, and because a number of such PAMPs are found in any one pathogen, it is difficult to study the molecular interactions in simple, isolated systems. Biomaterials can provide a blank slate on which antigen and a defined set of PAMPs can be displayed in defined amounts for mechanistic investigation. As an example, we refer to the previously mentioned study on nanoparticle-induced activation of complement by means of surface-tethered PEG, in which complement was the only DAMP in the absence of any pathogen-associated signals[28]. Investigation of the pairwise interactions of particle size and complement activation demonstrated that complement is a powerful trigger of humoral and cellular immunity as long as the complement-decorated particles are small enough (<50 nm in diameter) to enter the lymphatics and thus target lymph-node APCs[28]. Many other such mechanistic studies could use biomaterials as model systems, controlling the destination of PAMP presentation (plasma membrane or endolysosome, or both) and the identity of the PAMPs (for example with multiple TLR ligands), with independent control of the numbers of PAMP molecules, their clustering and spatial organization and even their duration of exposure.

Investigating the interactions between immune cells themselves presents another interesting challenge for materials science in immunobioengineering. For example, when activated dendritic cells present their antigen to T cells, a spatially organized structure referred to as the immunological synapse is created, in which the peptide-antigen-loaded MHC molecule on the dendritic cell is presented to the TCR on the T cell; this receptor pair is organized after cellular contact such that the peptide–MHC–TCR pair (one receptor on each of the two cells) is clustered and is surrounded by an adhesion-receptor pair consisting of intercellular adhesion molecule 1 (ICAM1) on the dendritic cell binding to lymphocyte function-associated antigen 1 (LFA1) on the T cell[82]. It was previously unknown whether this evolving geometric patterning was required for the function of the immunological synapse or was only associated with it. Using supported lipid membranes in which bound peptide–MHC and ICAM1 could freely move laterally, a functional immunological synapse could be formed, but when barriers were created to limit lateral reorganization of peptide–MHC and ICAM1, function was inhibited[83]; this suggested that the biological structuring is required for function. Likewise, when a TCR ligand was lithographically patterned in spots surrounded by ICAM1, a functional immunological synapse resulted, whereas when the geometry was reversed, no such function was possible[84]. In both of these examples, sophisticated materials science approaches allowed well-controlled models of immune-cell interactions to probe biological hypotheses.

Future prospects

Materials science has a great deal to offer the field of immunology: immunobioengineering of prophylactic and therapeutic vaccine platforms, and model systems with which to explore the molecular and cellular interactions between immune cells and pathogens (and between different classes of immune cell), are just a few examples of this. The application of materials science towards therapy and prophylaxis is instructed by immunobiology, and we have shown how materials science can in return be used to instruct immunobiology. The key feature that materials offer is that of design: design for encapsulation, design for immobilization or release of one or several biomolecular regulators of immune interactions, design for material functionalities such as release or membrane disruption in particular parts of the cell, and design for doing this in particular cellular and tissue targets. Some avenues may be more fruitful than others. Functional polymersomes are particularly attractive, in that both antigen and danger-signal payloads may be incorporated in the watery vesicle core, and other danger signals or targeting ligands may be attached to the polymersome surface or within the hydrophobic leaflet of the membrane. With materials chemistries leading to disruption and release within the endosome, and even destabilization of the endosomal membrane itself, a means of delivery for presentation by both MHC class II molecules and MHC class I molecules seems within reach. Still more interesting is the possibility of combining such means with physiological routes for delivery, for example by using ultrasmall self-assembled nanostructures to deliver such advanced materials directly to the lymphatics or to lymphoid tissue associated with the mucosae.

Clearly, even with the objective of vaccination described above, much work remains to be done, in that only very early implementations are clinically available or are in advanced stages of testing, rather than still being researched. As our understanding of immunobiology grows, so will the range of principles for the design of materials and material–biomolecular conjugates used in immunotherapeutics. Moreover, the design principles will be different in various contexts: for example, in therapeutic vaccines against cancer, complex materials and formulations may be contemplated, as cost does not present a major consideration. By contrast, vaccination in a global context places severe constraints on logistics (wet formulations and unrefrigerated storage), use (preferably needle-free administration routes and few doses) and

© 2009 Macmillan Publishers Limited. All rights reserved

cost (materials that can be easily manufactured and contain a minimal number of biological molecules). Although constraints are an annoyance from a design perspective, they can also present exciting intellectual challenges for the materials scientist and the immunobioengineer. The opportunity to combine possibilities for translation with those of using materials systems to learn more about the underpinning science, immunobiology, is also tremendously exciting. ■

1. Trombetta, E. S. & Mellman, I. Cell biology of antigen processing *in vitro* and *in vivo*. *Annu. Rev. Immunol.* **23,** 975–1028 (2005).
2. Steinman, R. M. & Hemmi, H. Dendritic cells: translating innate to adaptive immunity. *Curr. Top. Microbiol. Immunol.* **311,** 17–58 (2006).
3. Perrigoue, J. G. *et al.* MHC class II-dependent basophil–CD4$^+$ T cell interactions promote T$_H$2 cytokine-dependent immunity. *Nature Immunol.* **10,** 697–705 (2009).
4. Wilson, N. S. *et al.* Most lymphoid organ dendritic cell types are phenotypically and functionally immature. *Blood* **102,** 2187–2194 (2003).
5. Randolph, G. J., Angeli, V. & Swartz, M. A. Dendritic-cell trafficking to lymph nodes through lymphatic vessels. *Nature Rev. Immunol.* **5,** 617–628 (2005).
6. Dudziak, D. *et al.* Differential antigen processing by dendritic cell subsets *in vivo*. *Science* **315,** 107–111 (2007).
7. Singh, A. *et al.* Efficient modulation of T-cell response by dual-mode, single-carrier delivery of cytokine-targeted siRNA and DNA vaccine to antigen-presenting cells. *Mol. Ther.* **16,** 2011–2021 (2008).
8. Cubas, R. *et al.* Virus-like particle (VLP) lymphatic trafficking and immune response generation after immunization by different routes. *J. Immunother.* **32,** 118–128 (2009).
9. Lammermann, T. & Sixt, M. The microanatomy of T-cell responses. *Immunol. Rev.* **221,** 26–43 (2008).
10. Itano, A. A. *et al.* Distinct dendritic cell populations sequentially present antigen to CD4 T cells and stimulate different aspects of cell-mediated immunity. *Immunity* **19,** 47–57 (2003).
11. Pape, K. A., Catron, D. M., Itano, A. A. & Jenkins, M. K. The humoral immune response is initiated in lymph nodes by B cells that acquire soluble antigen directly in the follicles. *Immunity* **26,** 491–502 (2007).
12. Förster, R., Davalos-Misslitz, A. & Rot, A. CCR7 and its ligands: balancing immunity and tolerance. *Nature Rev. Immunol.* **8,** 362–371 (2008).
13. Schneider, M. A., Meingassner, J. G., Lipp, M., Moore, H. D. & Rot, A. CCR7 is required for the *in vivo* function of CD4$^+$ CD25$^+$ regulatory T cells. *J. Exp. Med.* **204,** 735–745 (2007).
14. Holmgren, J. & Czerkinsky, C. Mucosal immunity and vaccines. *Nature Med.* **11,** S45–S53 (2005).
15. Heath, W. R. *et al.* Cross-presentation, dendritic cell subsets, and the generation of immunity to cellular antigens. *Immunol. Rev.* **199,** 9–26 (2004).
16. Lee, M. S. & Kim, Y. J. Signaling pathways downstream of pattern-recognition receptors and their cross talk. *Annu. Rev. Biochem.* **76,** 447–480 (2007).
17. Macagno, A., Napolitani, G., Lanzavecchia, A. & Sallusto, F. Duration, combination and timing: the signal integration model of dendritic cell activation. *Trends Immunol.* **28,** 227–233 (2007).
18. Seong, S. Y. & Matzinger, P. Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. *Nature Rev. Immunol.* **4,** 469–478 (2004).
19. Zarember, K. A. & Godowski, P. J. Tissue expression of human Toll-like receptors and differential regulation of Toll-like receptor mRNAs in leukocytes in response to microbes, their products, and cytokines. *J. Immunol.* **168,** 554–561 (2002).
20. Jiang, W. *et al.* The receptor DEC-205 expressed by dendritic cells and thymic epithelial cells is involved in antigen processing. *Nature* **375,** 151–155 (1995).
21. Nchinda, G. *et al.* The efficacy of DNA vaccination is enhanced in mice by targeting the encoded protein to dendritic cells. *J. Clin. Invest.* **118,** 1427–1436 (2008).
This paper gives a demonstration of a DNA vaccine encoding an antigen targeted at dendritic cells, through expression of an antigen fusion protein with single-chain variable antibody fragments directed against DEC205.
22. Kwon, Y. J., James, E., Shastri, N. & Fréchet, J. M. *In vivo* targeting of dendritic cells for activation of cellular immunity using vaccine carriers based on pH-responsive microparticles. *Proc. Natl Acad. Sci. USA* **102,** 18264–18268 (2005).
In this paper, ketal-crosslinked acid-sensitive nanoparticles are functionalized with anti-DEC205 antibodies, providing multiple levels of nanoparticle functionality.
23. Zhao, X., Jain, S., Larman, H. B., Gonzalez, S. & Irvine, D. J. Directed cell migration via chemoattractants released from degradable microspheres. *Biomaterials* **26,** 5048–5063 (2005).
24. Hori, Y., Winans, A. M. & Irvine, D. J. Modular injectable matrices based on alginate solution/microsphere mixtures that gel *in situ* and co-deliver immunomodulatory factors. *Acta Biomater.* **5,** 969–982 (2009).
25. Ali, O. A., Huebsch, N., Cao, L., Dranoff, G. & Mooney, D. J. Infection-mimicking materials to program dendritic cells *in situ*. *Nature Mater.* **8,** 151–158 (2009).
26. Hori, Y., Winans, A. M., Huang, C. C., Horrigan, E. M. & Irvine, D. J. Injectable dendritic cell-carrying alginate gels for immunization and immunotherapy. *Biomaterials* **29,** 3671–3682 (2008).
27. Swartz, M. A. The physiology of the lymphatic system. *Adv. Drug Deliv. Rev.* **50,** 3–20 (2001).
28. Reddy, S. T. *et al.* Exploiting lymphatic transport and complement activation in nanoparticle vaccines. *Nature Biotechnol.* **25,** 1159–1164 (2007).
29. Zubay, G. *Biochemistry* 1052–1053 (Addison-Wesley, 1983).
30. Rehor, A., Tirelli, N. & Hubbell, J. A. A new living emulsion polymerization mechanism: episulfide anionic polymerization. *Macromolecules* **35,** 8688–8693 (2002).
31. Rehor, A., Tirelli, N. & Hubbell, J. A. Novel carriers based on polysulfide nanoparticles: production via living emulsion polymerization, characterization and preliminary carrier assessment. *J. Control Release* **87,** 246–247 (2003).
32. Cerritelli, S., Velluto, D., Hubbell, J. A. & Fontana, A. Breakdown kinetics of aggregates from poly(ethylene glycol-*bl*-propylene sulfide) di- and triblock copolymers induced by a non-ionic surfactant. *J. Polym. Sci. A* **46,** 2477–2487 (2008).
33. Velluto, D., Demurtas, D. & Hubbell, J. A. PEG-*b*-PPS diblock copolymer aggregates for hydrophobic drug solubilization and release: cyclosporin A as an example. *Mol. Pharm.* **5,** 632–642 (2008).
34. Todd, C. W. *et al.* Development of an adjuvant-active nonionic block copolymer for use in oil-free subunit vaccines formulations. *Vaccine* **15,** 564–570 (1997).
35. Cerritelli, S., Velluto, D. & Hubbell, J. A. PEG-SS-PPS: reduction-sensitive disulfide block copolymer vesicles for intracellular drug delivery. *Biomacromolecules* **8,** 1966–1972 (2007).
36. Roy, P. & Noad, R. Virus-like particles as a vaccine delivery system: myths and facts. *Hum. Vaccin.* **4,** 5–12 (2008).
37. Gillies, E. R. & Frechet, J. M. J. Dendrimers and dendritic polymers in drug delivery. *Drug Discov. Today* **10,** 35–43 (2005).
38. Sheng, K. C. *et al.* Delivery of antigen using a novel mannosylated dendrimer potentiates immunogenicity *in vitro* and *in vivo*. *Eur. J. Immunol.* **38,** 424–436 (2008).
39. Yamaoka, T., Tabata, Y. & Ikada, Y. Distribution and tissue uptake of poly(ethylene glycol) with different molecular weights after intravenous administration to mice. *J. Pharm. Sci.* **83,** 601–606 (1994).
40. Cone, R. A. Barrier properties of mucus. *Adv. Drug Deliv. Rev.* **61,** 75–85 (2009).
41. Lai, S. K., Wang, Y. Y. & Hanes, J. Mucus-penetrating nanoparticles for drug and gene delivery to mucosal tissues. *Adv. Drug Deliv. Rev.* **61,** 158–171 (2009).
42. Lai, S. K. *et al.* Rapid transport of large polymeric nanoparticles in fresh undiluted human mucus. *Proc. Natl Acad. Sci. USA* **104,** 1482–1487 (2007).
43. Peppas, N. A., Hansen, P. J. & Buri, P. A. A theory of molecular diffusion in the intestinal mucus. *Int. J. Pharm.* **20,** 107–118 (1984).
44. Huang, Y. B., Szleifer, I. & Peppas, N. A. Gel–gel adhesion by tethered polymers. *J. Chem. Phys.* **114,** 3809–3816 (2001).
45. Sanders, N., Rudolph, C., Braeckmans, K., De Smedt, S. C. & Demeester, J. Extracellular barriers in respiratory gene therapy. *Adv. Drug Deliv. Rev.* **61,** 115–127 (2009).
46. Sanders, N. N., De Smedt, S. C., Cheng, S. H. & Demeester, J. Pegylated GL67 lipoplexes retain their gene transfection activity after exposure to components of CF mucus. *Gene Ther.* **9,** 363–371 (2002).
47. Meyer, M. & Wagner, E. pH-responsive shielding of non-viral gene vectors. *Expert Opin. Drug Deliv.* **3,** 563–571 (2006).
48. Napoli, A., Valentini, M., Tirelli, N., Muller, M. & Hubbell, J. A. Oxidation-responsive polymeric vesicles. *Nature Mater.* **3,** 183–189 (2004).
49. Napoli, A., Bermudez, H. & Hubbell, J. A. Interfacial reactivity of block copolymers: understanding the amphiphile-to-hydrophile transition. *Langmuir* **21,** 9149–9153 (2005).
50. Manickam, D. S. & Oupický, D. Multiblock reducible copolypeptides containing histidine-rich and nuclear localization sequences for gene delivery. *Bioconjug. Chem.* **17,** 1395–1403 (2006).
51. McKenzie, D. L., Smiley, E., Kwok, K. Y. & Rice, K. G. Low molecular weight disulfide cross-linking peptides as nonviral gene delivery carriers. *Bioconjug. Chem.* **11,** 901–909 (2000).
52. Wang, C. *et al.* Molecularly engineered poly(ortho ester) microspheres for enhanced delivery of DNA vaccines. *Nature Mater.* **3,** 190–196 (2004).
53. Paramonov, S. E. *et al.* Fully acid-degradable biocompatible polyacetal microparticles for drug delivery. *Bioconjug. Chem.* **19,** 911–919 (2008).
54. Cohen, J. A. *et al.* T-cell activation by antigen-loaded pH-sensitive hydrogel particles *in vivo*: the effect of particle size. *Bioconjug. Chem.* **20,** 111–119 (2009).
55. Heffernan, M. J., Kasturi, S. P., Yang, S. C., Pulendran, B. & Murthy, N. The stimulation of CD8$^+$ T cells by dendritic cells pulsed with polyketal microparticles containing ion-paired protein antigen and poly(inosinic acid)-poly(cytidylic acid). *Biomaterials* **30,** 910–918 (2009).
In this paper, acid-labile microparticles are developed to degrade quickly after endocytosis, releasing antigen and a ligand for an intracellular triglyceride-rich lipoprotein.
56. Flanary, S., Hoffman, A. S. & Stayton, P. S. Antigen delivery with poly(propylacrylic acid) conjugation enhances MHC-1 presentation and T-cell activation. *Bioconjug. Chem.* **20,** 241–248 (2009).
In this paper, a pH-sensitive polymer is demonstrated to destabilize the endosomal membrane after endocytosis, allowing access of an associated protein antigen to the cytosol and therefore MHC class I presentation.
57. Cohen, J. L. *et al.* Enhanced cell penetration of acid-degradable particles functionalized with cell-penetrating peptides. *Bioconjug. Chem.* **19,** 876–881 (2008).
58. Boussif, O. *et al.* A versatile vector for gene and oligonucleotide transfer into cells in culture and *in vivo*: polyethylenimine. *Proc. Natl Acad. Sci. USA* **92,** 7297–7301 (1995).
59. Hu, Y. *et al.* Cytosolic delivery of membrane-impermeable molecules in dendritic cells using pH-responsive core–shell nanoparticles. *Nano Lett.* **7,** 3056–3064 (2007).
60. Hu, Y. *et al.* Cytosolic delivery mediated via electrostatic surface binding of protein, virus, or siRNA cargos to pH-responsive core–shell gel particles. *Biomacromolecules* **13,** 756–765 (2009).
61. Verma, A. *et al.* Surface-structure-regulated cell-membrane penetration by monolayer-protected nanoparticles. *Nature Mater.* **7,** 588–595 (2008).
This paper suggests that organization of charge and hydrophobicity on nanoparticle surfaces on the single-nanometre scale can allow access to the cytoplasm without endocytosis.
62. Li, H., Nookala, S. & Re, F. Aluminum hydroxide adjuvants activate caspase-1 and induce IL-1β and IL-18 release. *J. Immunol.* **178,** 5271–5276 (2007).
63. Franchi, L. & Nunez, G. The Nlrp3 inflammasome is critical for aluminium hydroxide-mediated IL-1β secretion but dispensable for adjuvant activity. *Eur. J. Immunol.* **38,** 2085–2089 (2008).
64. Sharp, F. A. *et al.* Uptake of particulate vaccine adjuvants by dendritic cells activates the NALP3 inflammasome. *Proc. Natl Acad. Sci. USA* **106,** 870–875 (2009).
65. Champion, J. A. & Mitragotri, S. Role of target geometry in phagocytosis. *Proc. Natl Acad. Sci. USA* **103,** 4930–4934 (2006).
66. Kidane, A. & Park, K. Complement activation by PEO-grafted glass surfaces. *J. Biomed. Mater. Res.* **48,** 640–647 (1999).
67. Tang, L. P., Liu, L. & Elwing, H. B. Complement activation and inflammation triggered by model biomaterial surfaces. *J. Biomed. Mater. Res.* **41,** 333–340 (1998).
68. Gadjeva, M. *et al.* The covalent binding reaction of complement component C3. *J. Immunol.* **161,** 985–990 (1998).

© 2009 Macmillan Publishers Limited. All rights reserved

69. Kemper, C. & Atkinson, J. P. T-cell regulation: with complements from innate immunity. *Nature Rev. Immunol.* **7,** 9–18 (2007).
70. Carroll, M. C. The complement system in regulation of adaptive immunity. *Nature Immunol.* **5,** 981–986 (2004).
71. Dempsey, P. W., Allison, M. E., Akkaraju, S., Goodnow, C. C. & Fearon, D. T. C3d of complement as a molecular adjuvant: bridging innate and acquired immunity. *Science* **271,** 348–350 (1996).
72. Kolla, R. V. *et al.* Complement C3d conjugation to anthrax protective antigen promotes a rapid, sustained, and protective antibody response. *PLoS ONE* **2,** e1044 (2007).
73. Villiers, M. B., Villiers, C. L., Laharie, A. M. & Marche, P. N. Amplification of the antibody response by C3b complexed to antigen through an ester link. *J. Immunol.* **162,** 3647–3652 (1999).
74. Jokiranta, T. S. *et al.* Structure of complement factor H carboxyl-terminus reveals molecular basis of atypical haemolytic uremic syndrome. *EMBO J.* **25,** 1784–1794 (2006).
75. Pascual, M., Plastre, O., Montdargent, B., Labarre, D. & Schifferli, J. A. Specific interactions of polystyrene biomaterials with factor D of human complement. *Biomaterials* **14,** 665–670 (1993).
76. Singh, M. *et al.* Polylactide-co-glycolide microparticles with surface adsorbed antigens as vaccine delivery systems. *Curr. Drug Deliv.* **3,** 115–120 (2006).
77. Malyala, P. *et al.* The potency of the adjuvant, CpG oligos, is enhanced by encapsulation in PLG microparticles. *J. Pharm. Sci.* **97,** 1155–1164 (2008).
In this paper, co-delivery, in the same degradable particle, of antigen and TLR ligand is demonstrated to be beneficial relative to separate but simultaneous delivery.
78. Malyala, P., O'Hagan, D. T. & Singh, M. Enhancing the therapeutic efficacy of CpG oligonucleotides using biodegradable microparticles. *Adv. Drug Deliv. Rev.* **61,** 218–225 (2009).
79. Schlosser, E. *et al.* TLR ligands and antigen need to be coencapsulated into the same biodegradable microsphere for the generation of potent cytotoxic T lymphocyte responses. *Vaccine* **26,** 1626–1637 (2008).
80. Hamdy, S. *et al.* Co-delivery of cancer-associated antigen and Toll-like receptor 4 ligand in PLGA nanoparticles induces potent CD8$^+$ T cell-mediated anti-tumor immunity. *Vaccine* **26,** 5046–5057 (2008).
81. Blander, J. M. & Medzhitov, R. Toll-dependent selection of microbial antigens for presentation by dendritic cells. *Nature* **440,** 808–812 (2006).
82. Monks, C. R. F., Freiberg, B. A., Kupfer, H., Sciaky, N. & Kupfer, A. Three-dimensional segregation of supramolecular activation clusters in T cells. *Nature* **395,** 82–86 (1998).
83. Mossman, K. D., Campi, G., Groves, J. T. & Dustin, M. L. Altered TCR signaling from geometrically repatterned immunological synapses. *Science* **310,** 1191–1193 (2005).
84. Doh, J. & Irvine, D. J. Immunological synapse arrays: patterned protein surfaces that modulate immunological synapse structure formation in T cells. *Proc. Natl Acad. Sci. USA* **103,** 5700–5705 (2006).
In this paper, surface patterning methods are used to demonstrate the importance of ultrastructural organization of the interface between a T cell and a dendritic cell.
85. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF-κB by Toll-like receptor 3. *Nature* **413,** 732–738 (2001).
86. Lee, M. S. & Kim, Y. J. Pattern-recognition receptor signaling initiated from extracellular, membrane, and cytoplasmic space. *Mol. Cell* **23,** 1–10 (2007).
87. Strindelius, L., Filler, M. & Sjoholm, I. Mucosal immunization with purified flagellin from *Salmonella* induces systemic and mucosal immune responses in C3H/HeJ mice. *Vaccine* **22,** 3797–3808 (2004).
88. Nempont, C. *et al.* Deletion of flagellin's hypervariable region abrogates antibody-mediated neutralization and systemic activation of TLR5-dependent immunity. *J. Immunol.* **181,** 2036–2043 (2008).
89. Salman, H. H., Gamazo, C., Campanero, M. A. & Irache, J. M. *Salmonella*-like bioadhesive nanoparticles. *J. Control Release* **106,** 1–13 (2005).
90. Novak, N., Yu, C. F., Bieber, T. & Allam, J. P. Toll-like receptor 7 agonists and skin. *Drug News Perspect.* **21,** 158–165 (2008).
91. Thomsen, L. L., Topley, P., Daly, M. G., Brett, S. J. & Tite, J. P. Imiquimod and resiquimod in a mouse model: adjuvants for DNA vaccination by particle-mediated immunotherapeutic delivery. *Vaccine* **22,** 1799–1809 (2004).
92. Vollmer, J. *et al.* Characterization of three CpG oligodeoxynucleotide classes with distinct immunostimulatory activities. *Eur. J. Immunol.* **34,** 251–262 (2004).
93. Dostert, C. *et al.* Innate immune activation through Nalp3 inflammasome sensing of asbestos and silica. *Science* **320,** 674–677 (2008).
94. van Lookeren Campagne, M., Wiesmann, C. & Brown, E. J. Macrophage complement receptors and pathogen clearance. *Cell. Microbiol.* **9,** 2095–2102 (2007).
95. Liu, F. J. *et al.* Independent but not synergistic enhancement to the immunogenicity of DNA vaccine expressing HIV-1 gp120 glycoprotein by codon optimization and C3d fusion in a mouse model. *Vaccine* **22,** 1764–1772 (2004).
96. Villadangos, J. A. & Schnorrer, P. Intrinsic and cooperative antigen-presenting functions of dendritic-cell subsets *in vivo. Nature Rev. Immunol.* **7,** 543–555 (2007).
97. Conner, S. D. & Schmid, S. L. Regulated portals of entry into the cell. *Nature* **422,** 37–44 (2003).
98. Blander, J. M. & Medzhitov, R. On regulation of phagosome maturation and antigen presentation. *Nature Immunol.* **7,** 1029–1035 (2006).
99. Huang, Y. B., Leobandung, W., Foss, A. & Peppas, N. A. Molecular aspects of muco- and bioadhesion: tethered structures and site-specific surfaces. *J. Control Release* **65,** 63–71 (2000).
In this paper, the mechanism by which grafted polymer chains lead to adhesion to mucus is demonstrated to derive from entanglement, providing a basis for understanding stealth behaviour in the airways.
100. Wang, Y. Y. *et al.* Addressing the PEG mucoadhesivity paradox to engineer nanoparticles that 'slip' through the human mucus barrier. *Angew. Chem. Int. Edn Engl.* **47,** 9726–9729 (2008).
In this paper, the mechanistic understanding of ref. 99 is applied to provide vaccine particles that can penetrate mucus, providing an efficient route to mucosal vaccination.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to J.A.H. (jeffrey.hubbell@epfl.ch) or M.A.S. (melody.swartz@epfl.ch).

© 2009 Macmillan Publishers Limited. All rights reserved

Drivers of biodiagnostic development

David A. Giljohann[1] & Chad A. Mirkin[1]

The promise of point-of-care medical diagnostics — tests that can be carried out at the site of patient care — is enormous, bringing the benefits of fast and reliable testing and allowing rapid decisions on the course of treatment to be made. To this end, much innovation is occurring in technologies for use in biodiagnostic tests. Assays based on nanomaterials, for example, are now beginning to make the transition from the laboratory to the clinic. But the potential for such assays to become part of routine medical testing depends on many scientific factors, including sensitivity, selectivity and versatility, as well as technological, financial and policy factors.

Recent technological advances have markedly improved the way in which we study disease and point towards new opportunities for diagnosing disease. Researchers now have tools to observe phenomena at the level of the atom, to sequence entire genomes and to understand the molecular basis of disease. In addition, new materials, especially nanostructures, are providing novel ways of detecting markers of disease at low concentrations, in complex sample media (such as serum) and with a wide variety of assay read-outs. But many of the latest innovations are not yet being used in routine diagnostic testing, especially when point-of-care issues are considerable, for example when the cost of deploying an assay and training staff at the point of care is high. As biodiagnostic applications based on these new materials continue to be developed, it will be important to be conscious of the key factors that drive this process so that new tests are more likely to reach the clinic. In this Perspective, we assess the factors of assay sensitivity, selectivity and versatility, and robustness, cost and portability. We also discuss some of the materials that are allowing new assays to be designed and the consequences of developing such technologies.

Sensitivity

The diagnosis of a disease on the basis of the presence or concentration of certain biomolecules requires assays that can detect molecules of interest (or targets) sensitively. In this post-genomic era, the targets are most commonly nucleic acids or proteins. Researchers have developed two general strategies to achieve high sensitivity: target-based amplification and signal-based amplification. In target-based amplification, a recognition event triggers a catalytic process that generates more of the target being recognized or surrogates for this target. The polymerase chain reaction (PCR) is a classic example of target amplification, and modern PCR techniques can reliably detect the presence of just a few copies of a nucleic acid sequence[1]. By contrast, in signal-based amplification, a catalytic entity is often used to increase the signal that results from a single binding event. A typical example is the enzyme-linked immunosorbent assay (ELISA)[2], in which a target protein can be captured by an antibody and then sandwiched with a second antibody that incorporates (or is associated with) a catalytic, signal-generating entity. Certain techniques that do not involve amplification, for example single-molecule spectroscopy techniques, might seem to be sensitive; however, these types of spectroscopy typically require greater than nanomolar concentrations of the molecule to be present in order to find and probe it. Therefore, such approaches are not typically viewed as high-sensitivity methods in the context of medical diagnostics.

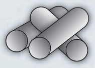
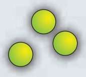
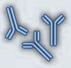
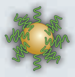
Target-based amplification is a more sensitive strategy than signal-based amplification and is generally considered to be a superior approach, because generating an exponential increase in target concentration leads to faster assay kinetics and pushes the thermodynamics of the probe–target capture reaction in favour of bound (detectable) target. In the short term, it seems that PCR will continue to be a benchmark for nucleic acid detection. But the instability and variability that are inherent in enzymatic processes limit its application outside an institutional setting, such as a research facility or a large central clinical lab. Another drawback to PCR is that lengthy optimization procedures are often required if several targets are to be amplified and detected at the same time, a process known as multiplexing, which is a desirable feature in the clinic, especially as panel assays (which test many disease markers simultaneously) grow in importance for diagnosing disease.

In the past decade, new materials and assays have been developed for signal-based amplification and detection, and assay sensitivities (Table 1) are now approaching those of target-based amplification. Many of these advances rely on nanoscale materials, which have attractive properties for such assays: they have unique and controllable size-dependent properties, have tunable chemical compositions, and in certain cases are chemically and physically robust structures[3,4]. The tailorable properties of nanomaterials, including their high surface-to-volume ratios, mean that target-binding events are often more easily transduced into detectable signals. An example is polyvalent nanoparticles that consist of gold particles modified with biomolecules; these can be used as diagnostic probes[5]. In one assay system, when the target binds to the biomolecules, the associated gold particles catalyse the reduction of silver, leading to a marked enhancement of signal, which can be read with a device that measures light scattered from the developed silver spots. This 'scanometric' strategy allows the detection of attomolar (10^{-18}) concentrations of nucleic acids and proteins in complex biological samples[6]. Indeed, the properties of DNA-functionalized gold nanoparticles (including their optical, catalytic and binding properties) have been taken advantage of in a variety of detection methodologies, including colorimetric[7], fluorescent[8–10], chemiluminescent[11], scanometric[5], surface-plasmon-resonance-based[12] and Raman-spectroscopy-based[13] strategies. For protein detection, certain assays using protein-modified gold nanoparticles have detection limits that are many orders of magnitude lower than those possible with a conventional technique, such as an ELISA[6]. Detection systems based on other nanomaterials have also been evaluated. For example, biomolecule-modified carbon nanotubes[14,15] and silicon nanowires[16] are systems in which changes in electrical conductance on target binding can be translated into a spectroscopic or electrical signal. These strategies take advantage of properties of the nanomaterial, for example their conductance, to cause a measurable change in electrical signal and have resulted in detection methods of moderately high sensitivity[17] (Table 1), which are likely to improve after further refinement.

[1] Department of Chemistry and International Institute for Nanotechnology, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA.

© 2009 Macmillan Publishers Limited. All rights reserved

Table 1 | Detection of protein biomolecules

Tool or technique	Read-out	Illustration	Detection limit*	Molecules per drop (60 μl)	In clinical use
Colorimetry	Visual		150 pM	10^9	Yes
Carbon nanotubes	Electrical		100 pM	10^9	No
Chemi-luminescence	Luminescence		30 pM	10^9	Yes
ELISA	Luminescence		1–10 pM	10^7	Yes
Quantum dots	Fluorescence		500 fM	10^7	No
Silicon nanowires	Electrical		1 fM	10^4	No
Metal nano-particles (bio-barcode)	Scanometric and/or light scattering		30 aM	900	Yes
Immuno-PCR	Fluorescence		20 aM	700	Yes

*Detection limits are best-case examples from the literature and can vary substantially depending on the target and the assay conditions.

The ability to detect targets more sensitively brings its own set of challenges. Consider the example of Alzheimer's disease. At present, markers for Alzheimer's disease, such as amyloid-β-derived diffusible ligands[18], are being identified through histological studies of brain tissue. However, a definitive diagnosis based on a brain biopsy is not a viable diagnostic option. Probes that allow imaging of the brain have been developed and might lead to powerful new biodiagnostic tools, but substantial additional work is needed to identify suitable markers and to correlate their presence and concentration with the state of the disease[19]. Importantly, some *in vitro* biodiagnostic tools based on new nanomaterials that have higher sensitivities than ELISAs might allow such markers to be identified in locations outside the brain, for example in the cerebrospinal fluid or blood, where the marker concentration is significantly lower[20] (Fig. 1). But the detection of new biomarkers is just the first step towards a new treatment protocol. New biomarkers must be validated, and this process requires intensive prospective and retrospective correlative research. These studies are costly and laborious, and researchers who discover new biomarkers are often ill-equipped to test the utility of these biomarkers in a clinical setting. Indeed, even for biomarkers that have been validated in the laboratory, the process of approval by regulatory bodies such as the US Food and Drug Administration (FDA) can be long and complicated, and cannot keep pace with scientific discovery. Although these challenges present considerable barriers to the development and implementation of new biomarkers, the pay-off for doctors and patients will be enormous when it becomes possible to diagnose diseases that are not detectable with conventional biodiagnostic tools as a result of lack of sensitivity.

When it becomes possible to detect a target at extremely low concentrations, barriers to sensitivity are no longer the motivating force behind the research and are superseded by more practical concerns about how the technology can be applied. Therefore, when researchers propose a new biomarker or reach a new detection limit, it is important for the medical community to determine the clinical utility of the biomarker. Important points to consider are the meaningful concentrations of a given biomarker, and how disease states and disease cure would be defined in the context of such previously undetectable biomarkers. In addition, it should be considered whether greater sensitivity allows biomarker detection in samples that are more distant from the source of biomarker production (such as blood, breath, urine and sweat) and what concentrations are clinically relevant (Fig. 1). Clearly, analytical benchmarks, which give clinical meaning to test results, will constantly need to be evaluated and redefined when this stage of the development process is reached.

An example from research in our laboratory highlights this challenge. We recently developed an ultrahigh-sensitivity biodiagnostic system called the bio-barcode assay. This assay involves the use of gold nanoparticle probes 'decorated' with DNA (the barcode DNA) that is end-functionalized with antibodies specific for a target of interest. The target in the assay is sandwiched between a gold particle probe and a magnetic particle probe. Isolation of the sandwich complex in a magnetic field, followed by chemical release of the barcode DNA, results in an amplification event (each protein target is 'traded' for hundreds of strands of barcode DNA). The barcode DNA is then identified using a scanometric assay and the Verigene System (Nanosphere, Northbrook, Illinois), a high-sensitivity, nanotechnology-enabled biodiagnostic platform for nucleic acid detection that is commercially available and has been cleared by the FDA. Using this assay, we have achieved unparalleled levels of sensitivity for detecting a variety of targets. Indeed, we have reached the point at which new definitions of disease states may be warranted as a result of the ability to detect previously unmeasurable concentrations of antigen.

At present, for example, the clinical limit of detection of prostate-specific antigen (PSA), which is often found in increased concentrations in the serum of men with prostate cancer, is $0.1\ \text{ng mL}^{-1}$, using a conventional (ELISA) immunoassay. In men who have undergone radical prostatectomy (removal of the prostate gland), serum PSA concentrations are still monitored, because PSA is considered a valid marker for disease recurrence, but the main source of PSA has been removed. So it is often not possible to detect PSA in the serum of these men, because its concentration falls below $0.1\ \text{ng mL}^{-1}$. These patients are therefore assigned the status 'undetectable', and disease recurrence is monitored in terms of PSA serum concentration increases above the $0.1\ \text{ng mL}^{-1}$ cutoff. With the development of the bio-barcode assay, it was concluded that 'undetectable' PSA is a limitation of current technologies: using the bio-barcode assay, PSA is measurable in the serum of nearly all patients who have had a prostatectomy, a finding that might lead to a re-evaluation of the clinical definitions of prostate cancer cure and recurrence[21]. Accordingly, this redefinition could change the ways in which patients are monitored and treated after radical prostatectomy, and these findings present a challenge to the medical community to make appropriate use of the new technique for the benefit of their patients. In this case, many important questions remain. Can disease cure be defined as a patient having a consistent PSA concentration after prostatectomy? Can changes be made to the schedule of post-operative follow-up visits if a patient has a favourable and consistent PSA concentration? What increase in PSA concentration constitutes clinically significant disease recurrence, and how should such a patient be treated? Can new clinical trials be developed such that increasing PSA concentrations (below the conventional detection limit) are used to initiate and validate new therapy and, importantly, track an individual's response to treatment?

This application of the bio-barcode assay highlights an interesting situation, in which it is possible to detect a disease marker at a much lower concentration than has traditionally been associated with a 'disease-positive' patient. The medical community now has a key role in further developing this assay and any related assays that might lead to changes in the definition of disease state, that might result in detection at an earlier stage, and that might improve the outcome for patients. Certainly, if doctors are to overcome the hurdles to the adoption of this technique and gain the advantage of more sensitive tests, the increase in sensitivity must provide meaningful information about their patients. Sceptics might ask, if the technologies for more sensitive assays are developed but no one knows what to do with the information, then why develop the tests? The answer is that enhancing sensitivity and changing analytical benchmarks are the first steps towards

© 2009 Macmillan Publishers Limited. All rights reserved

removing 'blind spots' in the ability to study, diagnose and treat disease, as well as in the ability to validate new therapeutics for diseases.

Selectivity and versatility

Not only do targets need to be detectable at low concentrations but, as a first step, they need to be identified specifically from among the numerous biomolecules present in each sample. Thus, one of the important drivers when developing new biodiagnostics is the need to recognize targets selectively, and for the next generation of biodiagnostics it will be crucial to develop methods that are versatile enough to be applicable to multiple diseases and disease states to allow their use in personalized diagnosis and medicine.

PCR is a prime example of a selective method that can be used to detect disease-specific targets present at a low copy number. PCR uses Watson–Crick base pairing to provide the selectivity needed for a nucleic acid probe to bind to a genomic DNA target and allow its subsequent amplification. PCR is also an excellent example of a technique that started as a research tool but was eventually translated into medical applications. Indeed, the versatility of PCR has allowed its use not only in clinical diagnostics but also in areas as diverse as forensics and archaeology.

Although PCR is highly selective, it lacks the versatility needed in medical diagnosis. It is clear that non-genetic targets, such as proteins, small molecules and ions, are also important biological indicators of disease. Recently, techniques such as immuno-PCR[22] and the bio-barcode assay have been developed, and these allow proteins to be detected selectively with much greater sensitivity than in the past. Immuno-PCR, which involves the coupling of nucleic acids to antibodies and then PCR-based amplification of the nucleic acid label, provides a limit of detection comparable to the bio-barcode assay, although it is hampered by the limitations of PCR, which requires specialized assay conditions. In addition, both assays, when used for complex sample media, are limited by the selectivity of the antibodies used to capture the target. New materials can play a prominent part, at least in reducing assay complexity and in increasing the versatility of assay read-out. For example, nanoshells or noble-metal nanoparticles[23] can be modified with agents that recognize biomolecules and used to detect proteins, elemental ions[24,25] and small molecules[26]. Assays based on such materials rely on non-enzymatic amplification of the signal with colorimetric or spectroscopic outputs. These strategies are adaptable enough to be used for recognizing many types of analyte, and as mentioned earlier the properties of the nanostructures, including their size and shape, can be manipulated, providing the needed versatility.

In the case of proteins, the issue of selectivity is considerably more challenging than for nucleic acids. Indeed, despite advances, there are many challenges still associated with identifying and making antibodies with the desired properties for binding to a target of interest. In addition, these structures are more fragile than their DNA counterparts and require special handling during storage and use. These problems are exacerbated when using antibody-labelled probes in the context of multiplexing assays. When antibodies are adsorbed on surfaces and used to assay complex sample media, scientists must contend with poor target affinity and selectivity, which can result in significant crossreactivity with targets (that is, an antibody can bind nonspecifically to proteins in the sample). In developing next-generation biodiagnostic systems, an important step in increasing assay selectivity and versatility will be to replace antibody-based systems with chemical systems that do not have the current limitations of antibodies but have the same or greater selectivity. For example, short oligonucleotides with high selectivity and affinity, termed aptamers, provide an alternative route to highly selective recognition of targets, including small molecules and proteins[27]. Through multiple rounds of *in vitro* selection, aptamers can be chosen for their ability to bind to a particular target, resulting in high target selectivity[28]. With further development, aptamers may rival antibodies in their ability to recognize targets, including proteins and small molecules. Importantly, with nucleic acids (as opposed to antibodies) as a basis for recognition, it is conceptually simpler for researchers to modify sequences and structures chemically, making them potentially suitable for signal-based amplification technologies. Finding such new robust target-recognition agents with strong binding and amplification capacities (especially for protein targets) will continue to motivate the development of biodiagnostics.

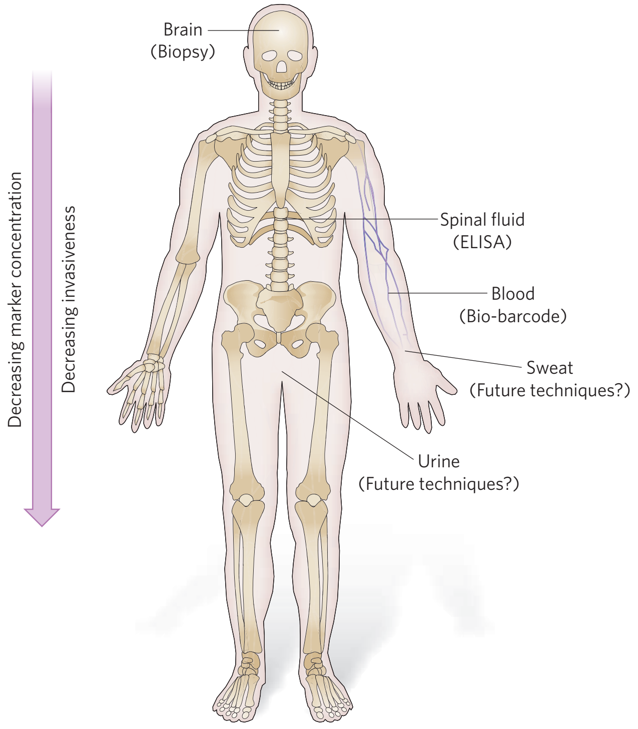


Figure 1 | High-sensitivity detection can allow less invasive disease diagnosis: Alzheimer's disease example. New technologies with higher sensitivities (lower limits of detection) allow markers to be detected at locations distant from the brain, for example in the cerebrospinal fluid, the blood, or even the urine or sweat. In this example, relevant technologies are indicated to illustrate the concept. Depending on the disease marker and the disease, the relevance of a given diagnostic tool will change.

Robustness and portability

At present, the medical diagnostics industry is highly centralized. Diagnostic testing is based mainly in laboratories that are remote from the site of patient care and are equipped with large and complex instruments that typically require highly skilled personnel to operate them. Therefore, a major challenge in the diagnostics field is the development of robust, portable and low-cost assays that will allow disease markers to be detected reliably in places as diverse as the battlefield, the developing world, community hospitals, the doctor's office, the post office or perhaps even at home. For example, a home test may require the use of a dipstick assay to test urine, whereas in a hospital a more complex assay to test blood may be warranted. Although robustness, portability and affordability are necessary features of any technological development, they are not always considered at the initial stages of assay conception and proof-of-concept demonstration. Yet inability to meet these criteria can cause some technologies to hit a dead end in their application, while other technologies accelerate past.

High-density microarrays are a good example of a new technology that generates a large amount of information per unit of time, information that can aid in disease diagnosis[29]. Conceptually, a microarray provides a plethora of genetic information and meets the requirements for gathering data from complex sample media. A natural question to ask is what has impeded the translation of high-density microarrays to the clinic. And, although it may seem inane, why cannot such a microarray be picked up at a local pharmacy? Ultimately, the limited applicability of microarrays lies in their high cost and lack of robustness, as well as in the difficulty in interpreting and acting on the information generated from them. In this case, the ability to measure targets at the molecular level in the laboratory has outpaced the ability to translate the data into useful information for

© 2009 Macmillan Publishers Limited. All rights reserved

the patient. Consequently, the modern field of diagnostic medicine has moved towards using lower-density structures, which allow the detection of a more manageable number of targets and provide a predictive outcome for the user.

In developing technologies that are robust, portability must be considered. Technology cannot be physically bulky, expensive or cumbersome in other ways, or else diagnostic testing will remain mostly in centralized laboratories and in the hands of research scientists. As an example, issues of cost, safety and field use stymied the development of radioactivity-based assays, leading to a switch to fluorescence-based techniques when this technology became available. However, even fluorescence-based assays have drawbacks: photobleaching (fading of the fluorophore), ease of read-out and cost have presented challenges to achieving widespread point-of-care use. Recently, the use of nanomaterials such as quantum dots has overcome some of these problems: these materials have tunable emission profiles that yield bright signals in the visible to near-infrared spectrum and are less susceptible to photobleaching than their molecular counterparts[30,31]. Although quantum dots have already shown their utility in a variety of laboratory detection applications, other advantages (apart from a reduction in photobleaching) must be identified, and methods for making these nanomaterials in an environmentally acceptable manner will be required if these materials are to be translated into widespread, commercially viable medical diagnostic applications. Efforts in this area are under way with silicon quantum dots[32]. Furthermore, structures such as nanowires that can sense analyte binding (for example through changes in electrical conductivity) are being developed and will provide a major step towards new detection assays, as electrical detection systems are conceptually more portable and robust than many optical-based systems[33]. Although these technologies based on nanomaterials are still in the early stages, they are showing promise in the laboratory and, with further design and development, could have a major impact on medical diagnostics.

Looking forward

The established techniques of PCR and ELISA will eventually give way to methods based on new technologies that offer comparable amplification, greater multiplexing capabilities and the prospects of lower cost and portability. New materials — in particular noble-metal nanoparticles, quantum dots, carbon nanotubes and silicon nanowires — are now redefining analytical benchmarks for sensitivity, selectivity and versatility. When proteins and other non-nucleic-acid analytes can be detected as robustly as nucleic acids and with a technology as sensitive as PCR, the field of molecular diagnostics and medicine will be revolutionized.

When developing a new biodiagnostic test, however, there are many stakeholders, including researchers, industry, doctors, regulatory bodies and patients. Each group faces challenges with regard to both the invention and adoption of new technologies that are sensitive, selective, robust and portable. From the perspective of the researcher, the process for taking a scientific result or a breakthrough in materials development and moving it into the clinic is not straightforward. In addition, the time frame is long, and the process is expensive and not one that a conventional research laboratory is equipped to complete. This is the stage at which business enterprises have the financial and intellectual capital available to take promising technologies from the research laboratory into the clinic. However, not all promising technologies are marketable. From the perspective of the medical community, novel technologies can be a welcome addition to the doctor's office. Nevertheless, change comes at a high cost. In addition to direct financial expenses, burdens include the time, infrastructure, regulatory issues, reimbursement considerations and experience needed to diagnose a disease on the basis of the latest technology. The disproportionate cost of adopting a technology relative to the immediate benefits to doctors and patients can prove high barriers to implementation. Regulatory agencies must find a way to keep pace with the rate of innovation and must provide incentives for generating new tests and validating disease markers. Finally, getting a technology into the hands of the patient (in a home test kit) comes with its own set of challenges, in which ease of use, cost and portability are key issues.

With such a diverse group of stakeholders with varied motivations, all of the drivers for the development of biodiagnostic assays need to be considered at each stage of development. Full understanding of these considerations is essential not only for gaining a fundamental scientific understanding of disease but also for designing and implementing innovative disease detection and treatment strategies. We predict that it will also yield the most successful biodiagnostic technologies in the future. ■

1. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6,** 986–994 (1996).
2. Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry* **8,** 871–874 (1971).
3. Rosi, N. L. & Mirkin, C. A. Nanostructures in biodiagnostics. *Chem. Rev.* **105,** 1547–1562 (2005).
This review highlights how nanomaterials contribute to diagnostics.
4. Alivisatos, P. The use of nanocrystals in biological detection. *Nature Biotechnol.* **22,** 47–52 (2004).
5. Taton, T. A., Mirkin, C. A. & Letsinger, R. L. Scanometric DNA array detection with nanoparticle probes. *Science* **289,** 1757–1760 (2000).
6. Nam, J. M., Thaxton, C. S. & Mirkin, C. A. Nanoparticle-based bio-bar codes for the ultrasensitive detection of proteins. *Science* **301,** 1884–1886 (2003).
7. Elghanian, R., Storhoff, J. J., Mucic, R. C., Letsinger, R. L. & Mirkin, C. A. Selective colorimetric detection of polynucleotides based on the distance-dependent optical properties of gold nanoparticles. *Science* **277,** 1078–1081 (1997).
8. You, C.-C. *et al.* Detection and identification of proteins using nanoparticle–fluorescent polymer 'chemical nose' sensors. *Nature Nanotechnol.* **2,** 318–323 (2007).
9. Dubertret, B., Calame, M. & Libchaber, A. J. Single-mismatch detection using gold-quenched fluorescent oligonucleotides. *Nature Biotechnol.* **19,** 365–370 (2001).
10. Seferos, D. S., Giljohann, D. A., Hill, H. D., Prigodich, A. E. & Mirkin, C. A. Nano-flares: probes for transfection and mRNA detection in living cells. *J. Am. Chem. Soc.* **129,** 15477–15479 (2007).
11. Wang, Z., Hu, J., Jin, Y., Yao, X. & Li, J. *In situ* amplified chemiluminescent detection of DNA and immunoassay of IgG using special-shaped gold nanoparticles as label. *Clin. Chem.* **52,** 1958–1961 (2006).
12. He, L. *et al.* Colloidal Au-enhanced surface plasmon resonance for ultrasensitive detection of DNA hybridization. *J. Am. Chem. Soc.* **122,** 9071–9077 (2000).
13. Cao, Y. C., Jin, R. & Mirkin, C. A. Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection. *Science* **297,** 1536–1540 (2002).
14. Wang, J., Liu, G. & Jan, M. R. Ultrasensitive electrical biosensing of proteins and DNA: carbon-nanotube derived amplification of the recognition and transduction events. *J. Am. Chem. Soc.* **126,** 3010–3011 (2004).
15. Chen, R. J. *et al.* Noncovalent functionalization of carbon nanotubes for highly specific electronic biosensors. *Proc. Natl Acad. Sci. USA* **100,** 4984–4989 (2003).
16. Cui, Y., Wei, Q., Park, H. & Lieber, C. M. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science* **293,** 1289–1292 (2001).
17. Soman, C. P. & Giorgio, T. D. Quantum dot self-assembly for protein detection with sub-picomolar sensitivity. *Langmuir* **24,** 4399–4404 (2008).
18. Catalano, S. M. *et al.* The role of amyloid-β derived diffusible ligands (ADDLs) in Alzheimer's disease. *Curr. Top. Med. Chem.* **6,** 597–608 (2006).
19. Perrin, R., Fagan, A. & Holtzman, D. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* **461,** 916–922 (2009).
20. Georganopoulou, D. G. *et al.* Nanoparticle-based detection in cerebral spinal fluid of a soluble pathogenic biomarker for Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **102,** 2273–2276 (2005).
21. Thaxton, C. S. *et al.* Nanoparticle-based bio-barcode assay redefines 'undetectable' PSA and biochemical recurrence after radical prostatectomy. *Proc. Natl Acad. Sci. USA* **106,** 18437–18442 (2009).
22. Sano, T., Smith, C. L. & Cantor, C. R. Immuno-PCR: very sensitive antigen detection by means of specific antibody–DNA conjugates. *Science* **258,** 120–122 (1992).
23. McFarland, A. D. & Van Duyne, R. P. Single silver nanoparticles as real-time optical sensors with zeptomole sensitivity. *Nano Lett.* **3,** 1057–1062 (2003).
24. Liu, J. & Lu, Y. A colorimetric lead biosensor using DNAzyme-directed assembly of gold nanoparticles. *J. Am. Chem. Soc.* **125,** 6642–6643 (2003).
25. Lee, J.-S., Han, M. S. & Mirkin, C. A. Colorimetric detection of mercuric ion in aqueous media using DNA-functionalized gold nanoparticles. *Angew. Chem. Int. Edn Engl.* **46,** 4093–4096 (2007).
26. Hirsch, L. R., Jackson, J. B., Lee, A., Halas, N. J. & West, J. L. A whole blood immunoassay using gold nanoshells. *Anal. Chem.* **75,** 2377–2381 (2003).
27. Jayasena, S. D. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. *Clin. Chem.* **45,** 1628–1650 (1999).
28. Ellington, A. D. & Szostak, J. W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).
This paper describes the identification of RNA aptamers.
29. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
This study demonstrated the use of microarray analysis for profiling gene expression patterns.
30. Chan, W. C. & Nie, S. Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science* **281,** 2016–2018 (1998).
31. Bruchez, M. Jr, Moronne, M., Gin, P., Weiss, S. & Alivisatos, A. P. Semiconductor nanocrystals as fluorescent biological labels. *Science* **281,** 2013–2016 (1998).
32. Park, J. H. *et al.* Biodegradable luminescent porous silicon nanoparticles for *in vivo* applications. *Nature Mater.* **8,** 331–336 (2009).
33. Zheng, G., Patolsky, F., Cui, Y., Wang, W. U. & Lieber, C. M. Multiplexed electrical detection of cancer markers with nanowire sensor arrays. *Nature Biotechnol.* **23,** 1294–1301 (2005).

Acknowledgements We thank P. Patel for assistance with preparing Table 1. We also thank the US National Science Foundation, the National Cancer Institute and the National Institutes of Health for research support.

Author Information Reprints and permission information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to C.A.M. (chadnano@northwestern.edu).

© 2009 Macmillan Publishers Limited. All rights reserved

Structure of the formate transporter FocA reveals a pentameric aquaporin-like channel

Yi Wang[1]*, Yongjian Huang[2]*, Jiawei Wang[2]*, Chao Cheng[2], Weijiao Huang[2], Peilong Lu[1], Ya-Nan Xu[3], Pengye Wang[3], Nieng Yan[2] & Yigong Shi[1]

FocA is a representative member of the formate–nitrite transporter family, which transports short-chain acids in bacteria, archaea, fungi, algae and parasites. The structure and transport mechanism of the formate–nitrite transporter family remain unknown. Here we report the crystal structure of *Escherichia coli* FocA at 2.25 Å resolution. FocA forms a symmetric pentamer, with each protomer consisting of six transmembrane segments. Despite a lack of sequence homology, the overall structure of the FocA protomer closely resembles that of aquaporin and strongly argues that FocA is a channel, rather than a transporter. Structural analysis identifies potentially important channel residues, defines the channel path and reveals two constriction sites. Unlike aquaporin, FocA is impermeable to water but allows the passage of formate. A structural and biochemical investigation provides mechanistic insights into the channel activity of FocA.

Formate is a major carbon source for methanogenic archaea such as *Methanobacterium formicicum*[1,2]. It is also a signature metabolite of enteric bacteria under anaerobic conditions, during which pyruvate is cleaved by pyruvate–formate lyase (PFL) to yield acetyl CoA and formate[3,4]. As much as one-third of the carbon in the sugar was thought to be converted to formate during fermentative growth[3,4], reaching a concentration of up to 20 mM in the cytoplasm[5]. Intracellular accumulation of formate may lead to a substantial decrease in cytoplasmic pH. In *Escherichia coli*, formate is metabolized by three formate dehydrogenases (FDHs)[4–7], of which FDH-N and FDH-O have their active sites located in the periplasm and FDH-H in the cytoplasm as part of a multiprotein complex named formate hydrogenlyase (FHL). Formate must therefore be able to pass through the cell membrane. However, with a pK_a of 3.75, formate exists predominantly in the deprotonated anionic form at physiological pH, necessitating a transport system. The integral membrane protein FocA was identified as a putative formate transporter in *E. coli*[7].

FocA is a representative member of the formate–nitrite transporter (FNT) family (transporter classification 2.A.44), which was thought to transport structurally similar short-chain acids such as formate and nitrite[8]. Other known members of the FNT family include NirC of *E. coli* for nitrite uptake and export[9–12], and FdhC of *M. thermoformicicum* for formate uptake[8,13]. FocA homologues have been identified in bacteria, archaea, fungi, algae and parasites (Supplementary Fig. 1). Although members of the FNT family share considerable sequence homology, they have no apparent sequence similarity to other proteins. The structure and transport mechanism of the FNT family remain largely unknown.

We purified the full-length, recombinant FocA protein from *E. coli* and generated crystals under several conditions. However, these crystals persistently diffracted X-rays to low resolutions. We subjected the full-length FocA to V8 protease digestion, which removed 21 amino-acid residues from the amino terminus. The truncated FocA was crystallized in two space groups, $P2_12_12_1$ and $P3_2$, each with an improved diffraction limit. The structure in $P2_12_12_1$ was determined by platinum-based single-wavelength anomalous dispersion (SAD) (Supplementary Table 1). The experimental electron density was of adequate quality (Supplementary Fig. 2a), and the final atomic model was refined to a free R-factor of 0.22 at 2.25 Å resolution (Supplementary Table 1 and Supplementary Fig. 2b–d). We also solved the structure of FocA in the $P3_2$ space group at 3.2 Å resolution (Supplementary Table 1 and Supplementary Fig. 3a).

Overall structure

Each asymmetric unit in $P2_12_12_1$ contains five molecules of FocA, arranged as a symmetric homopentamer through a five-fold axis perpendicular to the plane of the lipid membrane (Fig. 1a). The five FocA protomers associate with each other through extensive interactions, resulting in the burial of 15,800 Å^2 of surface area. This analysis suggests that FocA may exist as a pentamer in lipid membrane. Supporting this notion is the observation that there are ten molecules of FocA in an asymmetric unit of the $P3_2$ space group, organized into two homopentamers (Supplementary Fig. 3a). These homopentamers are nearly identical to each other and to that in the $P2_12_12_1$ space group (Supplementary Fig. 3b), with a pairwise root mean squared deviation (r.m.s.d.) of about 0.7 Å. For simplicity, we limit our discussion to the FocA pentamer in the $P2_12_12_1$ space group.

The five protomers of FocA form a short cylinder, with a diameter of about 80 Å and a height of about 55 Å (Fig. 1b). The cylindrical outer surface is hydrophobic. By contrast, the periplasmic and cytoplasmic faces of the FocA cylinder have negative and positive electrostatic potentials, respectively (Fig. 1b). The charged amino acids are located mainly at the periphery of the FocA cylinder, which seems to correlate with the lipid composition of the membrane[14]. Each FocA protomer contains an axial passage that is roughly perpendicular to the plane of the lipid membrane (Fig. 1a).

[1]Ministry of Education Protein Science Laboratory, [2]State Key Laboratory of Biomembrane and Membrane Biotechnology, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [3]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

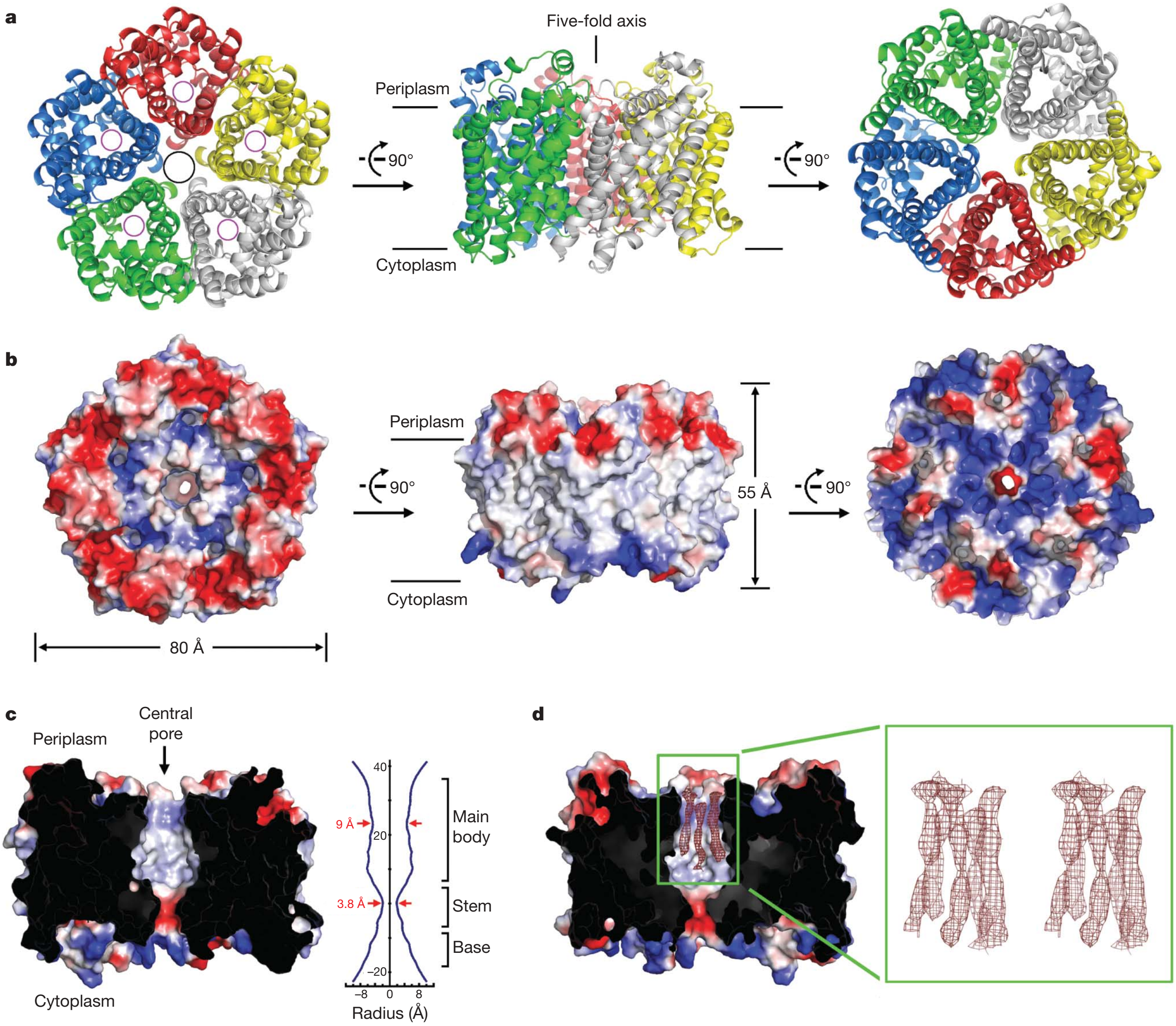


Figure 1 | The overall structure of FocA. **a**, Ribbon representation of the FocA pentamer, shown in three perpendicular views. Note the pore in the centre of the pentameric assembly (black circle) and the axial passage in each of the five protomers (magenta circles). **b**, Surface electrostatic potential of the FocA pentamer. The three views shown here correspond to those in **a**. The periplasmic and cytoplasmic sides of FocA are negatively and positively charged, respectively. **c**, The FocA pentamer contains a central pore. A cut-through section of the FocA pentamer is shown to indicate the shape and surface features of the pore. The calculated pore diameters are shown at the right. **d**, The central pore is filled with five strings of linear electron density that are characteristic of the hydrophobic tails of detergent molecules. The $2F_o - F_c$ electron density is contoured at 2σ and shown in stereo. All structural figures were prepared with PyMol[36].

Intriguingly, the pentameric assembly of FocA also contains a central pore, whose shape resembles that of a wineglass (Fig. 1c). The main body of this pore is exclusively hydrophobic, whereas the stem is acidic and the base on the cytoplasmic side is positively charged (Fig. 1c). The narrowest point in the stem has a diameter of 3.8 Å as calculated by HOLE[15]. The hydrophobic main body is filled with five strings of linear electron density that is characteristic of the hydrophobic tails of detergent molecules (Fig. 1d). This structural feature suggests that the central pore is likely to be occupied by lipid molecules within the plasma membrane of *E. coli*. However, as suggested by others[16], we cannot rule out the possibility that the central pore may serve as a channel of some defined function.

Structure of the FocA protomer

The five FocA protomers have identical structure, with a pairwise r.m.s.d. of 0.19–0.34 Å. As previously predicted[7], each FocA protomer consists of six transmembrane segments (TMs) (Fig. 2a). The amino and carboxy termini of FocA are both placed on the cytoplasmic side, which is consistent with the predicted topology of FocA[7]. TM1, TM3, TM4 and TM6 each consist of a single α-helix (Supplementary Fig. 4). TM2 and TM5, in contrast, consist of two α-helices connected by an extended loop, which is highly conserved among the FNT family members. These two signature loops, placed roughly parallel to the plane of lipid membrane, are located in the axial passage of the FocA protomer. Two additional α-helices, both on the periplasmic side, connect TM3 and TM4 (α3–4) and TM5 and TM6 (α5–6) (Supplementary Fig. 4).

Each FocA protomer contains an internal structural repeat. The N-terminal half of the protomer, TM1–TM3 (residues 29–136), are structurally related to the C-terminal half, TM4–TM6 (residues 160–276), with a quasi-two-fold axis in the plane of the lipid membrane (Supplementary Figs 4 and 5). Despite a low sequence identity of about 8%, these two halves can be superimposed with an r.m.s.d. of 3.3 Å.

In the pentameric FocA cylinder, TM3 and TM6 of each protomer constitute the bulk of the outer layer (Fig. 2b). The central elements

©2009 Macmillan Publishers Limited. All rights reserved

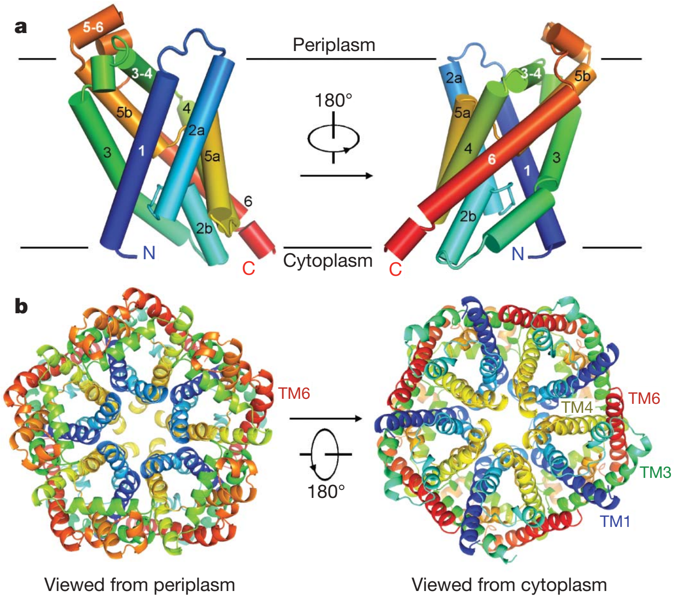


Figure 2 | Structural features of the FocA protomer. a, Ribbon representation of the FocA protomer, shown in two views related by a 180° rotation. FocA is rainbow-coloured, with its N terminus in blue and its C terminus in red. **b**, Role of individual transmembrane segments in the FocA protomer. The FocA pentamer is shown in two views, with the transmembrane segments colour-coded.

of the interface between adjacent protomers are TM1 and TM4, which traverse the full span of the lipid membrane bilayer and radiate from the centre to the outer layer of the FocA cylinder (Fig. 2b). Additional elements of this interface include TM2a and TM5a, as well as portions of TM3 and TM6 (Supplementary Fig. 6). Formation of the inter-protomer interface involves a large number of hydrophobic amino acids that are moderately conserved among the FocA homologues (Supplementary Figs 1 and 6). For example, residues Phe 44/Ile 47/Phe 51 on TM1 and Met 174/Leu 177/Met 181 on TM4, which reside at the centre of the interface, are preserved between 17% and 67% in each of the FocA homologues. The central pore of the FocA pentamer is formed exclusively by amino acids from TM2a and TM5a of the FocA protomers (Supplementary Fig. 7).

Structural mimicry with aquaporins

The pentameric architecture and the extensive packing interactions between adjacent FocA protomers are more indicative of a channel than a transporter. Supporting this notion, structural features of FocA are reminiscent of those of the aquaporin family of proteins (AQPs), which are a family of structurally conserved channels permeable to water or other small organic molecules such as glycerol[17–20]. To examine this situation systematically, we searched for structural homologues of FocA in the Protein Data Bank with DALI[21]. The result was unequivocal: all hits with a *Z* score (similarity score) of 10 or higher are AQPs. The FocA protomer can be superimposed on that of the *E. coli* water channel AqpZ (PDB code 1RC2)[22] and the glycerol channel GlpF (PDB code 1FX8)[16] with r.m.s.ds of 3.2 and 3.3 Å, respectively (Fig. 3a and Supplementary Fig. 8). In both FocA and AQPs, the six transmembrane segments of a protomer comprise two structural repeats: TM1–TM3 and TM4–TM6, with the second transmembrane segment in each repeat (TM2 or TM5) disrupted by a highly conserved loop. FocA is an integral membrane protein with no sequence homology with AQPs but sharing the same fold.

Despite the overall structural similarity, FocA has prominent features that distinguish it from AQPs. First, in contrast to the homotetrameric AQPs, FocA is organized as a homopentamer (Supplementary Fig. 9). In comparison with AQPs, the transmembrane segments within each FocA protomer are arranged at slightly different orientations to accommodate the extra protomer in the assembly. The size of the central pore of the FocA pentamer is larger than that of the AQP tetramer. Second, unlike AQPs, the two internal structural repeats of FocA share little sequence similarity, and neither contains an Asn-Pro-Ala (NPA) motif, which is characteristic of AQPs (Supplementary Fig. 1). Third, and importantly, the two loops that disrupt TM2 and TM5, named L2

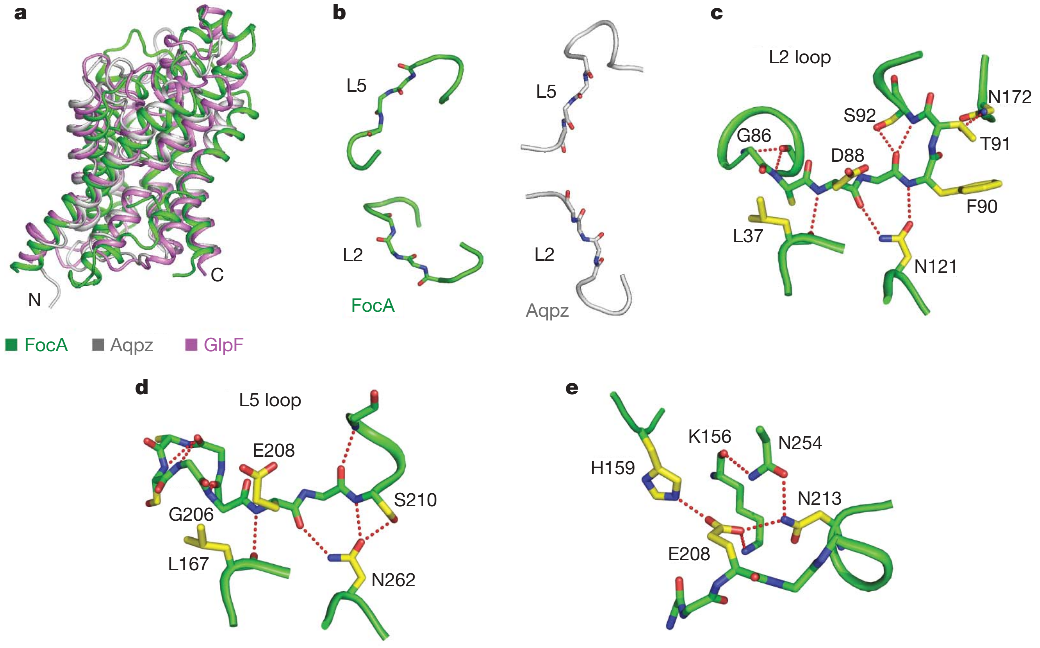


Figure 3 | The FocA protomer is structurally similar to aquaporin. a, Structural overlay of FocA (green), AqpZ[22] (grey) and GlpF[16] (magenta). **b**, The L2 and L5 loops of FocA (green) show different structural features from those in the corresponding loops of AqpZ (grey). Highlighted here are the overall conformation and main-chain carbonyl groups from these loops. **c**, A close-up view of the L2 loop in FocA. Hydrogen bonds are indicated by dashed red lines. **d**, A close-up view of the L5 loop in FocA. Similarly to the L2 loop, the L5 loop is hydrogen-bonded by a conserved asparagine residue, Asn 262. **e**, A network of hydrogen bonds around Glu 208 in the L5 loop. Glu 208 in the L5 loop and Asp 88 in the L2 loop are both highly conserved in FocA homologues.

©2009 Macmillan Publishers Limited. All rights reserved

and L5, respectively, have different configurations from those of AQPs. For example, in AQPs but not in FocA, the backbone carbonyls in these loops are positioned along the same side of the channel and point into the channel passage (Fig. 3b). Consequently, these carbonyl oxygen atoms directly hydrogen-bond to water or glycerol molecules in the channel and presumably facilitate their movement[16,22–25].

In FocA, the central portion of the L2 loop is constrained by three hydrogen bonds (Fig. 3c). The amide nitrogen and carbonyl oxygen of Asp 88 are hydrogen-bonded to the carbonyl oxygen of Leu 37 and the side-chain nitrogen of Asn 121, respectively. The side-chain oxygen atom of Asn 121 accepts an additional hydrogen bond from the amide nitrogen of Phe 90. In addition, Thr 91 at the C-terminal end of the L2 loop donates a hydrogen bond to the side-chain nitrogen atom of Asn 172. The central segment of the L5 loop is similarly confined by hydrogen bonds (Fig. 3d). In this case, the amide nitrogen and carbonyl oxygen of Glu 208 interact with the carbonyl oxygen of Leu 167 and the side-chain nitrogen of Asn 262, respectively. The side-chain oxygen atom of Asn 262 accepts two additional hydrogen bonds from the amide nitrogen and side-chain oxygen atoms of Ser 210. Furthermore, the side chain of Glu 208 organizes a network of hydrogen bonds, with its carboxylate group accepting three bonds from the side chains of Lys 156, Asn 213 and His 159 (Fig. 3e).

Unlike the AQPs, the L2 and L5 loops of FocA are unrelated by primary sequences, yet they share a number of notable structural features. Both loops are constrained by conserved hydrogen bonds, of which two are mediated by a similarly positioned Asn residue: Asn 121 in TM3 and Asn 262 in TM6 (Fig. 3c, d). These two Asn residues are highly conserved among members of the FNT family, suggesting functional significance. A highly conserved acidic residue is positioned in the middle of the loops: Asp 88 in L2 and Glu 208 in L5.

Channel-lining residues

The axial channel of the FocA protomer contains a mixture of polar, charged and hydrophobic amino acids, which render the passage amphipathic (Fig. 4a). The hydrophilic side is formed by the L2 and L5 loops, the connecting helix α3–4, TM2b, TM4 and TM5b (Fig. 4b). This side comprises nine polar or charged residues and several main-chain groups. The other side, formed by TM1, TM2a, TM4 and TM5a, is considerably more hydrophobic, especially in the central portion of the channel. The amphipathic nature of the FocA channel is consistent with the chemical property of its putative substrate: formate or other short-chain acids.

Amino acids that line the channel are highly conserved in FocA homologues across several species, with five invariant residues in the central portion of the channel. The hydrophilic side of the channel contains three invariant residues, Leu 89, Val 175 and His 209, and six conserved amino acids, of which three are polar or charged (Ser 92, Lys 156 and Asn 213). The hydrophobic side comprises two invariant residues, Phe 75 and Phe 202, and seven conserved amino acids, one of which is charged (Lys 68). The conservation of channel-lining residues suggests a shared mechanism for the channel activity of these FocA homologues.

In the periplasmic side of each FocA channel there is an elongated stretch of well-defined electron density (Fig. 4c). The bottom portion of the electron density is surrounded by conserved amino acids Phe 75, Lys 156, Phe 202, His 209 and Asn 213 (Supplementary Fig. 1). Although we could have modelled water or formate molecules into the electron density, we chose not to because we could not differentiate unambiguously between formate and water at the present resolution. In addition, the shape of this electron density remains little changed with or without formate in the crystallization buffer. Nonetheless, a formate or water molecule placed into the bottom portion of the electron density would be well within hydrogen-bond distances of Lys 156, Asn 213 and the carbonyl oxygen of Phe 207.

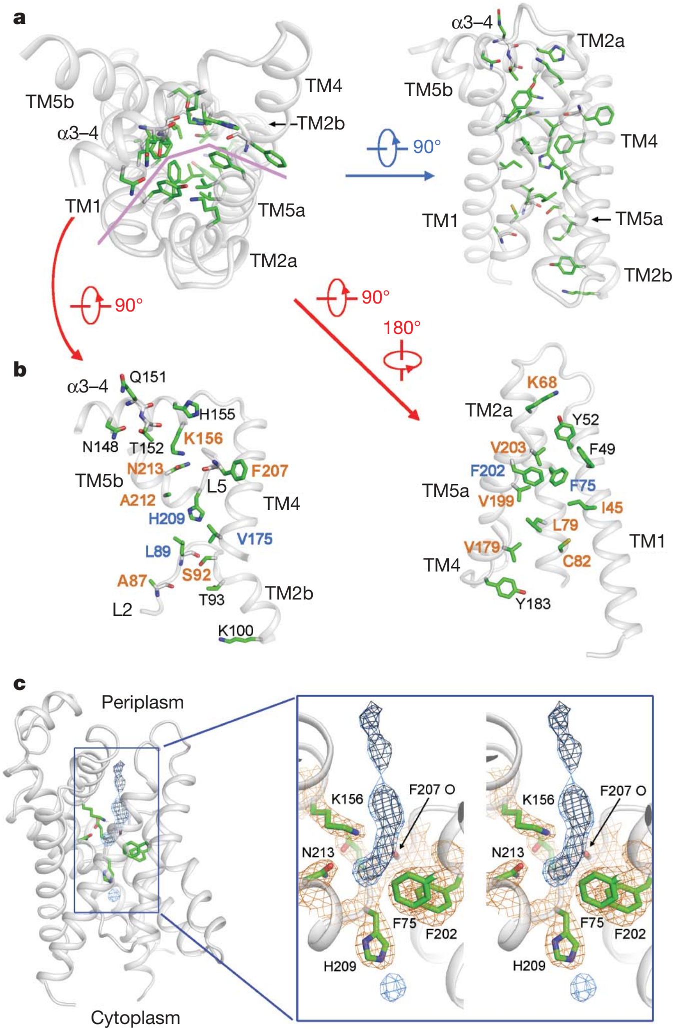


Figure 4 | Features of the axial channel in the FocA protomer. a, The FocA channel is amphipathic. Amino acids along the channel are shown in two perpendicular views. **b,** The FocA channel opened up into two halves to display the amphipathic feature. Amino acids identical in all 12 FocA homologues are shown with blue labels; those conserved in at least 6 FocA homologues are labelled in orange. **c,** The channel on the periplasmic side is occupied by an elongated stretch of electron density. The $2F_o - F_c$ electron density, contoured at 1.5σ, is coloured brown in FocA and black in the middle. The $F_o - F_c$ density, contoured at 3σ, is coloured cyan.

Constriction sites

We calculated the channel diameter along the axial passage with HOLE[15]. Similarly to AQPs, the FocA channel consists of two vestibules, located on the periplasmic and cytoplasmic ends, and a narrow pore in between (Fig. 5a). A central portion of the pore, about 15 Å in length, has a diameter of 3.5 Å or less. There are two constriction sites. One site was formed mainly by the side chains of two invariant, aromatic residues: Phe 75 from TM2a and Phe 202 from TM5a. At this site, the pore diameter is narrowed to about 1.8 Å. The other constriction site is located 7.5 Å to the cytoplasmic side of the first constriction, with an even narrower diameter of 1.35 Å. This constriction was created mainly by the side chains of two highly conserved residues: Leu 79 from TM2a and Leu 89 from the L2 loop. The two pairs of constriction residues, Phe 75/Phe 202 and Leu 79/Leu 89, are positioned diagonally from each other (Fig. 5b). Neither constriction would allow the passage of water molecules, let alone formate or other solutes. Thus we conclude that the observed structure may represent that of FocA in a closed-pore state.

The observed pore diameters of the FocA protomer are generally within the range calculated for AQPs (Fig. 5a). The diameters of the

©2009 Macmillan Publishers Limited. All rights reserved

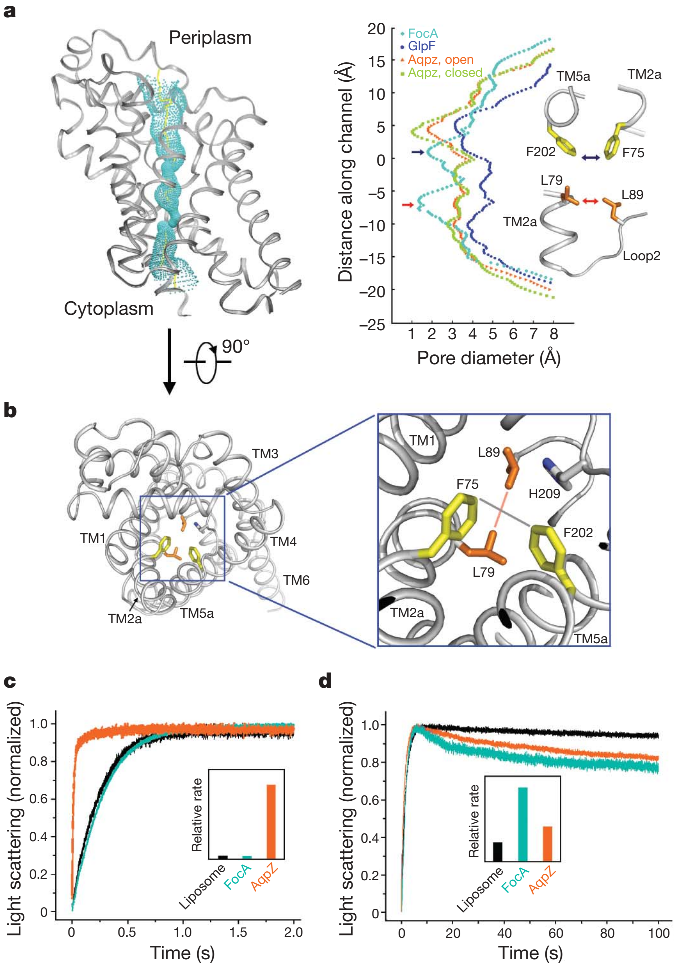


Figure 5 | FocA contains two constriction sites and may exist in a closed-pore state. **a**, The central channel in the FocA protomer contains two constriction sites. The channel passage (left panel), calculated by HOLE[15], is indicated by cyan dots along a central yellow line. The diameters of the channel are tabulated in the right panel and compared with those from GlpF and AqpZ. **b**, The two pairs of amino acid residues that contribute to the two constriction sites, Phe 75/Phe 202 and Leu 79/Leu 89, are roughly diagonal to each other when viewed along the channel axis. **c**, FocA may be impermeable to water. Protein-free liposomes or FocA-loaded proteoliposomes were mixed with 500 mM sucrose in a stopped-flow apparatus. In response to high osmotic pressure, water molecules diffused through the lipid, causing the vesicles to deflate. The rapid changes in vesicle size are reflected by changes of light scattering. **d**, FocA is permeable to formate. The experiments conducted here are similar to those in **c** except that 20 mM sodium formate was used instead of 500 mM sucrose. FocA allowed the passage of formate, as demonstrated by the swelling of the vesicles again. AqpZ also allowed the passage of formate, though more slowly than FocA did. The relative rates of formate conduction were calculated on the basis of a published protocol[16].

two constriction sites of FocA are slightly larger than that of the closed-pore AqpZ[22] but smaller than that of the open-pore AqpZ or GlpF[16,25]. Opening of the pore may require the putative gating residues, Phe 75/Phe 202 and Leu 79/Leu 89, to adopt other rotamer conformations. Gating by a hydrophobic amino acid has previously been observed for the water-selective channel AQP0 (ref. 26), in which the side chain of Met 176 regulates the opening and closure of the water pore.

To examine whether FocA is permeable to water, we generated FocA-loaded proteoliposomes and reconstituted a water permeability assay with the use of a stopped-flow apparatus (Fig. 5c). On mixing with 500 mM sucrose, water molecules diffused through the lipid, causing the lipid vesicles to deflate, as demonstrated by the increasing signal of light scattering. FocA-loaded proteoliposomes allowed a similar rate of water passage to that of the empty liposomes, indicating that FocA may be impermeable to water. By contrast, proteoliposomes loaded with AqpZ allowed the passage of water molecules at a rate about 20-fold faster than that of the empty liposomes or FocA-loaded proteoliposomes.

Next, using a similar experimental setup, we examined whether FocA allowed the passage of formate molecules (Fig. 5d). On mixing with 20 mM formate, both empty liposomes and the FocA-loaded proteoliposomes deflated as a result of a rapid efflux of water. If FocA allows formate to pass through, the proteoliposomes will swell again over time, and the formate conductivity can be calculated by an established protocol[16]. As expected, the empty liposomes allowed rapid water efflux but little formate uptake. By contrast, the proteoliposomes loaded with FocA swelled again after the initial deflation phase, suggesting that FocA is permeable to formate. AqpZ also allowed the passage of formate, although more slowly than FocA.

Perspective

How does FocA mediate the passage of formate? Although a conclusive answer remains to be found, the current structure and analysis provide tantalizing clues. To allow the passage of formate or another solute, the constriction sites of FocA must open. Movement of the aromatic side chains of Phe 75/Phe 202 or Leu 79/Leu 89 may widen the constriction site on the periplasmic or cytoplasmic side. Mutation of these bulky, hydrophobic residues to amino acids with smaller side chains is predicted to enlarge the constriction sites. Consistent with this prediction, the FocA mutants L79V/L89V and F202A both showed a markedly increased capacity for formate passage (Supplementary Fig. 10).

Substrate specificity is determined largely by the selectivity filter in GlpF[16,25] and the ar/R constriction site in AQPs[22,23], both of which comprise hydrophobic and positively charged amino acids (Supplementary Fig. 11). Phe 43, His 174 and Arg 189 constitute the ar/R constriction site in AqpZ. Trp 48, Phe 200 and Arg 206 contribute to the selectivity filter in GlpF, where Arg 206 forms hydrogen bonds to the hydroxyl groups of glycerol and Trp 48/Phe 200 provide a hydrophobic wedge to accommodate the carbon atoms of glycerol. The selectivity filter of GlpF corresponds to the constriction site on the periplasmic side of FocA, with Trp 48/Phe 200 of GlpF replaced by Phe 75/Phe 202 of FocA. Arg 206 is replaced by Ala 212, but its functionality might be substituted for by Lys 156 or Asn 213 (Supplementary Fig. 11). This analysis suggests that these residues in FocA may constitute the selectivity filter. Another notable residue in this region is His 209, which is invariant among all FocA homologues. The imidazole side chain of His 209 is aligned with the channel passage and points to the cytoplasmic side (Fig. 4c). The side-chain rotamer conformation of His 209 is probably important in the channel activity of FocA and its homologues.

We provide strong evidence that the previously classified FNTs may constitute a previously unrecognized class of solute channels. FocA, a representative member of the FNT family, is structurally similar to AQPs, with the same membrane topology and some detailed structural features. FocA and its sequence homologues probably attained their current structure through convergent evolution. In fact, previous biochemical characterization suggested that FocA might function as a bidirectional formate channel[7]. Taken together, these observations suggest that FocA and its homologues may be reclassified, perhaps as a family of formate–nitrite channels (FNCs).

The FNC family members have been found in bacteria, archaea, fungi, algae and parasites, but not in higher eukaryotes[8]. This observation is consistent with our current understanding that, as with several other small molecules such as carbon dioxide and ammonia, formate is likely to have been a key molecule in the early evolution of life on Earth. FocA may therefore represent a member of an ancient and important family of channels that were responsible for the transfer of small molecules across cell membranes. The evolution of channel proteins that facilitate organic acid transfer across the cell membrane without an ATP requirement would be particularly beneficial to

©2009 Macmillan Publishers Limited. All rights reserved

organisms that live under fermentative conditions, in which energy conservation is at a premium. Understanding the functional mechanisms of FocA may reveal important insights into the mechanism of these channel proteins.

METHODS SUMMARY

The recombinant FocA protein was overexpressed in *E. coli* and purified to homogeneity. Proteolysis by V8 protease removed the N-terminal 21 amino acids. The N-terminally truncated FocA was crystallized by the hanging-drop vapour-diffusion method. All data sets were collected at the Spring-8 beamline BL41XU and processed with HKL2000 (ref. 27) and the CCP4 suite[28]. The experimental phase in the $P2_12_12_1$ space group was generated by Pt-SAD with SHELXD[29] and improved by DM[30] and DMMulti[30]. The model was built with BUCCANEER[31] and *Coot*[32]. The final model was refined with PHENIX[33]. The structure of FocA in the $P3_2$ space group was solved by molecular replacement with the program PHASER[34] and refined with PHENIX[33]. The preparation of liposomes and the liposome-based assays were performed as described previously[16,35].

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

Received 16 August; accepted 26 October 2009.

1. Stams, A. J. & Plugge, C. M. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nature Rev. Microbiol.* **7,** 568–577 (2009).
2. White, W. B. & Ferry, J. G. Identification of formate dehydrogenase-specific mRNA species and nucleotide sequence of the *fdhC* gene of *Methanobacterium formicicum*. *J. Bacteriol.* **174,** 4997–5004 (1992).
3. Leonhartsberger, S., Korsa, I. & Bock, A. The molecular biology of formate metabolism in enterobacteria. *J. Mol. Microbiol. Biotechnol.* **4,** 269–276 (2002).
4. Sawers, R. G. Formate and its role in hydrogen production in *Escherichia coli. Biochem. Soc. Trans.* **33,** 42–46 (2005).
5. Sawers, G. The hydrogenases and formate dehydrogenases of *Escherichia coli. Antonie Van Leeuwenhoek* **66,** 57–88 (1994).
6. Stephenson, M. & Stickland, L. H. Hydrogenlyases: bacterial enzymes liberating molecular hydrogen. *Biochem. J.* **26,** 712–724 (1932).
7. Suppmann, B. & Sawers, G. Isolation and characterization of hypophosphite-resistant mutants of *Escherichia coli*: identification of the FocA protein, encoded by the *pfl* operon, as a putative formate transporter. *Mol. Microbiol.* **11,** 965–982 (1994).
8. Saier, M. H. Jr *et al.* Phylogenetic characterization of novel transport protein families revealed by genome analyses. *Biochim. Biophys. Acta* **1422,** 1–56 (1999).
9. Jia, W. & Cole, J. A. Nitrate and nitrite transport in *Escherichia coli. Biochem. Soc. Trans.* **33,** 159–161 (2005).
10. Jia, W., Tovell, N., Clegg, S., Trimmer, M. & Cole, J. A single channel for nitrate uptake, nitrite export and nitrite uptake by *Escherichia coli* NarU and a role for NirC in nitrite export and uptake. *Biochem. J.* **417,** 297–304 (2009).
11. Clegg, S., Yu, F., Griffiths, L. & Cole, J. A. The roles of the polytopic membrane proteins NarK, NarU and NirC in *Escherichia coli* K-12: two nitrate and three nitrite transporters. *Mol. Microbiol.* **44,** 143–155 (2002).
12. Clegg, S. J., Jia, W. & Cole, J. A. Role of the *Escherichia coli* nitrate transport protein, NarU, in survival during severe nutrient starvation and slow growth. *Microbiology* **152,** 2091–2100 (2006).
13. Nolling, J. & Reeve, J. N. Growth- and substrate-dependent transcription of the formate dehydrogenase (*fdhCAB*) operon in *Methanobacterium thermoformicicum* Z-245. *J. Bacteriol.* **179,** 899–908 (1997).
14. von Heijne, G. & Gavel, Y. Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* **174,** 671–678 (1988).
15. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. The pore dimensions of gramicidin A. *Biophys. J.* **65,** 2455–2460 (1993).
16. Fu, D. *et al.* Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **290,** 481–486 (2000).
17. Agre, P. The aquaporin water channels. *Proc. Am. Thorac. Soc.* **3,** 5–13 (2006).
18. Gonen, T. & Walz, T. The structure of aquaporins. *Q. Rev. Biophys.* **39,** 361–396 (2006).
19. Stroud, R. M., Nollert, P. & Miercke, L. The glycerol facilitator GlpF its aquaporin family of channels, and their selectivity. *Adv. Protein Chem.* **63,** 291–316 (2003).
20. Carbrey, J. M. & Agre, P. Discovery of the aquaporins and development of the field. *Handb. Exp. Pharmacol.* **190,** 3–28 (2009).
21. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233,** 123–138 (1993).
22. Savage, D. F., Egea, P. F., Robles-Colmenares, Y., O'Connell, J. D. III & Stroud, R. M. Architecture and selectivity in aquaporins: 2.5 Å X-ray structure of aquaporin Z. *PLoS Biol.* **1,** E72 (2003).
23. Sui, H., Han, B. G., Lee, J. K., Walian, P. & Jap, B. K. Structural basis of water-specific transport through the AQP1 water channel. *Nature* **414,** 872–878 (2001).
24. Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407,** 599–605 (2000).
25. Tajkhorshid, E. *et al.* Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science* **296,** 525–530 (2002).
26. Gonen, T. *et al.* Lipid–protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **438,** 633–638 (2005).
27. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
28. Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
29. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
30. Cowtan, K. dm: an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsl. Protein Crystallogr.* **31,** 34–38 (1994).
31. Cowtan, K. The Buccaneer software for automated model building. *Acta Crystallogr. D* **62,** 1002–1011 (2006).
32. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
33. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
35. Borgnia, M. J., Kozono, D., Calamita, G., Maloney, P. C. & Agre, P. Functional reconstitution and characterization of AqpZ, the *E. coli* water channel protein. *J. Mol. Biol.* **291,** 1169–1179 (1999).
36. DeLano, W. L. PyMOL Molecular Viewer. http://www.pymol.org (2002).

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

Acknowledgements We thank N. Shimizu, S. Baba and T. Kumasaka at the Spring-8 beamline BL41XU for assistance, and J. He and S. Huang for help at Shanghai Synchrotron Radiation Facility (SSRF). This work was supported by funds from the Ministry of Science and Technology of China (grants 2009CB918801 and 2009CB918802), Tsinghua University 985 Phase II funds, Project 30888001 supported by National Natural Science Foundation of China, and the Beijing Municipal Commissions of Education and Science and Technology. N.Y. acknowledges support from the Yuyuan Foundation and Li's Foundation.

Author Contributions Experiments were performed by Y.W., Y.H., J.W., C.C., W.H., P.L., Y.-N.X., P.W. and N.Y. Data were analysed by Y.W., Y.H., J.W., N.Y. and Y.S. The manuscript was prepared by N.Y. and Y.S.

Author Information The atomic coordinates of FocA in the $P2_12_12_1$ and $P3_2$ space groups have been deposited in the Protein Data Bank under accession codes 3KCU and 3KCV, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.Y. (nyan@tsinghua.edu.cn) or Y.S. (shi-lab@tsinghua.edu.cn).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Protein preparation. The cDNA of full-length FocA from *E. coli* strain O157:H7 was subcloned into pET21b (Novagen). Overexpression of FocA was induced in *E. coli* BL21(DE3) by 0.2 mM isopropyl β-D-thiogalactoside (IPTG) when the cell density reached an attenuance (D_{600}) of 1.5. After growth for 16 h at 22 °C, the cells were harvested, homogenized in buffer containing 25 mM Tris-HCl pH 8.0 and 150 mM NaCl, and lysed by sonication. Cell debris was removed by low-speed centrifugation for 10 min. The supernatant was collected and applied to ultracentrifugation at 150,000*g* for 1 h. Membrane fraction was harvested and incubated with 1.5% (w/v) *n*-octyl-β-D-glucopyranoside (β-OG; Anatrace) for 3 h at 4 °C. After another ultracentrifugation step at 150,000*g* for 30 min, the supernatant was collected and loaded on Ni^{2+}-nitrilotriacetate affinity resin (Qiagen) and washed with 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 10 mM imidazole, 1.06% β-OG. The protein was eluted from the affinity resin by 25 mM Tris-HCl pH 8.0, 300 mM imidazole, 1.06% β-OG, and concentrated to about 10 mg ml^{-1} before further purification by gel filtration (Superdex-200; GE Healthcare). The buffer for gel filtration contained 25 mM Tris-HCl pH 8.0, 150 mM NaCl and detergents. The peak fraction was collected and concentrated to about 6 mg ml^{-1} for crystallization.

The FocA mutants were generated with a standard PCR-based strategy and were subcloned, overexpressed and purified in the same way as the wild-type protein. Limited proteolysis was used to identify the structural core of FocA. Mass spectrometry revealed that the N-terminal 21 amino-acid residues were removed by digestion with V8. A new construct (residues 22–285) was made to express the N-terminally truncated protein. Truncation of the N-terminal 21 residues had no apparent effect on the channel activity of FocA. AqpZ was cloned from *E. coli* genome DNA to pET21b, overexpressed, and purified as described for FocA.

Crystallization. Crystals were grown at 18 °C by the hanging-drop vapour-diffusion method. Full-length FocA protein purified in 0.4% (w/v) decyl-β-D-maltopyranoside (DM; Anatrace) gave rise to large diamond-shaped crystals in multiple poly(ethylene glycol) (PEG) conditions. However, the best data set collected at beamline BL41XU of Spring-8 for these crystals, at a nominal resolution of 4 Å, was unsuitable for structural determination. The truncated FocA purified in 0.8% (w/v) β-OG and 0.046% (w/v) *n*-dodecyl-*N*,*N*-dimethylamine-*N*-oxide (LDAO; Anatrace) gave rise to crystals of trigonal plates. The crystals appeared overnight in the well buffer containing 0.1 M MOPS pH 7.5, 36% (w/v) PEG400, 200 mM NaCl or sodium formate, and grew to full size in one week. The crystals from space group $P3_2$ diffracted to 3.2 Å at BL41XU. To further improve the resolution, a third detergent was screened as an additive. Finally, 0.2% Cymal-2 was shown to be essential for improving the diffracting resolution from 3.5 Å to 2.2 Å. The crystals belong to space group $P2_12_12_1$, with unit cell dimensions of $a = 102.31$ Å, $b = 107.07$ Å, $c = 164.24$ Å, and $\alpha = \beta = \gamma = 90°$. Derivative crystals were obtained by soaking crystals for 24 h in mother liquor containing 2 mM K_2PtCl_4 followed by back-soaking for 3 min in well buffer plus 1.0% β-OG. Both native and heavy-atom-derived crystals were directly flash-frozen in a cold nitrogen stream at 100 K.

Data collection and structure determination. All data sets were collected at the Spring-8 beamline BL41XU, and processed with HKL2000 (ref. 27). Additional processing was performed with programs from the CCP4 suite[28]. Data collection statistics are summarized in Supplementary Table 1. The platinum sites were located with SHELXD[29] from the Bijvoet differences in the Pt-SAD data. The identified positions were refined and the phases were calculated with SAD experimental phasing module of PHASER[34]. The real-space constraints, including solvent flattening, histogram matching and non-crystallographic symmetry averaging, were applied to the electron density map in DM[30]. Cross-crystal averaging in DMMulti[30] gave rise to electron density maps of sufficient quality for model building, using the Pt-SAD and $P2_12_12_1$ native data. The initial model was built with BUCCANEER[31]. Additional missing residues in the automatically built model were added manually in *Coot*[32]. The final model in the $P2_12_12_1$ space group was refined with PHENIX[33]. Of the amino acids in the final atomic model, 94.1%, 5.8% and 0.1% are in the most favourable, additional allowed, and generously allowed regions of the Ramachandran plots, respectively. No amino acid is in the disallowed region. The refined model for one FocA protomer was used for molecular replacement with the program PHASER[34] into the hexagonal crystal form, and ten protomers per asymmetric unit were found. The $P3_2$ structure was also refined with PHENIX[33]. Of the amino acids in the final atomic model, 87.5%, 11.5%, 0.8% and 0.2% in the $P3_2$ space group are in the most favourable, additional allowed, generously allowed, and disallowed regions of the Ramachandran plots, respectively.

Preparation of liposome and proteoliposome. *E. coli* polar lipids (Avanti Polar Lipids) were dissolved in chloroform/methanol mixture (3:1, v/v) at 50 mg ml^{-1} and dried under a nitrogen stream. The lipids were then dissolved at 20 mg ml^{-1} in buffer containing 20 mM HEPES pH 7.0 and 0.4 mM dithiothreitol, and incubated at 22 °C for 1 h followed by sonication for 2 h.

The liposomes and proteoliposomes used for the water permeability assay were reconstituted as described in ref. 35; those used for the formate permeability assay were prepared as described in ref. 16.

Assay for water permeability. To measure the water permeability of different target proteins, the osmotic responses of proteoliposomes were monitored as described previously[16,35,37,38]. In brief, 75 μl of liposome or proteoliposome solution was rapidly mixed with an equal volume of 500 mM sucrose in the same buffer (20 mM HEPES pH 7.0). The osmotic pressure then causes water efflux from the lipid vesicle, leading to a decrease in the vesicle volume and an increase in the light scattering signal. The experiment was performed on a stopped-flow device (Applied Photophysics PiStar180) at 5 °C. The light scattering signal was recorded at an emission wavelength of 440 nm. The data were processed with software Origin to fit the equation $Y = ae^{-kt} + b$, where t is time and Y is the signal of light scattering. The relative rate of water conduction (k value) was calculated to compare the water permeability of AqpZ and FocA.

Assay for formate permeability. To measure the formate permeability, the liposomes or proteoliposomes were rapidly mixed with 20 mM formate at equal volume. The change in vesicle size was detected by recording the light scattering signal at 440 nm. The above experiments were performed on a stopped-flow apparatus (Applied Photophysics PiStar180) at 10 °C. On the basis of a published protocol[16], changes in light scattering could be fitted by two exponentials in the equation $Y = [a(1 - e^{-kt}) - b]e^{-\mu t} + c$, where t is time and Y is the signal of light scattering. The first time-constant (k) corresponds to the rapid water efflux. The second time-constant (μ) corresponds to the relative rates of swelling again due to formate conduction[16]. The data were analysed with Origin.

37. Borgnia, M. J. & Agre, P. Reconstitution and functional comparison of purified GlpF and AqpZ, the glycerol and water channels from *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **98**, 2888–2893 (2001).

38. Khademi, S. *et al.* Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. *Science* **305**, 1587–1594 (2004).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 462 | 26 November 2009 | doi:10.1038/nature08542

Structure and hydration of membranes embedded with voltage-sensing domains

Dmitriy Krepkiy[1]*, Mihaela Mihailescu[2,5]*, J. Alfredo Freites[2,3], Eric V. Schow[4], David L. Worcester[2,5,6], Klaus Gawrisch[7], Douglas J. Tobias[3], Stephen H. White[2,5] & Kenton J. Swartz[1]

Despite the growing number of atomic-resolution membrane protein structures, direct structural information about proteins in their native membrane environment is scarce. This problem is particularly relevant in the case of the highly charged S1–S4 voltage-sensing domains responsible for nerve impulses, where interactions with the lipid bilayer are critical for the function of voltage-activated ion channels. Here we use neutron diffraction, solid-state nuclear magnetic resonance (NMR) spectroscopy and molecular dynamics simulations to investigate the structure and hydration of bilayer membranes containing S1–S4 voltage-sensing domains. Our results show that voltage sensors adopt transmembrane orientations and cause a modest reshaping of the surrounding lipid bilayer, and that water molecules intimately interact with the protein within the membrane. These structural findings indicate that voltage sensors have evolved to interact with the lipid membrane while keeping energetic and structural perturbations to a minimum, and that water penetrates the membrane, to hydrate charged residues and shape the transmembrane electric field.

Membrane-embedded S1–S4 voltage-sensing domains are used by membrane proteins to sense and react to changes in membrane voltage (Fig. 1a). In the voltage-activated potassium (K_v), sodium and calcium channels, these domains drive opening and closing of an associated ion-conducting pore domain (Fig. 1a) to generate electrical signals[1]. In the *Ciona intestinalis* voltage-sensitive phosphatase, an S1–S4 domain controls the hydrolysis of phospholipids by an associated phosphatase[2], and in voltage-activated proton channels, the S1–S4 domain contains the permeation pathway for protons[3]. X-ray structures of S1–S4 domains show that the protein domain comprises four transmembrane α-helices (Fig. 1a) and that its structure is well conserved in organisms ranging from archaebacteria to mammals[1,4,5].

A fundamental feature of S1–S4 domains is that they contain basic and acidic residues that enable the protein to change conformation in response to rapid fluctuations in membrane voltage[1,6,7]. In these voltage sensors, interactions with the surrounding lipid membrane have crucial roles. The S3b–S4 paddle motif within S1–S4 domains, for example, moves at the protein–lipid interface[5,8–13], and alterations in the composition of the lipid membrane alter voltage-sensor activation[14–17]. The polar nature of voltage sensors and their intimate interactions with the bilayer raise the possibility that these domains perturb the structure of the surrounding lipid bilayer. In addition, although spectroscopic and functional studies suggest that the electric field is focused across voltage sensors[18–21], the structural basis for focusing is unclear. It is not known whether deformations of the membrane contribute to focusing the electric field or whether the shape and chemistry of the protein are mainly responsible. Crevices observed in X-ray structures of S1–S4 domains would be expected to reshape the electric field, but only if they persist and are filled with water when the domain is embedded in a lipid membrane. Although water penetration of the membrane has been inferred from accessibility studies[9,13,19,22–26] and simulations[12,27–29], hydration of voltage sensors has not been measured.

Structure and hydration of membranes containing voltage sensors

To address these fundamental issues, we developed a homogeneous preparation of voltage-sensing domains incorporated into lipid membranes for use with neutron diffraction[30–33]. The neutron scattering length gives the relative amplitude of the de Broglie wave scattered from a nucleus and is analogous to the X-ray scattering length of an electron. We focused our efforts on the S1–S4 domain of K_vAP, an archaebacterial channel from *Aeropyrum pernix* that can be robustly expressed, stably purified and reconstituted into lipid membranes[4,9,10,13,34]. After expression and purification of the S1–S4 domain of K_vAP (Methods), circular dichroism spectroscopy reveals that the domain has high α-helical content in detergent micelles or reconstituted into liposomes (Fig. 1b), consistent with the X-ray structure of the domain[4] and electron paramagnetic resonance studies[9,13,34]. To investigate the topology of the S1–S4 domain in liposomes, we measured the fluorescence of the single Trp 70 residue near the middle of the S2 helix. The emission spectrum of Trp 70 is shifted towards shorter wavelengths relative to that of free Trp in aqueous solution (Fig. 1c), indicating that Trp 70 resides in a non-polar environment[35]. Moreover, its fluorescence is efficiently quenched by bromine atoms covalently bound to lipid hydrocarbon tails[36] but not by the aqueous quencher acrylamide (Fig. 1c–e), consistent with the S2 helix having a transmembrane topology (see below).

To determine the profile structure of bilayers containing S1–S4 domains, we produced oriented lipid multilayers by deposition of proteoliposomes or neat liposomes on glass substrates. When these multilayers were hydrated (86 or 93% relative humidity) and mounted in a cold neutron beam[37], we observed strong lamellar

[1] Molecular Physiology and Biophysics Section, Porter Neuroscience Research Center, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA. [2] Department of Physiology and Biophysics, and Center for Biomembrane Systems, [3] Department of Chemistry and Institute for Surface and Interface Science, [4] Department of Physics and Astronomy and Institute for Genomics and Bioinformatics, University of California, Irvine, California 92697, USA. [5] NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA. [6] Biology Division, University of Missouri, Columbia, Missouri 65211, USA. [7] Laboratory of Membrane Biochemistry and Biophysics, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health, Bethesda, Maryland 20892, USA.

*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

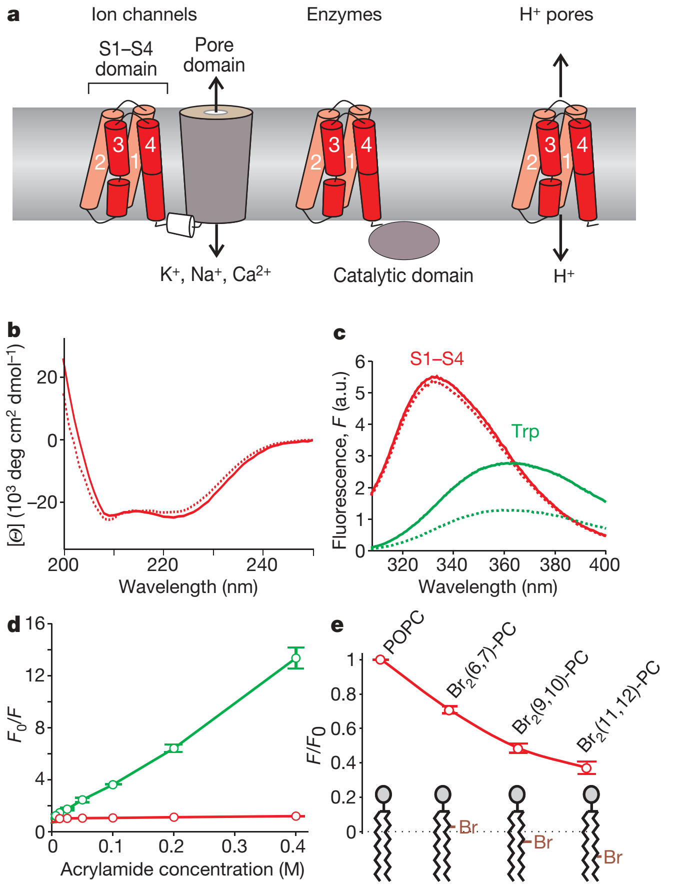


Figure 1 | S1–S4 voltage-sensing domains and their biophysical properties in lipid bilayers. a, Representation of membrane proteins containing S1–S4 voltage-sensing domains (red), embedded in the lipid bilayer (light grey). **b,** Circular dichroism spectra of the S1–S4 voltage-sensing domain of K_vAP in octyl glucoside micelles (dotted red line) and reconstituted into 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) and 1-palmitoyl-2-oleoyl-*sn*-glycero-3-[phospho-*rac*-(1-glycerol)] (POPG) proteoliposomes at a protein/lipid molar ratio of 1:130 (solid red line). The spectra indicate high (~85%) helical content (Methods). [Θ], mean residue ellipticity. **c,** Fluorescence emission spectra of Trp 70 within the S1–S4 voltage-sensing domain after reconstitution in POPC–POPG (protein/lipid molar ratio of 1:100) in the absence (solid red line) and presence (dotted red line) of 50 mM aqueous acrylamide. Also shown are the emission spectra of free Trp (25 μM) mixed with POPC–POPG liposomes of identical lipid concentration in the absence (solid green line) and presence (dotted green line) of 50 mM aqueous acrylamide. a.u., arbitrary units. **d,** Stern–Volmer plots for acrylamide quenching of fluorescence emission of Trp 70 within the S1–S4 voltage-sensing domain (red) and of free Trp (green). Error bars, s.e.m. ($n = 3$). The Stern–Volmer constant for quenching was 0.4 ± 0.02 M^{-1} for Trp 70 and 26.2 ± 0.2 M^{-1} for free Trp. F_0, fluorescence in absence of a quencher. **e,** Quenching of fluorescence emission of Trp 70 within the S1–S4 voltage-sensing domain by bromine atoms attached at different positions along the lipid hydrocarbon tail. The protein was reconstituted into the 1:1 mixture of POPG and either POPC or one of three dibrominated phosphatidylcholines (PCs), Br_2(6,7)-PC, Br_2(9,10)-PC or Br_2(11,12)-PC (Methods). Error bars, s.e.m. ($n = 3$).

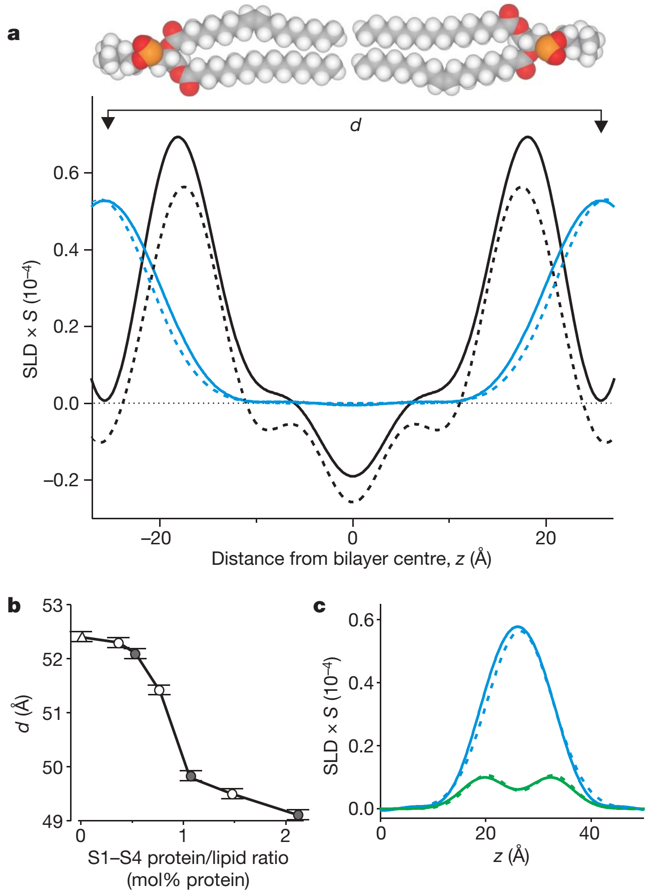


Figure 2 | Scattering-length density profiles for bilayers containing S1–S4 voltage-sensing domains. a, Scattering-length density profiles, on an absolute scale[32], of the S1–S4 voltage-sensing domain in lipid bilayers (black solid line) and the water distribution (blue solid line). The protein/lipid molar ratio is 1:130 (0.77 mol%) and relative humidity is 86%. Profiles for lipid in the absence of protein (dashed lines) are shown for comparison. The density profile amplitudes are presented in units of scattering length per unit length, corresponding to the scattering-length density (SLD) of a unit cell (POPC/POPG/protein ratio, 0.4962:0.4962:0.0076; plus 8.5 water molecules) multiplied by the area per lipid (*S*; Methods). The *x* axis shows the distance from the bilayer centre (*z*), with the origin positioned in the middle of the bilayer. A space-filling model of POPC is shown above the plot (hydrogen, white; carbon, grey; oxygen, red; phosphorous, yellow). **b,** Effect of S1–S4 voltage sensing domains on the bilayer thickness at different protein/lipid ratios. The triangle marks the value of *d* for neat-lipid bilayers. Open circles are for voltage-sensing domains with His tag removed (Methods) and filled circles are for the His-tagged protein. Error bars represent 1 s.d., obtained from least-squares fitting of angular positions of the Bragg peaks. **c,** Distribution of deuterium atoms in head-group-labelled phosphocholine ($\text{-C}^2\text{H}_2\text{-C}^2\text{H}_2\text{-}$; D4 lipid) in bilayers containing S1–S4 domains (solid green line) and comparison with the distribution of water (solid blue line). Dashed lines show the distributions of D4 lipid and water in the absence of the protein. The protein/lipid ratio is as in **a** and 25-mol% D4-POPC is used in the mixture of POPC and POPG. The relative humidity is 93%.

diffraction patterns with Bragg spacing *d* (Supplementary Fig. 1a, b). One-dimensional, absolute-scale, scattering-length density profiles along the normal of the lipid bilayer plane were then computed from the observed structure factors (Fig. 2a). The constructed neutron scattering-length density profiles for neat-lipid bilayers show the distribution of lipid (black dashed line), with positive densities for the head-group region, a trough for the hydrocarbon tails and negative densities near the terminal methyl groups (Fig. 2a). The scattering lengths for most of the relevant nuclear species (carbon, nitrogen, oxygen and phosphorous) have similarly positive values; a notable exception is hydrogen, which has a negative scattering length. The average scattering-length density of the bilayer hydrocarbon core is close to zero because the scattering length of carbon is positive and that of hydrogen is negative. The head-group peaks appear closer together than they would in an equivalent X-ray experiment[33] because X-rays scatter most strongly from head-group phosphates, whereas neutrons scatter most strongly from the carbonyl groups owing to their relative lack of hydrogen atoms.

The overall scattering-length density of the bilayer increases in the presence of the protein (Fig. 2a, solid black line), consistent with the S1–S4 domain having a transmembrane topology (see below).

©2009 Macmillan Publishers Limited. All rights reserved

Comparison of the lipid bilayer profiles with and without S1–S4 domains reveals how the structure of the bilayer is influenced by the protein (Fig. 2a). Although the S1–S4 domain does not radically alter the structure of the lipid bilayer, examination of the profiles shows that the voltage sensors produce a detectable thinning of the bilayer, as revealed by a decrease in *d* (Fig. 2b). The thickness decrease depends on the concentration of the protein in the membrane, with a maximal decrease of about 3 Å at protein/lipid molar ratios greater than 1:100 (Fig. 2b). These results indicate that lipid molecules in the membrane maintain a bilayer-like arrangement around voltage sensors, consistent with the lipids resolved in the recent crystal structure of the K_v1.2/2.1 paddle chimaera[5].

Next we performed experiments to determine the distribution of water and to quantify the number of water molecules per lipid by using contrast variation between water (^1H_2O) and deuterium oxide (^2H_2O) and by comparing with lipids containing four deuterium atoms in the head-group region (D4 lipids; Fig. 2c; Methods). This approach takes advantage of the fact that deuterated nuclei have a positive scattering length whereas that of hydrogen is negative; selective substitution of deuterium for hydrogen therefore allows the deuterium atoms to be easily detected against the low scattering-length density of the hydrocarbon core. Although the water distributions show that thinning of the bilayer brings water on the two sides of the bilayer closer together (Fig. 2a, blue lines), we could not detect a change in the shape of the water distribution or the total water content. At 86% relative humidity, the unit cell of the membrane contains 8.1 ± 0.7 water molecules per lipid for neat-lipid bilayers, compared with 8.5 ± 0.5 in the presence of 0.77-mol% protein. At 93% relative humidity, water content was 10.6 ± 0.2 molecules per lipid in the absence of 0.77-mol% protein and 11.0 ± 0.2 molecules per lipid in its presence (Supplementary Fig. 2).

Distributions of S1–S4 voltage-sensing domains and water across membranes

To investigate the membrane topology and hydration of S1–S4 domains directly, we determined the protein distribution using contrast variation between protonated and deuterated S1–S4 domains. The S1–S4 domain of K_vAP was uniformly deuterated to 74% (Fig. 3a) and multilayers were formed with either protonated or deuterated protein at the same protein/lipid ratio and lipid composition. The two types of sample display similar diffraction patterns, with the same number of observed diffraction orders and repeat spacing (Supplementary Fig. 3). Subtraction of scaled profiles to obtain the protein density distribution reveals the distribution of the protein across the bilayer (Fig. 3b, red line). Maxima in the density distribution are observed in the head-group region of the bilayer and minima are observed in the inter-bilayer space. This finding firmly establishes that S1–S4 domains adopt a transmembrane topology when embedded in a lipid membrane, with the four helices oriented roughly normal to the membrane plane. It is not surprising to find a significant protein density in the inter-bilayer space, given that the dimensions of the S1–S4 helices[4,5] are similar to the thickness of the bilayer (and that the helices may protrude somewhat outside the membrane).

Having determined the distribution of protein in the bilayer (Fig. 3b, red), we then compared it to that of water (Fig. 3b, blue) to ascertain whether S1–S4 domains are hydrated. Strikingly, the two distributions have an extensive overlap within the confines of the lipid membrane, in particular for the outer halves of the bilayer. Because the voltage sensor does not detectably alter water content or the shape of the water distribution, the hydration detected in these neutron diffraction experiments is mainly from water that is already present in the bilayer. The voltage sensors may bring additional water molecules into the bilayer, but it is unlikely that we would have yet detected them in the experiments. For example, molecular dynamics simulations predict that 45–47 water molecules intimately associate with each voltage sensor (see below); such association would not

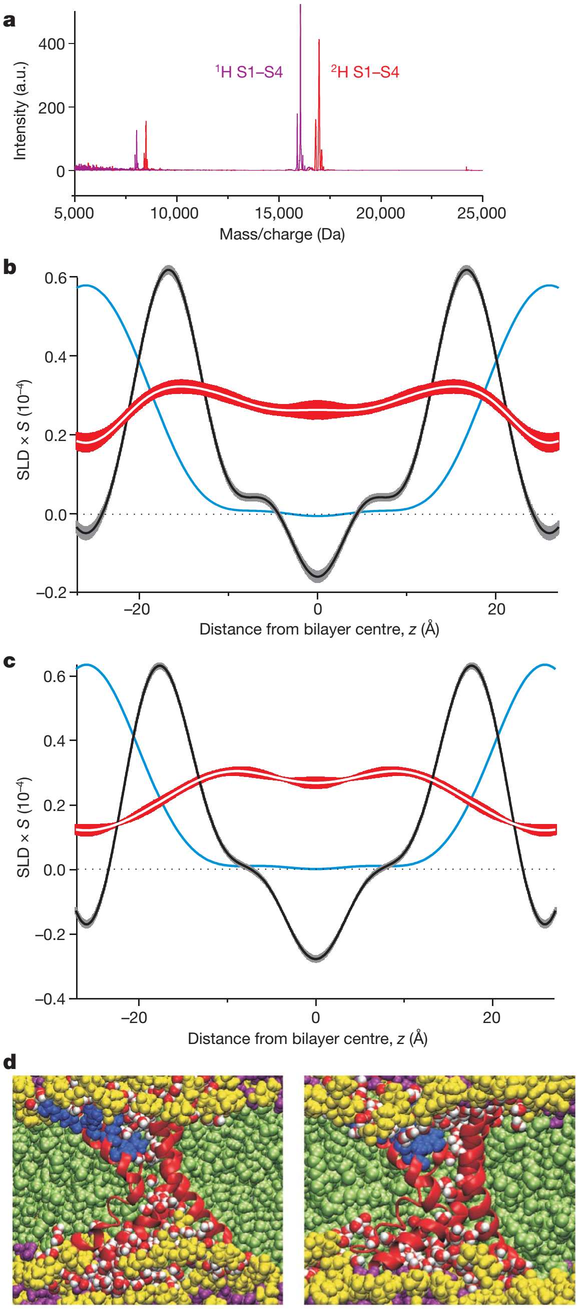


Figure 3 | Deuteration of S1–S4 voltage-sensing domains and distribution of the protein in lipid membranes. a, Matrix-assisted laser desorption/ionization (MALDI) mass spectra of the protonated S1–S4 domain of K_vAP (^{1}H S1–S4, purple) and the uniformly deuterated S1–S4 domain (^{2}H S1–S4, red). The difference in mass indicates that the protein is deuterated to 74%. **b,** Trans-bilayer distribution of the S1–S4 domain (white line surrounded by a broad red band) obtained in neutron diffraction experiments from the profile difference between deuterated and protonated S1–S4 domains. Profiles are shown on an absolute (per lipid) scale. Water distribution is shown in blue and lipid as a black line surrounded by a grey band. The broad bands represent estimates of experimental uncertainty computed using the methods of ref. 33. The protein/lipid ratio is 1:130 (0.77 mol%) and the relative humidity is 93%. **c,** Neutron scattering-length density profiles for the simulation system with 11 water molecules per lipid. Trans-bilayer distribution of the S1–S4 domain is shown as a white line surrounded by a broad red band (estimated experimental uncertainty), water is shown in blue and lipid is shown in black. **d,** Snapshots of the region in the vicinity of one of the two voltage-sensing domains from the molecular dynamics simulation of a stack of two bilayers with 11 water molecules per lipid (left) and excess water (right). Water molecules within 6 Å of protein are shown as red–white spheres and all other water molecules are coloured purple. Phosphocholine head groups are coloured yellow and the acyl chains are coloured light green. The ribbon diagram of the S1–S4 domain is coloured red with the outer four Arg residues in S4 shown as blue CPK models.

©2009 Macmillan Publishers Limited. All rights reserved

detectably alter the shape of the water distribution determined in neutron diffraction experiments because these molecules constitute less than 4% of all water molecules in the system for each voltage sensor (at 0.77-mol% protein and 93% humidity, the protein/lipid/water ratio is 1:130:1,430).

Predicted distributions of water, lipid and protein with varying hydration

To explore whether the distributions of water, lipid and protein observed in neutron diffraction experiments are compatible with those predicted from molecular dynamics simulations, we calculated neutron diffraction structure factors from molecular dynamics simulation trajectories for the S1–S4 domain of K_vAP embedded in a lipid bilayer in a transmembrane orientation (Methods). The resulting Bragg spacing is in excellent agreement with the experimental results, and the overall bilayer scattering-length density profile and water distributions, determined by applying the same procedures as in the reduction of the experimental data, are in good agreement with the experimental results (Fig. 3c). Simulation and experiment show comparable overall protein density distributions in the membrane interior, as well as overlap between the distributions of protein and water, suggesting a similar disposition of protein and water in the lipid bilayer. Compared with that used for simulations, the protein studied experimentally contains 18 additional residues on the amino terminus (Methods), precluding a quantitative comparison of the experimental and simulated protein profiles. On the basis of the location of the corresponding segment in the structure of the K_v1.2/2.1 paddle chimaera[5], the excess scattering-length density in the experimental data relative to the simulation near the membrane–water interface ($|z| > 10$ Å) can reasonably be attributed to the 18-residue amphipathic segment absent in the simulation.

To explore whether hydration of the preparation influences these distributions, we compared a simulation with 11 water molecules per lipid (corresponding to 93% relative humidity) with a previously reported simulation in a lipid bilayer in excess water[27] (Fig. 3d), finding that the scattering-length density profiles observed in the two cases are similar (Supplementary Fig. 4). The structure of the S1–S4 domain was also relatively insensitive to hydration level (Supplementary Fig. 5), and in each case a similar number of water molecules (45–47 at 11 water molecules per lipid and 48–49 in excess water) are intimately associated with the protein domain within the hydrophobic core of the bilayer (Supplementary Fig. 6). In addition, solvation of crucial S4 Arg residues by both water and phosphate head groups is similarly observed at varying hydration levels (Supplementary Fig. 7). Together these observations suggest that the extent to which hydration of the preparation influences the structure of lipid membranes containing voltage sensors is minor, and would not be discernible in the neutron scattering profiles at the protein concentrations studied.

Interaction between water and S1–S4 voltage-sensing domains

Although the neutron diffraction experiments indicate that the distribution of water and protein in the bilayer overlap, they do not directly address the question of whether water is intimately associated with voltage sensors (Fig. 4a). To explore this possibility, we used solid-state NMR spectroscopy to measure magnetization transfer from water to lipid by means of intermolecular ^{1}H dipole–dipole interactions in the presence of the voltage sensor. Well-resolved lipid resonances (Fig. 4b, black spectra) were observed when the sample was rapidly spun (10 kHz) at an angle of 54.7° (the magic angle) to the magnetic field, a procedure that averages out anisotropic dipolar interactions that broaden resonance lines.

We performed saturation-transfer difference experiments[38,39] by comparing lipid spectra before (Fig. 4b, black spectra) and after (Fig. 4b, blue spectra) applying saturating radio-frequency pulses at the ^1H_2O resonance frequency (chemical shift, 4.79 p.p.m.). Magnetization transfer to lipid would cause a decrease in the intensity of

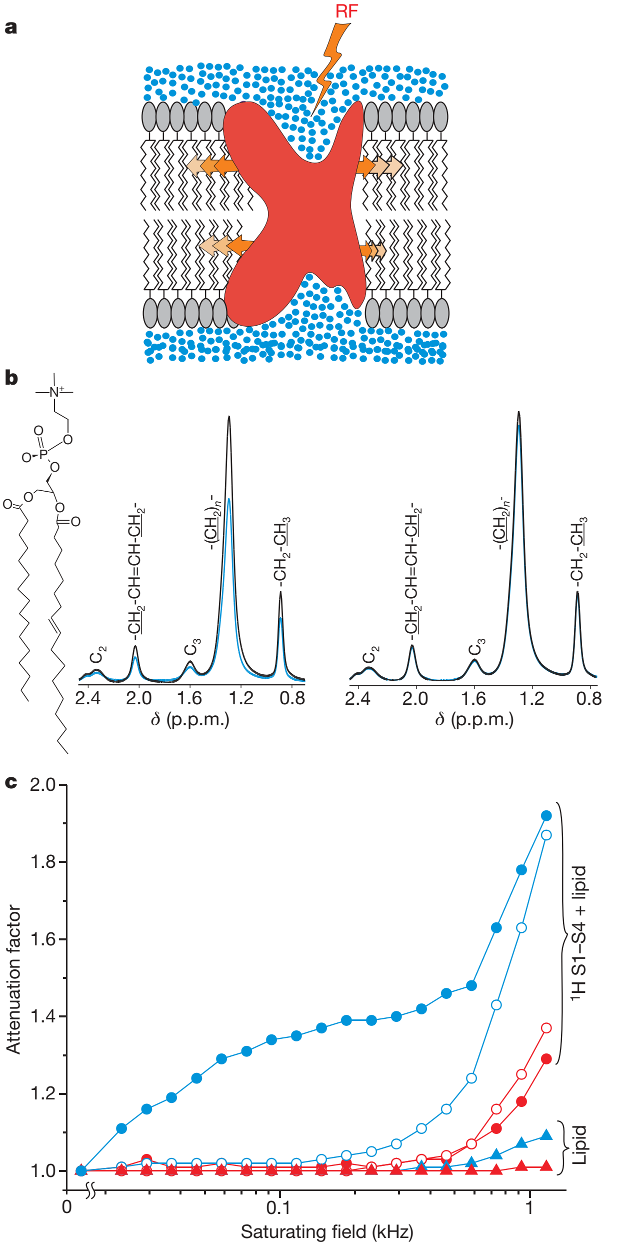


Figure 4 | Interaction of water and S1–S4 voltage-sensing domains within lipid membranes. **a,** Schematic representation of crevices in S1–S4 voltage-sensing domains (red) filled with water (blue) and an experiment in which selective, saturating radio-frequency (RF) pulses were applied at the water resonance (4.79 p.p.m.). Magnetization transfer (arrows) from water through protein to the surrounding lipid results in the attenuation of the lipid ^{1}H NMR signals. **b,** Aliphatic region of magic-angle-spinning ^{1}H NMR spectra of a lipid sample containing S1–S4 voltage-sensing domains in the presence of ^1H_2O (left-hand spectrum) or ^2H_2O (right-hand spectrum). Lipid resonances for both POPC and POPG (present in a 1:1 mix) are indicated on the spectra, with peaks corresponding to underlined ^{1}H atoms. Attenuation of the methylene resonance (1.3 p.p.m.) is observed (blue traces) when saturating radio-frequency pulses (field strength, 232 Hz) are applied at 4.79 p.p.m. to a sample containing ^1H_2O (left), but not to one containing ^2H_2O (right). Field strength is reported in frequency units as $(\gamma/2\pi)B_1$, where γ is the proton gyromagnetic ratio and B_1 is the magnetic component of the radio-frequency pulse. Attenuation is defined as signal intensity recorded without saturation divided by signal intensity recorded with saturation. δ, chemical shift. The structure of POPC is shown on the left. **c,** Attenuation factors plotted as a function of radio-frequency field strength. The carrier radio frequency was set at either the water resonance (4.79 p.p.m., blue) or in the protein amide region (8.5 p.p.m., red). Circles show data for S1–S4 domains in lipid bilayers where the protein/lipid ratio is 1:100. Triangles show data for samples containing lipid alone. Samples containing ^1H_2O are indicated by filled symbols, whereas those containing ^2H_2O are indicated by open symbols.

©2009 Macmillan Publishers Limited. All rights reserved

lipid resonances, which can be quantified as an attenuation factor for different saturating field strengths (Fig. 4c). Control experiments in which neat-lipid membranes were studied show that magnetization transfer from ^1H_2O to lipid is inefficient when radio-frequency pulses are applied to ^1H_2O (Fig. 4c, blue triangles). In the presence of the protein, magnetization transfer is similar in ^1H_2O and ^2H_2O when radio frequency pulses are applied directly to the protein amide resonance frequency (8.5 p.p.m.; Fig. 4c, red filled (^1H_2O) and open (^2H_2O) circles). In contrast, magnetization transfer from water to lipid is very efficient in membranes containing S1–S4 voltage-sensing domains in the presence of ^1H_2O (Fig. 4c, blue filled circles). Much weaker transfer is observed when ^1H_2O is replaced with ^2H_2O (Fig. 4c, blue open circles), demonstrating that most of the attenuation results from magnetization transfer starting from ^1H_2O rather than from resonances of aliphatic hydrogen atoms of the protein that overlap with the water resonances at 4.79 p.p.m. (The attenuation observed in ^2H_2O at a saturating field of >200 Hz probably originate from aliphatic hydrogen atoms.) When considered together with the overlap in water and protein distributions observed in neutron diffraction experiments (Fig. 3b), these NMR results indicate that water intimately associates with the protein within the bilayer.

Discussion

The objective of the present study was to investigate the structure and hydration of lipid membranes containing S1–S4 voltage-sensing domains. Previously, only computational approaches have been used to explore how different types of membrane protein influence the structure of the bilayer[12,27–29,40,41]. We adapted neutron diffraction techniques to determine how voltage sensors influence membrane structure; these protein domains are highly polar and exhibit important interactions with the lipid membrane, making them a particularly interesting test case. We succeeded in reconstructing the bilayer profile in the presence of the S1–S4 domain of K_vAP, which shows that the structure of the lipid bilayer remains largely intact around the protein. The most notable change is that the protein causes a thinning of the bilayer by about 3 Å. Neutron diffraction measurements reflect changes in the structure of the bilayer, averaged over the entire membrane plane. Molecular dynamics simulations, yielding a value of the Bragg spacing consistent with the diffraction experiments, predict that the distortion of the lipid bilayer by the protein is restricted to the lipids immediately surrounding the voltage sensors (Fig. 5a; Supplementary Fig. 8). Taken together, the modest membrane-averaged thinning and local adaptation of the lipid bilayer to the presence of the voltage sensor suggest that the protein has evolved to interact with lipid molecules while minimizing the energetic and structural perturbations of the bilayer.

Our neutron diffraction, solid-state NMR and simulation results indicate that S1–S4 voltage-sensing domains are hydrated in the bilayer and that water interacts intimately with the protein. The observed hydration can explain the accessibility of water-soluble reagents to residues in S1–S4[9,13,19,22–26] and suggests that the crevices seen in X-ray structures[4,5], which house the Arg residues that carry gating charge[6,7], actually contain water when the protein is embedded in a lipid membrane, as illustrated in Fig. 3d. Hydration of these critical residues will raise the local dielectric constant within the bilayer, ensuring that the Arg residues remain charged and thereby move in response to changes in membrane voltage. Consideration of the observed average water distributions in the presence of S1–S4 domains indicates that the membrane electric field decreases over a distance of no more than about 25 Å, which is the hydrophobic thickness of the bilayer in the presence of the voltage-sensing domain. The water-filled crevices in the structure of the S1–S4 domain would be expected to focus the electric field further, in agreement with our simulations showing that the transmembrane potential is contoured by the structure of the protein and decreases over a distance of about 20 Å (Fig. 5b). These simulations do show significant distortions of the lipid bilayer in the local vicinity of the protein (Fig. 5a), but these do not have pronounced effects on the transmembrane voltage.

Although the effects of S1–S4 voltage-sensing domains on the physical thickness of the bilayer are not large, the bilayer thinning we observe indicates that the protein and bilayer do interact, thereby providing a basis for understanding how lipid modification can influence voltage-sensor function[14–17]. For example, on the basis of theoretical considerations and studies of gramicidin channels[42–44], the thinning we observed would be expected to have profound effects on the mechanical properties of channels containing S1–S4 domains and may help to explain the sensitivity of voltage-activated ion channels to alterations in the mechanical properties of the lipid bilayer[41,45–47].

The hydration and reshaping of the lipid membrane that we observe for voltage sensors will probably be relevant for other classes of membrane protein. For example, the presence of binding sites for water-soluble ligands deep within G-protein-coupled receptors[48] and transporters[49] implies that these proteins are hydrated within the membrane. In the case of intramembrane enzymes, such as the rhomboid protease[50], hydrolysis of the peptide bond requires the presence of water molecules in an active site located in the hydrophobic interior of the membrane. In each of these instances, little is

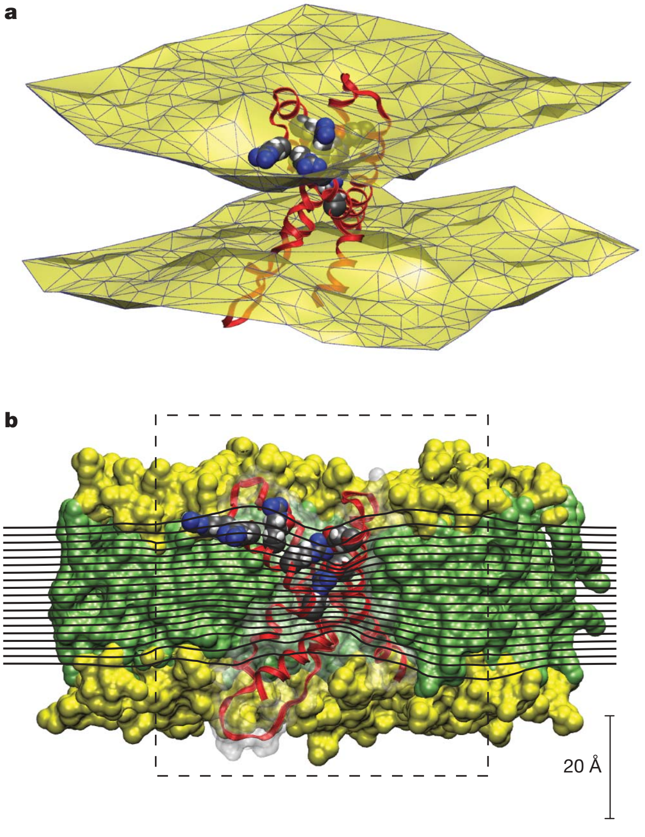


Figure 5 | Effects of the voltage-sensing domain on a lipid bilayer as revealed by molecular dynamics simulation. a, The lipid bilayer interface, represented as a two-dimensional Delaunay triangulation for the average positions of lipid carbonyl carbon atoms, reveals local distortions around the voltage-sensing domain. **b,** The transmembrane equipotential surfaces (black lines; contour interval is 5% of the applied potential) on a slice passing through the system centre (the Arg 133–Asp 62 salt bridge) show focusing features in the cavities of the voltage-sensing domain. Contributions to the molecular surface from aliphatic chains (green), polar groups (yellow), and proteins (white, transparent) of the corresponding cut-away view are shown as background. The dashed box indicates the region of the system considered atomistic in the calculation of the transmembrane potential (Methods). In both panels, the voltage-sensing domain is in ribbon representation (red) with the outer four Arg residues in S4, and their salt-bridge partners, shown in CPK representation and coloured by atom (carbon, grey; nitrogen, blue; oxygen, red; hydrogen, white).

©2009 Macmillan Publishers Limited. All rights reserved

known about the structure of the surrounding lipid membrane or whether adaptations may be important in the mechanism of the protein.

METHODS SUMMARY

The S1–S4 domain of $\mathrm{K_vAP}$ was expressed and purified as previously described[4] and reconstituted into a 1:1 mixture of POPC and POPG using rapid dilution of detergent micelles containing the protein and lipids. We prepared lipid multilayer samples for neutron diffraction by deposition of aqueous dispersions of liposomes on glass slides. A solution containing ~4 mg of lipid was applied to the glass surface (15-mm diameter), dried under vacuum and rehydrated using water vapour in a sealed chamber containing saturated salt solutions. Three types of sample were prepared, the first using a protonated voltage-sensing domain and protonated lipid, the second using a deuterated voltage-sensing domain and protonated lipid, and the third using a protonated voltage-sensing domain and head-group-deuterated POPC lipid (D4 lipid: $(\mathrm{CH_3})_3$-N-$\mathrm{C^2H_2}$-$\mathrm{C^2H_2}$-).

Oriented multilayers (~10-μm thick, corresponding to 2,000–3,000 bilayers) were transferred into the sample chamber of the neutron diffractometer and hydrated through the vapour phase at a temperature of 298 K. Neutron diffraction measurements were performed at the Advanced Neutron Diffractometer/Reflectometer[37], located at the US National Institute of Standards and Technology Center for Neutron Research, Maryland. Monochromatic cold neutrons of wavelength $\lambda = 5$ Å and a wavelength spread of $\Delta\lambda/\lambda = 1\%$ were diffracted by the sample and counted with a pencil-type ^{3}He detector. Specular $\Theta - 2\Theta$ scans were made to ensure that the momentum transfer (Q_z, typically 0–1.2 Å^{-1}) between the incident and diffracted neutron wavevectors was always perpendicular to the multilayer plane, thus probing the direction normal to the bilayer (Supplementary Fig. 1). We performed solid-state NMR experiments on proteoliposome pellets in $^1\mathrm{H_2O}$ or $^2\mathrm{H_2O}$ at a water/lipid ratio of 30:1. Lipid spectra were recorded on an 800-MHz Bruker AV800 spectrometer equipped with a 4-mm ^{1}H/^{13}C/^{2}H cross-polarization magic-angle-spinning probe (Bruker BioSpin). All-atom molecular dynamics simulations of the S1–S4 domain of $\mathrm{K_vAP}$ were carried out in POPC bilayers hydrated with, respectively, 9 and 11 water molecules per lipid (corresponding to experiments at relative humidities of 86% and 93%). Each system consisted of two stacked lipid bilayers, each containing a single S1–S4 domain, arranged to form a single pseudo-centrosymmetric unit cell. The simulations were run at a constant temperature of 295 K and a constant pressure of 1 bar.

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

Received 13 July; accepted 24 September 2009.

1. Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **456,** 891–897 (2008).
2. Murata, Y., Iwasaki, H., Sasaki, M., Inaba, K. & Okamura, Y. Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor. *Nature* **435,** 1239–1243 (2005).
3. Lee, S. Y., Letts, J. A. & MacKinnon, R. Functional reconstitution of purified human Hv1 H+ channels. *J. Mol. Biol.* **387,** 1055–1060 (2009).
4. Jiang, Y. *et al.* X-ray structure of a voltage-dependent K$^+$ channel. *Nature* **423,** 33–41 (2003).
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. 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).
7. Aggarwal, S. K. & MacKinnon, R. Contribution of the S4 segment to gating charge in the Shaker K$^+$ channel. *Neuron* **16,** 1169–1177 (1996).
8. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K$^+$ channel. *Nature* **423,** 42–48 (2003).
9. Cuello, L. G., Cortes, D. M. & Perozo, E. Molecular architecture of the KvAP voltage-dependent K$^+$ channel in a lipid bilayer. *Science* **306,** 491–495 (2004).
10. Ruta, V., Chen, J. & MacKinnon, R. Calibrated measurement of gating-charge arginine displacement in the KvAP voltage-dependent K$^+$ channel. *Cell* **123,** 463–475 (2005).
11. Alabi, A. A., Bahamonde, M. I., Jung, H. J., Kim, J. I. & Swartz, K. J. Portability of paddle motif function and pharmacology in voltage sensors. *Nature* **450,** 370–375 (2007).
12. Sands, Z. A. & Sansom, M. S. How does a voltage sensor interact with a lipid bilayer? Simulations of a potassium channel domain. *Structure* **15,** 235–244 (2007).
13. Chakrapani, S., Cuello, L. G., Cortes, D. M. & Perozo, E. Structural dynamics of an isolated voltage-sensor domain in a lipid bilayer. *Structure* **16,** 398–409 (2008).
14. Schmidt, D., Jiang, Q. X. & MacKinnon, R. Phospholipids and the origin of cationic gating charges in voltage sensors. *Nature* **444,** 775–779 (2006).
15. Ramu, Y., Xu, Y. & Lu, Z. Enzymatic activation of voltage-gated potassium channels. *Nature* **442,** 696–699 (2006).
16. Milescu, M. *et al.* Interaction between lipids and voltage sensor paddles detected with tarantula toxins. *Nature Struct. Mol. Biol.* **16,** 1080–1085 (2009).
17. Xu, Y., Ramu, Y. & Lu, Z. Removal of phospho-head groups of membrane lipids immobilizes voltage sensors of K$^+$ channels. *Nature* **451,** 826–829 (2008).
18. Asamoah, O. K., Wuskell, J. P., Loew, L. M. & Bezanilla, F. A fluorometric approach to local electric field measurements in a voltage-gated ion channel. *Neuron* **37,** 85–97 (2003).
19. Starace, D. M. & Bezanilla, F. A proton pore in a potassium channel voltage sensor reveals a focused electric field. *Nature* **427,** 548–553 (2004).
20. Chanda, B., Asamoah, O. K., Blunck, R., Roux, B. & Bezanilla, F. Gating charge displacement in voltage-gated ion channels involves limited transmembrane movement. *Nature* **436,** 852–856 (2005).
21. Ahern, C. A. & Horn, R. Focused electric field across the voltage sensor of potassium channels. *Neuron* **48,** 25–29 (2005).
22. Yang, N. & Horn, R. Evidence for voltage-dependent S4 movement in sodium channels. *Neuron* **15,** 213–218 (1995).
23. Larsson, H. P., Baker, O. S., Dhillon, D. S. & Isacoff, E. Y. Transmembrane movement of the shaker K$^+$ channel S4. *Neuron* **16,** 387–397 (1996).
24. Starace, D. M. & Bezanilla, F. Histidine scanning mutagenesis of basic residues of the S4 segment of the Shaker K$^+$ channel. *J. Gen. Physiol.* **117,** 469–490 (2001).
25. Neale, E. J., Rong, H., Cockcroft, C. J. & Sivaprasadarao, A. Mapping the membrane-aqueous border for the voltage-sensing domain of a potassium channel. *J. Biol. Chem.* **282,** 37597–37604 (2007).
26. Tombola, F., Pathak, M. M. & Isacoff, E. Y. Voltage-sensing arginines in a potassium channel permeate and occlude cation-selective pores. *Neuron* **45,** 379–388 (2005).
27. Freites, J. A., Tobias, D. J. & White, S. H. A voltage-sensor water pore. *Biophys. J.* **91,** L90–L92 (2006).
28. Jogini, V. & Roux, B. Dynamics of the Kv1.2 voltage-gated K$^+$ channel in a membrane environment. *Biophys. J.* **93,** 3070–3082 (2007).
29. Treptow, W. & Tarek, M. Environment of the gating charges in the Kv1.2 Shaker potassium channel. *Biophys. J.* **90,** L64–L66 (2006).
30. Worcester, D. L. & Franks, N. P. Structural analysis of hydrated egg lecithin and cholesterol bilayers. II. Neutrol diffraction. *J. Mol. Biol.* **100,** 359–378 (1976).
31. Blasie, J. K., Schoenborn, B. P. & Zaccai, G. Direct methods for the analysis of lamellar neutron diffraction from oriented multilayers: a difference Patterson deconvolution approach. *Brookhaven Symp. Biol.* **27,** 58–67 (1976).
32. Jacobs, R. E. & White, S. H. The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochemistry* **28,** 3421–3437 (1989).
33. Wiener, M. C. & White, S. H. Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure. *Biophys. J.* **61,** 434–447 (1992).
34. Vamvouka, M., Cieslak, J., Van Eps, N., Hubbell, W. & Gross, A. The structure of the lipid-embedded potassium channel voltage sensor determined by double-electron-electron resonance spectroscopy. *Protein Sci.* **17,** 506–517 (2008).
35. Lee, A. G. Measurement of lipid-protein interactions in reconstituted membrane vesicles using fluorescence spectroscopy. *Methods Mol. Biol.* **27,** 101–107 (1994).
36. McIntosh, T. J. & Holloway, P. W. Determination of the depth of bromine atoms in bilayers formed from bromolipid probes. *Biochemistry* **26,** 1783–1788 (1987).
37. Dura, J. A. *et al.* AND/R: advanced neutron diffractometer/reflectometer for investigation of thin films and multilayers for the life sciences. *Rev. Sci. Instrum.* **77,** 074301 (2006).
38. Grossfield, A., Pitman, M. C., Feller, S. E., Soubias, O. & Gawrisch, K. Internal hydration increases during activation of the G-protein-coupled receptor rhodopsin. *J. Mol. Biol.* **381,** 478–486 (2008).
39. Ader, C. *et al.* Structural rearrangements of membrane proteins probed by water-edited solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **131,** 170–176 (2009).
40. Lindahl, E. & Sansom, M. S. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **18,** 425–431 (2008).
41. Phillips, R., Ursell, T., Wiggins, P. & Sens, P. Emerging roles for lipids in shaping membrane-protein function. *Nature* **459,** 379–385 (2009).
42. Huang, H. W. Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.* **50,** 1061–1070 (1986).
43. Nielsen, C., Goulian, M. & Andersen, O. S. Energetics of inclusion-induced bilayer deformations. *Biophys. J.* **74,** 1966–1983 (1998).
44. Goulian, M. *et al.* Gramicidin channel kinetics under tension. *Biophys. J.* **74,** 328–337 (1998).
45. Tabarean, I. V., Juranka, P. & Morris, C. E. Membrane stretch affects gating modes of a skeletal muscle sodium channel. *Biophys. J.* **77,** 758–774 (1999).
46. Laitko, U., Juranka, P. F. & Morris, C. E. Membrane stretch slows the concerted step prior to opening in a Kv channel. *J. Gen. Physiol.* **127,** 687–701 (2006).
47. Schmidt, D. & Mackinnon, R. Voltage-dependent K$^+$ channel gating and voltage sensor toxin sensitivity depend on the mechanical state of the lipid membrane. *Proc. Natl Acad. Sci. USA* **105,** 19276–19281 (2008).
48. Rosenbaum, D. M., Rasmussen, S. G. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459,** 356–363 (2009).
49. Krishnamurthy, H., Piscitelli, C. L. & Gouaux, E. Unlocking the molecular secrets of sodium-coupled transporters. *Nature* **459,** 347–355 (2009).
50. Erez, E., Fass, D. & Bibi, E. How intramembrane proteases bury hydrolytic reactions in the membrane. *Nature* **459,** 371–378 (2009).

©2009 Macmillan Publishers Limited. All rights reserved

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

Acknowledgements We thank the US National Institutes of Health (NIH) National Institute of Neurological Disorders and Stroke (NINDS) DNA sequencing facility for DNA sequencing, H. Jaffe in the NINDS protein sequencing facility for mass spectrometry and peptide sequencing and T. Kitaguchi for cloning K_vAP. We also thank T. Kimura, M. Mayer, M. Milescu, J. Mindell and S. Silberberg for discussions. This work was supported by the Intramural Research Programs of the NINDS, NIH (K.J.S.), and the National Institute on Alcohol Abuse and Alcoholism, NIH (K.G.); NIH grants GM74737 (S.H.W.) and Program Project GM86685 from the NINDS and the National Institute of General Medical Science (S.H.W., D.J.T.); and US National Science Foundation (NSF) grant CHE-0750175 (D.J.T.). We are grateful for allocation of computer time on the NSF-supported Teragrid resources provided by the Texas Advanced Computing Center, and the support of the National Institute of Standards and Technology, US Department of Commerce, in providing the neutron research facilities used for neutron diffraction experiments. The identification of any commercial product or trade name does not imply endorsement or recommendation by the US National Institute of Standards and Technology.

Author Contributions D.K. performed the biochemistry experiments; M.M., D.K. and D.L.W. performed the neutron diffraction experiments; D.K. and K.G. performed the solid-state NMR experiments; and J.A.F. and E.V.S. performed molecular dynamics simulations. All authors contributed to the study design and to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.H.W. (stephen.white@uci.edu) or K.J.S. (swartzk@ninds.nih.gov).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Voltage-sensing-domain expression, solubilization and purification. The K$_v$AP gene was amplified from *Aeropyrum pernix* genomic DNA using PCR and cloned into the pQE-60 vector (Qiagen). A pQE-60 plasmid containing the S1–S4 voltage-sensing domain of K$_v$AP (amino acid residues Met 1–Lys 147) was obtained by deletion of the carboxy-terminal region of K$_v$AP using PCR and its sequence was confirmed by DNA sequencing. The recombinant voltage-sensing domain was expressed in XL1-Blue strain of *Escherichia coli* as previously described[4]. The plasmid was transformed into the chemically competent *E. coli* cells (subcloning grade; Stratagene). One colony was inoculated into 100 ml LB broth supplemented with ampicillin (100 μg ml^{-1}) and grown overnight at 37 °C with continuous shaking at 200 r.p.m. One litre of LB–ampicillin broth was inoculated with 10 ml of the starting culture and the protein expression was induced with 1 mM isopropyl-β-D-thiogalactopyranoside (Calbiochem) when the absorbance of the cells at 600 nm reached 0.6. After 3 h of induction, cells were harvested by centrifugation at 5,000*g* for 20 min.

Cells were resuspended in 10 mM EDTA solution and collected by centrifugation, and then twice resuspended in 20 mM Tris/HCl, 100 mM KCl, pH 7.8 and collected by centrifugation. Cells were then resuspended in 40 ml of 20 mM Tris/HCl, 100 mM KCl, pH 7.8 buffer, supplemented with 100 μl of protease inhibitor cocktail (Sigma) and 100 μl of 26 mg ml^{-1} PMSF in isopropanol. Cells were sonicated for 5 min on ice, 50 μl of protease inhibitor cocktail was added, and the S1–S4 domain was extracted by solubilization of the homogenate in 2.5% (w/v) decylmaltoside (Anatrace) in 20 mM Tris/HCl, 100 mM KCl, pH 7.8. Lysate was then centrifuged at 100,000*g* for 1 h at 4 °C and the supernatant collected. Supernatant was mixed with Co-TALON resin (Clontech), the mixture was transferred to a chromatography column (Bio-Rad) and the solution was allowed to pass through. The resin with bound protein was washed with 0.25% (w/v) decylmaltoside, 10 mM imidazole in 20 mM Tris/HCl, 100 mM KCl, pH 7.8. Buffer was then exchanged for 3% (w/v) *n*-octyl-β-D-glucopyranoside (OG) in 20 mM Tris/HCl, 100 mM KCl, pH 7.8 and the protein eluted with 400 mM imidazole in the same buffer. One unit of thrombin (Sigma) was added per milligram of protein and the mixture was dialyzed against 3% (w/v) OG in 20 mM Tris/HCl, 100 mM KCl, pH 7.8 overnight at 4 °C in a dialysis cassette with a molecular-weight cut-off of 10 kDa (Pierce).

The protein was analysed by SDS–PAGE electrophoresis, followed by Coomassie staining (Invitrogen) and by MALDI-MS using the Invitrosol MALDI protein solubilizer kit (Invitrogen). The efficiency of the thrombin cleavage and the His-tag removal during the dialysis was confirmed using the His-Probe HRP reagent kit (Pierce) and by MALDI-MS. S1–S4 K$_v$AP protein identity was confirmed by mass spectrometric analysis of protein digest fragments and N-terminal Edman sequencing. Amino-terminal sequencing revealed that the first five amino acids are removed during protein expression, consistent with earlier observations[4]. The uniformly deuterated protein was obtained by expression in BioExpress media (Cambridge Isotope Laboratories) supplemented with 80% ^2H_2O, and the molecular weight of the purified protein was determined by MALDI-MS using the Invitrosol MALDI protein solubilizer kit (Invitrogen). The concentration of the protein was determined spectrophotometrically using an extinction coefficient ($\varepsilon_{280\,\mathrm{nm}} = 17{,}210\ \mathrm{M}^{-1}\,\mathrm{cm}^{-1}$) calculated from the deduced protein composition[51].

Lipid reconstitution of the voltage-sensing domain of K$_v$AP. The S1–S4 voltage-sensing domain of K$_v$AP were reconstituted to different molar ratios of protein and lipid as previously described[14,52–54], using a 1:1 mixture of POPC and POPG. All lipids were purchased from Avanti Polar Lipids and mixtures were dried from solution in chloroform under a stream of nitrogen gas and desiccated under vacuum overnight. Lipid films were solubilized in 20 mM Tris/HCl, 100 mM KCl buffer, pH 7.8 with 3% (w/v) OG, and protein was added to the lipid to form mixed detergent–lipid micelles. Proteoliposomes were formed by rapid dilution of the mixed protein–detergent–lipid micelles well below the critical micelle concentration of the OG detergent. Proteoliposome pellets were collected by ultracentrifugation at 200,000*g* at 4 °C using an Optima TL 100 TLX ultracentrifuge and 100.3 TLA rotor (Beckman). Proteoliposome pellets were resuspended in H_2O and sedimented by ultracentrifugation, resuspended and mildly sonicated for 1 min in a water-bath sonicator. The resultant proteoliposomes were analysed for lipid content using the method of ref. 55, and residual detergent contents were determined using the modified phenol-sulphuric acid assay[56] and by dissolving aliquots of the sample in deuterated MeOH and analysis of the components of the mixture by ^{1}H NMR.

Circular dichroism spectroscopy. Circular dichroism spectra were recorded in 20 mM Tris/HCl, 100 mM KCl buffer, pH 7.8, using a JASCO J-815 spectropolarimeter equipped with a thermally controlled cuvette holder. Spectra were recorded on voltage-sensing-domain samples in 0.1–1-mm quartz cuvettes, from 180 nm to 250 nm with 1-nm step resolution and 4-s integration time. The helix content of the protein sample was calculated following ref. 57, and indicated that the protein had high (~85%) helical content both in OG micelles and when reconstituted in lipid, consistent with the X-ray structure of the S1–S4 domain of K$_v$AP[4] and EPR results on the S1–S4 domain and full-length K$_v$AP channel[9,13].

Fluorescence spectroscopy. Fluorescence emission spectra for Trp 70 within the S2 helix of the voltage-sensing domain of K$_v$AP were recorded for the protein in either OG micelles or when reconstituted into lipid. (Trp 70 is the only Trp residue within S1–S4.) Fluorescence spectra were recorded in 20 mM Tris/Cl, 100 mM KCl, pH 7.4 at 25 °C with stirring in a total volume of 2 ml using the SPEX FluoroMax 3 spectrofluorometer. Quartz 1 cm × 1 cm cuvettes were used for all fluorescence measurements. An excitation wavelength of 295 nm (5-nm band pass) was used and the emission was scanned between 300 and 400 nm (5-nm band pass) with an increment of 0.5 nm. The polarizer was configured to excitation-90°, emission-0° (ref. 58) and emission spectra of OG buffer or lipid alone were subtracted. Quenching of Trp 70 fluorescence was examined by titration with acrylamide, an aqueous quencher of Trp fluorescence. Stern–Volmer quenching constants, K_{SV}, were calculated from the best fits of $F_0/F = 1 + K_{SV}[Q]$, where F_0 and F are the fluorescences of the Trp 70 in the absence and presence of a quencher, respectively, and $[Q]$ is the concentration of the quencher. To determine the disposition of the Trp 70 in model membranes, we compared quenching by bromine atoms attached to different positions on the hydrocarbon tail[35,36,59]. For these experiments, protein was reconstituted into proteoliposomes using a 1:1 mixture of POPG and either Br$_2$(6,7)-PC (1-palmitoyl-2-stearoyl(6-7)dibromo-*sn*-glycero-3-phosphatidylcholine, C16:0, C18:0), Br$_2$(9,10)-PC (1-palmitoyl-2-stearoyl(9-10)dibromo-*sn*-glycero-3-phosphatidylcholine, C16:0, C18:0) or Br$_2$(11,12)-PC (1-palmitoyl-2-stearoyl(11-12)dibromo-*sn*-glycero-3-phosphatidylcholine, C16:0, C18:0).

Determining structure from neutron diffraction data and deuterium contrast variation. Lamellar diffraction patterns yield trans-bilayer distributions of scattering length projected onto the bilayer normal (*z* axis), which we call bilayer profiles and write as $\rho(z)$. The profiles presented here have been placed on the absolute per-lipid scale. The simplest profiles are obtained by Fourier transformation of the measured structure factors $F_M(h) = \sqrt{I_h}$, where I_h is the corrected intensity of the *h*th diffracted intensity. In this case, $\rho(z)$ varies along the bilayer normal and has an average value of zero when integrated over the unit cell defined by the Bragg spacing, *d*. The amplitude of $\rho(z)$ is arbitrary, determined only by the units used to measure intensities, such as neutron counts observed in a given time period. This simple approach provides limited information about the disposition of molecules dissolved within the bilayer.

Useful information can be obtained only when the profiles are placed on an absolute scale, meaning that the average value of $\rho(z)$, ρ_0, corresponds to the total scattering length of the unit cell and that the variation of $\rho(z)$ around ρ_0 shows absolute changes in scattering-length density. To determine ρ_0, the contents of the unit cell (lipid, water and protein) must be known. To calibrate the variation of $\rho(z)$, an isomorphous substitution of atoms of known scattering length, b_{sub}, must be introduced into the sample. In this case, the integral over the unit cell of the 'difference profile', $\Delta\rho(z) \equiv \rho_{\mathrm{sub}}(z) - \rho(z)$, must equal b_{sub}. This procedure, described in detail in refs 32, 60, 61, yields instrumental constants, $k(h)$, that lead to absolute-scale structure factors, $F(h) = k(h)F_M(h)$. For a centrosymmetric bilayer containing two lipids per unit cell, the average scattering-length density is given by $\rho_0 = (2/Sd)\sum b_i$, where S is the area per lipid and $\sum b_i$ is the sum of the scattering lengths of all of the atoms in the unit cell. The value of S is rarely known. To circumvent this problem, we can use the 'relative absolute scale'[32], in which $\rho^*(z) = S\rho(z)$. A better and more descriptive term is the per-lipid scale, because the scattering-length density describes the scattering length per lipid rather than per unit volume. To use this scale, we need to know only the average numbers per lipid of water molecules and other components in the unit cell. As stated, the profiles here have been placed on the absolute per-lipid scale.

The absolute per-lipid scattering-length density is given by

$$\rho * (z) = \rho *_0 + \frac{2}{d}\sum_{h=1}^{h_{\max}} \varphi(h)|F(h)|\cos\left(\frac{2\pi hz}{d}\right) \qquad (1)$$

where $\varphi(h)$ is the sign of the absolute-scale structure factor, $F(h)$ (whose absolute value is $|F(h)|$), and $h_{\max}$ indexes the highest observable structure factor. The unit cell is centrosymmetric; consequently, in the data presented here $\varphi(h) = \pm 1$ (that is, $\cos(0)$ or $\cos(\pi)$). Methods for determining the signs have been discussed in detail elsewhere[30,31].

Treatment of diffraction data. In the kinematical approximation, the structure factors, $F(h)$, at all observed diffraction orders are determined as the square root of the integrated peak intensities, after background correction, absorption correction (B) and Lorentz-factor correction ($\sin(2\theta_h)$), where θ_h denotes the angle

©2009 Macmillan Publishers Limited. All rights reserved

of incidence corresponding to the *h*th order of diffraction. The absorption correction was calculated as follows, where *t* is the sample thickness and α is the linear absorption coefficient[30]: $B(h) = \sin(\theta_h)/(2\alpha t)[1 - \exp(-2\alpha t/\sin(\theta_h))]$. We determined phases by measuring a sample in a series of different ^1H_2O–^2H_2O contrasts (that is, using different mole fractions of ^2H_2O in the salt solutions for hydrating the sample). Assuming a Gaussian distribution of water hydrating the lipid head groups[60], the difference structure factors corresponding to a lipid bilayer hydrated with ^2H_2O (F_D) and, respectively, ^1H_2O (F_H) can be modelled according to

$$\Delta F(h) = k_D F_D(h) - k_H F_H(h) = x_D \exp[-(\pi h A/d)^2] \cos(2\pi h Z/d) \qquad (2)$$

where *A* and *Z* respectively denote the 1/e half-width and the mean position of the Gaussian describing the labelled component distribution (water in this case), and k_D and k_H are scale factors. The prefactor x_D scales with the amount of deuterium per lipid: $x_D = 2(b_D - b_H)f_D n_D$, where b_H and b_D are the scattering lengths of deuterium and hydrogen, respectively, f_D is the fraction of deuterated component and n_D is the number of deuterium atoms per molecule in the labelled component. The cosine factor in equation (2) determines the slope of the linear dependence, $\Delta F(h) = f(x_D)$, and thus the phases $\varphi(h)$.

To determine the trans-bilayer distributions of water, the -CH$_2$-CH$_2$- group of the phosphorylcholine (PC), and the S1–S4 domain of K$_v$AP, we substituted those molecular components for their deuterated counterparts, for both protein-containing or pure-lipid samples, and compared their density profiles with those of protonated samples using the absolute per-lipid scale.

Scaling of the neutron data and determination of the amount of water per lipid. Because the raw diffraction intensities collected are not normalized, the absolute (per-lipid) scale is determined on the basis of the sample composition (for example protein concentration and amount of water). The protein concentration was determined from ultraviolet absorbance at 280 nm, but the amount of water accumulated in the protein-containing membranes at the relative humidity of our experiments is unknown and has to be determined by additional experiments. We used a lipid-deuteration scheme (manuscript in preparation) that includes deuteration of the water of hydration (^2H_2O) and the PC-C$^2H_2$-C$^2H_2$- group (D4 lipid) to resolve the absolute scale and the number of waters per lipid. The two homologous samples (protonated and D4 lipid) were each measured under at least two different ^1H_2O–^2H_2O contrast conditions by exchanging ^1H_2O with ^2H_2O in different proportions (for example $^1H_2O/^2H_2O$ ratios of 100:0, 50:50 and 80:20) in the saturated salt solutions used in the chamber. Both the water and the D4 lipid are components that can be described by Gaussian distributions[33,60]. In equation (2), we compare the data from the protonated- and D4-lipid homologous samples, measured in either ^1H_2O or 20% ^2H_2O, to determine the parameters describing the D4-lipid distribution. Knowing the prefactor x_D, we determine the position *Z* and width *A*, as well as the scale factors k_H (protonated lipid) and k_D (D4 lipid), by a least-squares minimization procedure. Once scaled, the data collected in ^1H_2O and 20% ^2H_2O for a given sample are compared in equation (2) to determine the water distribution parameters and the number of water molecules per lipid, represented by the prefactor x_D in equation (2). The water content of neat-POPC multilayers determined using this approach is indistinguishable from that determined by independent methods[62].

Molecular dynamics simulations. Two simulation systems with 9 and 11 water molecules per lipid, respectively, and 130 lipid molecules per protein (corresponding to relative humidities of 86% and 93% and a protein/lipid molar ratio of 0.77 mol%) were prepared from the end configuration of a simulation trajectory of the S1–S4 domain of K$_v$AP in a POPC bilayer with excess water[27]. The initial atomistic model in the excess-water simulation corresponded to residues 24 to 147 in the model of the K$_v$AP full channel proposed in ref. 63. The pore domain of the full-channel model provides an unambiguous constraint for the orientation of the S1–S4 domain in the lipid bilayer. The final placement of the protein in the lipid bilayer along the transmembrane direction was determined by ensuring that the five Tyr side chains in the S2–S3 connecting turn and the S3–S4 end were simultaneously in contact with the head-group region on opposite sides of the lipid bilayer. Further details of the set-up of the excess-water simulation system and the generation of molecular dynamics trajectories can be found in ref. 27.

The low-hydration simulation systems consisted of two stacked lipid bilayers, each containing a single S1–S4 domain, arranged to form a single pseudo-centrosymmetric unit cell. The system with nine water molecules per lipid was prepared by removing the necessary water and lipid molecules from the end configuration of the excess-water simulation. The initial equilibration consisted of 1,000 steps of energy minimization followed by a 1-ns molecular dynamics run at constant volume and constant temperature (295 K), with the protein backbone held fixed. The full simulation was then carried out at a constant temperature of 295 K and a constant pressure of 1 atm. The protein was progressively released from its initial configuration over the first 5.5 ns using harmonic restraints. The simulation was carried out in the absence of restraints for 37.5 ns. The system with 11 water molecules per lipid was prepared from the end configuration of the system with nine water molecules per lipid by adding the necessary number of water molecules. The initial equilibration consisted of 1,000 steps of energy minimization followed by a 20-ps run at constant volume and constant temperature (295 K) over the newly added waters, and an 80-ps run over the whole system. The full simulation was then carried out at a constant temperature of 295 K and a constant pressure of 1 atm for 25.2 ns.

All of the molecular dynamics trajectories were generated with the NAMD 2.6 software package[64]. The CHARMM22 and the revised CHARMM27 force fields[65–67] were used for the peptide and the lipids, respectively, and the TIP3P model was used for water[68]. The smooth particle mesh Ewald (PME) method[69,70] was used to calculate electrostatic interactions, and the short-range real-space interactions were cut off at 11 Å, using a switching function. A reversible multiple-time-step algorithm[71] was used to integrate the equations of motion with time steps of 4 fs for electrostatic forces, 2fs for short-range non-bonded forces and 1 fs for bonded forces. All bond lengths involving hydrogen atoms were held fixed using the SHAKE and SETTLE algorithms. A Langevin dynamics scheme was used for thermostatting. Nose–Hoover–Langevin pistons were used for pressure control[72,73]. Molecular graphics and simulation analyses were performed with the VMD 1.8.6 software package[74] over the last 10 ns of each simulation.

To compare simulations with 11 water molecules per lipid directly with the experimental data, neutron diffraction structure factors[75] for the *n*th order of diffraction, *F*(*n*), were computed from the molecular dynamics trajectory according to

$$F(n) = \sum_{j}^{\text{cell}} b_j \exp(2\pi i n z_j/d) \qquad (4)$$

where the sum is over all the atoms in the simulation unit cell; b_j and z_j are the neutron scattering length and *z* coordinate of the *j*th atom, respectively; the Bragg spacing, *d*, is taken to be half of the simulation cell length along the transmembrane direction; and *n* is an integer. The scaling factor of half the simulation cell length for the spatial coordinates corresponds to the repeat distance (Bragg spacing) of an oriented stack of bilayers. The oriented bilayers diffract as centrosymmetric objects independent of the presence of the protein. The purpose of the double-bilayer simulation system is to model the two equally probable orientations of the protein in the lipid bilayer. Therefore, because the total scattering length of a single simulation cell is twice that of a single repeat unit in the diffraction experiment, each atom in the simulation cell is considered to have an occupation factor of one-half. Structure factors were averaged over ten system configurations (one per nanosecond of trajectory time), and the total scattering-density profiles were constructed from the structure factors exactly as in the analysis of the experimental diffraction data (see equation (1)). Component densities were computed following the experimental protocol (see equation (2)), assuming uniform labelling at the same mole fraction as in the neutron diffraction experiments. The average length of the simulation cell in the dimension perpendicular to the membranes was 104 Å, corresponding to a spacing of $d = 51.8$ Å for a single bilayer containing the S1–S4 voltage-sensing domain of K$_v$AP, in excellent agreement with the experimental value of 52 Å obtained at 93% relative humidity (corresponding to 11 water molecules per lipid).

The electrostatic potential in the excess-hydration simulation was calculated using the linearized Poisson–Boltzmann theory, treating all the system components as linear, isotropic dielectrics under an applied potential difference across the membrane, as previously described[76–78]. For a given configuration along the simulation trajectory, the electrostatic potential was calculated over a composite system consisting of a cuboid region of space (dashed rectangle in Fig. 5b) containing the atomistic configurations of the protein and most of the lipids. This region was considered to be embedded in a continuum composed of a semi-infinite planar slab, representing the membrane, between two half-spaces that represent the electrolyte solution[77]. The calculations were performed over the last 16 ns of the simulation trajectory, taking one configuration per nanosecond, using the PBEQ module of the CHARMM 32a2 software package[79]. The linearized Poisson–Boltzmann equation was solved by finite differences, using the successive over-relaxation method, over a cubic grid of 161 nodes with a grid spacing of 1 Å. The continuum slab thickness was set to equal the separation between carbonyl distributions in the atomistic system. A dielectric constant of two was assigned to lipids and protein. The solvent dielectric constant was set to 80 and the salt concentration was set to 150 mM. The molecular surface was used to define the atomistic dielectric boundaries using the van der Waals radii from the CHARMM force field[65].

NMR saturation transfer difference using magic-angle spinning. Saturation transfer difference experiments[80] were used to study hydration of the voltage-sensing domains in membranes. Magic-angle-spinning (MAS) conditions were

©2009 Macmillan Publishers Limited. All rights reserved

used as previously described[38,81] to resolve lipid resonances. ^{1}H NMR spectra of lipids were recorded and resonance attenuation measured in response to radio-frequency pulses. The saturating radio-frequency pulses (field strength, $(\gamma/2\pi)B_1 = 0$–1.2 kHz) consisted of twenty Gaussian-shaped 50-ms pulses. The saturation frequency was set to the amide region of the protein (8.5 p.p.m.) or the water resonance (4.79 p.p.m.). The attenuation of the lipid methylene signal, defined as resonance amplitude recorded without saturation divided by the amplitude recorded with saturation, was followed as indicator of magnetization transfer to lipid. The proteoliposomes were packed into 4-mm MAS rotors (Bruker) and hydrated with either ^2H_2O or ^1H_2O to the final water/lipid ratio of 30:1. Sixteen scans with a recycle delay of 10 s were acquired at 295.1 K. All spectra were recorded on an 800-MHz Bruker AV800 spectrometer equipped with a 4-mm ^{1}H/^{13}C/^{2}H CP-MAS probe (Bruker BioSpin) at a MAS frequency of 10 kHz.

51. Gill, S. C. & von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182,** 319–326 (1989).
52. Heginbotham, L., Kolmakova-Partensky, L. & Miller, C. Functional reconstitution of a prokaryotic K$^+$ channel. *J. Gen. Physiol.* **111,** 741–749 (1998).
53. Perozo, E., Cortes, D. M. & Cuello, L. G. Three-dimensional architecture and gating mechanism of a K$^+$ channel studied by EPR spectroscopy. *Nature Struct. Biol.* **5,** 459–469 (1998).
54. Cuello, L. G., Romero, J. G., Cortes, D. M. & Perozo, E. pH-dependent gating in the *Streptomyces lividans* K$^+$ channel. *Biochemistry* **37,** 3229–3236 (1998).
55. Hurst, R. O. The determination of nucleotide phosphorus with a stannous chloride-hydrazine sulphate reagent. *Can. J. Biochem. Physiol.* **42,** 287–292 (1964).
56. Dubois, M., Gilles, K., Hamilton, J. K., Rebers, P. A. & Smith, F. A colorimetric method for the determination of sugars. *Nature* **168,** 167 (1951).
57. Chen, Y. H., Yang, J. T. & Chau, K. H. Determination of the helix and beta form of proteins in aqueous solution by circular dichroism. *Biochemistry* **13,** 3350–3359 (1974).
58. Ladokhin, A. S., Jayasinghe, S. & White, S. H. How to measure and analyze tryptophan fluorescence in membranes properly, and why bother? *Anal. Biochem.* **285,** 235–245 (2000).
59. Abrams, F. S. & London, E. Extension of the parallax analysis of membrane penetration depth to the polar region of model membranes: use of fluorescence quenching by a spin-label attached to the phospholipid polar headgroup. *Biochemistry* **32,** 10826–10831 (1993).
60. Wiener, M. C., King, G. I. & White, S. H. Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. I. Scaling of neutron data and the distributions of double bonds and water. *Biophys. J.* **60,** 568–576 (1991).
61. Wiener, M. C. & White, S. H. Transbilayer distribution of bromine in fluid bilayers containing a specifically brominated analogue of dioleoylphosphatidylcholine. *Biochemistry* **30,** 6997–7008 (1991).
62. Hristova, K. & White, S. H. Determination of the hydrocarbon core structure of fluid dioleoylphosphocholine (DOPC) bilayers by X-ray diffraction using specific bromination of the double-bonds: effect of hydration. *Biophys. J.* **74,** 2419–2433 (1998).
63. Lee, S. Y., Lee, A., Chen, J. & MacKinnon, R. Structure of the KvAP voltage-dependent K$^+$ channel and its dependence on the lipid membrane. *Proc. Natl Acad. Sci. USA* **102,** 15441–15446 (2005).
64. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).
65. MacKerell, A. D. Jr *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).
66. Feller, S. E. & MacKerell, A. D. Jr. An improved empirical potential energy function for molecular simulations of phospholipids. *J. Phys. Chem. B* **104,** 7510–7515 (2000).
67. Klauda, J. B., Brooks, B. R., MacKerell, A. D. Jr, Venable, R. M. & Pastor, R. W. An ab initio study on the torsional surface of alkanes and its effect on molecular simulations of alkanes and a DPPC bilayer. *J. Phys. Chem. B* **109,** 5300–5311 (2005).
68. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).
69. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N*.log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
70. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).
71. Grubmüller, H., Heller, H., Windemuth, A. & Schulten, K. Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol. Simul.* **6,** 121–142 (1991).
72. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant-pressure molecular-dynamics algorithms. *J. Chem. Phys.* **101,** 4177–4189 (1994).
73. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **103,** 4613–4621 (1995).
74. Humphrey, W., Dalke, W. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
75. Benz, R. W., Castro-Román, F., Tobias, D. J. & White, S. H. Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophys. J.* **88,** 805–817 (2005).
76. Roux, B. Influence of the membrane potential on the free energy of an intrinsic protein. *Biophys. J.* **73,** 2980–2989 (1997).
77. Roux, B. The membrane potential and its representation by a constant electric field in computer simulations. *Biophys. J.* **95,** 4205–4216 (2008).
78. Grabe, M., Lecar, H., Jan, Y. N. & Jan, L. Y. A quantitative assessment of models for voltage-dependent gating of ion channels. *Proc. Natl Acad. Sci. USA* **101,** 17640–17645 (2004).
79. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).
80. Forsen, S. H. A. Study of moderately rapid chemical exchange reactions by means of nuclear magnetic double resonance. *J. Chem. Phys.* **39,** 2892–2901 (1963).
81. Soubias, O. & Gawrisch, K. Probing specific lipid-protein interaction by saturation transfer difference NMR spectroscopy. *J. Am. Chem. Soc.* **127,** 13110–13111 (2005).

©2009 Macmillan Publishers Limited. All rights reserved

Enrichment by supernovae in globular clusters with multiple populations

Jae-Woo Lee[1], Young-Woon Kang[1], Jina Lee[1] & Young-Wook Lee[2]

The most massive globular cluster in the Milky Way, ω Centauri, is thought to be the remaining core of a disrupted dwarf galaxy[1,2], as expected within the model of hierarchical merging[3,4]. It contains several stellar populations having different heavy elemental abundances supplied by supernovae[5]—a process known as metal enrichment. Although M 22 appears to be similar to ω Cen[6], other peculiar globular clusters do not[7,8]. Therefore ω Cen and M 22 are viewed as exceptional, and the presence of chemical inhomogeneities in other clusters is seen as 'pollution' from the intermediate-mass asymptotic-giant-branch stars expected in normal globular clusters[9]. Here we report Ca abundances for seven globular clusters and compare them to ω Cen. Calcium and other heavy elements can only be supplied through numerous supernovae explosions of massive stars in these stellar systems[10], but the gravitational potentials of the present-day clusters cannot preserve most of the ejecta from such explosions[11]. We conclude that these globular clusters, like ω Cen, are most probably the relics of more massive primeval dwarf galaxies that merged and disrupted to form the proto-Galaxy.

The Sejong/ARCSEC Ca uvby survey programme was initiated in 2006 to investigate the homogenous metallicity scale for globular clusters and to obtain the complete metallicity distribution function of the Galactic bulge using the hk index (ref. 12): $hk = (Ca - b) - (b - y)$. The Ca filter in the hk index measures ionized calcium H and K lines, which have been frequently used to calibrate metallicity scale for globular clusters[13,14]. The utility of the hk index is that it is known to be about three times more sensitive to metallicity than the m_1 index is, for stars more metal-poor than the Sun, and it has half the sensitivity of the m_1 index to interstellar reddening[12]. During the past three years, we have used more than 85 nights of CTIO 1.0-m telescope time for this project. The telescope was equipped with an STA 4k × 4k CCD camera, providing a plate scale of 0.289 arcsec per pixel and a field of view of 20 × 20 arcmin. All of our targets accompanied with standards were observed under the photometric weather conditions and most of targets were repeatedly visited between separate runs. The photometry of our targets and standards were analysed using DAOPHOT II, ALLSTAR and ALLFRAME[15,16].

In the course of metallicity calibration of red giant branch (RGB) stars in globular clusters, we found that many such clusters show a split in the RGB in their hk versus V colour–magnitude diagrams (Figs 1 and 2). The prime examples are M 22 and NGC 1851. In particular, the double RGB sequence in M 22 is very intriguing. The differential reddening effect and the contamination from the off-cluster populations cannot explain the double RGB sequences in M 22 (see Supplementary Information). It has been debated for decades whether this cluster is chemically inhomogeneous or not, but the recent high resolution spectroscopic study of 17 RGB stars in the cluster suggests that it contains chemically inhomogeneous subpopulations[6]. The bimodality in the m_1 index of M 22 RGB stars was also known, but it has been argued that it is most probably due to the bimodal CN abundances (where CN absorption strengths strongly affect the m_1 index), and not due to the bimodal distribution of heavy elements in the cluster[17–19]. The star-to-star light elemental abundance (C, N, O, Na, Mg and Al) variations have been known for decades and they are now generally believed to be resulted from chemical pollution by intermediate-mass asymptotic giant branch stars[9] or fast rotating massive stars[20]. However, it should be emphasized that our hk measurements for RGB stars in M 22, NGC 1851 and other globular clusters show discrete or bimodal distributions in calcium abundance, which cannot be supplied by intermediate-mass asymptotic giant branch stars or fast rotating massive stars.

As shown in Fig. 3, the difference in calcium, silicon, titanium and iron abundances between the calcium weak (Ca-w hereafter) group with smaller hk index and the calcium strong (Ca-s hereafter) group with larger hk index in M 22 and NGC 1851 suggests that they are indeed chemically distinct[21–24]. (It is not shown in the figure but europium also has a bimodal abundance distribution in M 22, in the sense that the Ca-s group has a higher europium abundance.) As for the origin of chemical inhomogeneity in globular clusters, at least four viable chemical enrichment mechanisms have been proposed up to date. They are, in the order of time required to emerge; (1) fast rotating massive stars, (2) type II supernovae, (3) intermediate-mass asymptotic

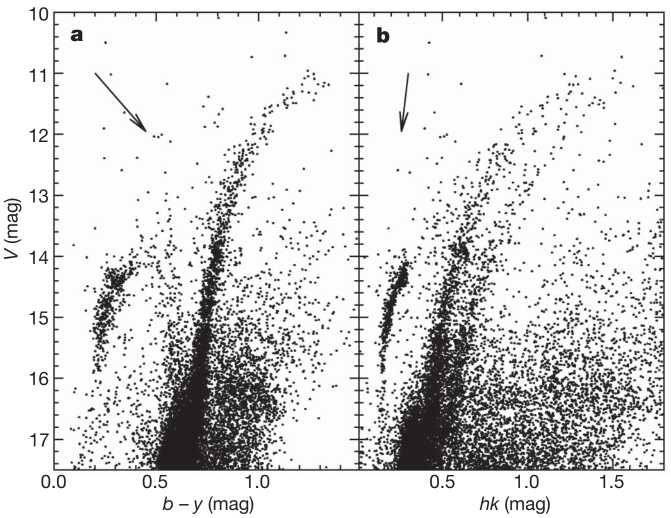


Figure 1 | Colour–magnitude diagrams for M 22. **a**, V versus $b - y$; **b**, V versus hk. In **b**, we note the distinct and discrete double RGB sequences. This cannot be due to a differential reddening effect across the cluster or contamination from the off-cluster field, but is due to the difference in calcium abundance, which was synthesized in supernovae, between the two RGB sequences. The number ratio between the Ca-w group with smaller hk index and the Ca-s group with larger hk index is about 70:30. Black arrows in each panel denote reddening vectors.

[1]Department of Astronomy and Space Science, ARCSEC, Sejong University, Seoul 143-747, Korea. [2]Center for Space Astrophysics, Yonsei University, Seoul 120-749, Korea.

©2009 Macmillan Publishers Limited. All rights reserved

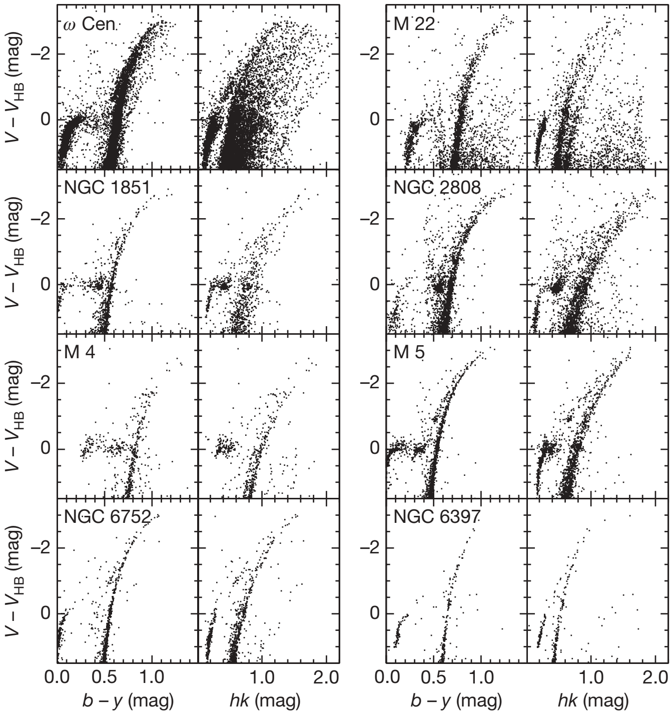


Figure 2 | Colour–magnitude diagrams for ω Cen, M 22, NGC 1851, NGC 2808, M 4, M 5, NGC 6752 and NGC 6397. Note that, while the distributions of the RGB sequences in the $b - y$ colour are relatively narrow, those in the hk index are either discrete or broad. This is evidence for the multiple stellar populations with distinct calcium abundances. Among these globular clusters, NGC 6397 appears to be the only normal globular cluster with a simple population (that is, coeval and monometallic).

giant branch stars, and (4) type Ia supernovae. If the current understanding of supernovae explosions is correct, only type Ia and II supernovae can supply the heavy elements such as calcium and iron[10]. To explain the discrete calcium abundances seen in M 22 and NGC 1851, however, the contribution from type Ia supernovae can be ruled out for two reasons. First, the longer timescale ($\gtrsim$1–2 Gyr) before the onset of type Ia supernova explosions, which would produce detectable age spread between two populations; and second, the enhanced α-elemental abundances, indicative of absence of contributions from type Ia supernovae[10]. Qualitatively, the differences in elemental abundances between the two stellar populations in M 22 and NGC 1851 can be naturally explained by invoking chemical enrichment by type II supernovae, where α-elements (silicon, calcium, and titanium) and r-process elements (such as europium) are dominantly produced. However, our results do not necessarily imply that type II supernovae are solely responsible for the chemical enrichment in M 22 and NGC 1851, since all four above-mentioned mechanisms may be involved. We emphasize that the crux of our results is the undeniable evidence for type II supernova contributions to chemical enrichment of some globular clusters, in sharp contrast to the widely accepted idea of chemical pollution only by intermediate-mass asymptotic giant branch stars or fast rotating massive stars, with which the chemical enhancement of the α- and r-process elements in the second generation of the stars cannot be easily explained.

More than half of the 37 globular clusters in our sample shows discrete or broad distributions of the hk index in their RGB sequences. In Fig. 2, we show colour–magnitude diagrams for some of the exemplary globular clusters in the order of hk widths of RGB sequences at V_{HB}, the V magnitude level at the horizontal branch: ω Cen, M 22, NGC 1851, NGC 2808, M 4, M 5, NGC 6752 and NGC 6397 (see also Supplementary Table 3 and Supplementary Figs 6–13). NGC 2808 is known to have multiple main-sequences but no multiple RGB sequences have been reported to date. Our new results

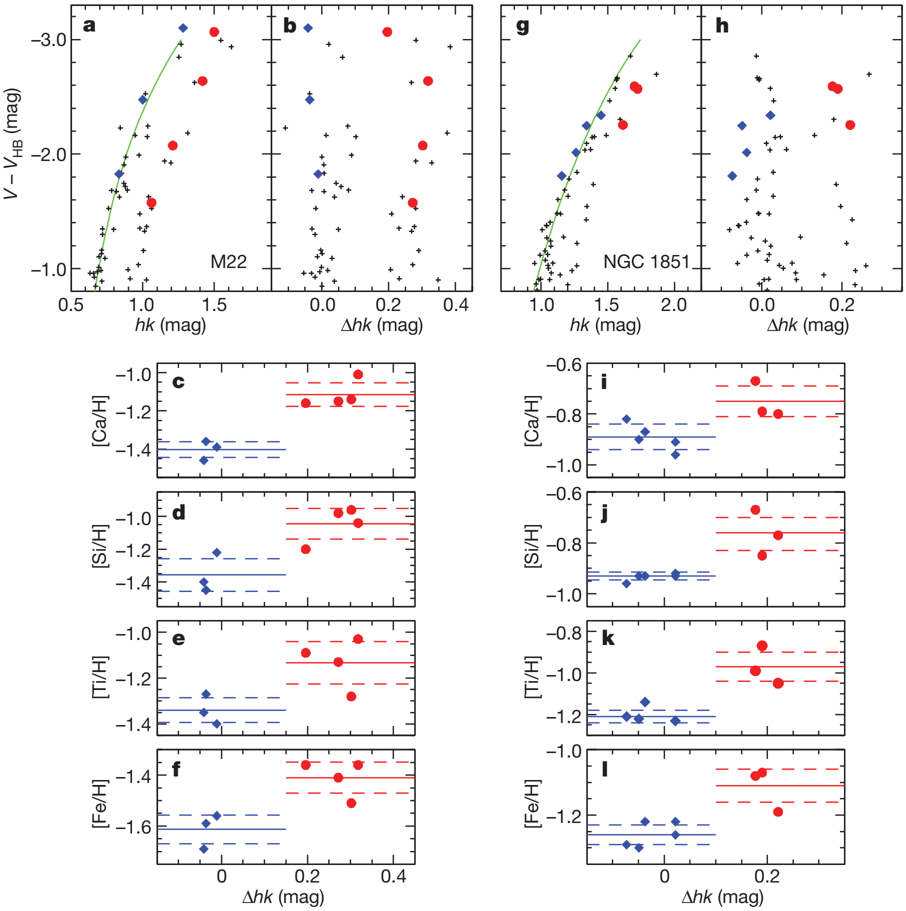


Figure 3 | Differences in chemical compositions between double RGB sequences in M 22 and NGC 1851. a, b, Black 'plus' signs denote stars in M 22 with proper motion membership probabilities $P \geq 90\%$; blue filled diamonds and red filled circles denote RGB stars studied with high-resolution spectroscopy in the Ca-w and the Ca-s groups, respectively[23,24]. The green solid line denotes the fiducial sequence of RGB stars and Δhk denotes the difference in the hk index against the fiducial sequence. The double RGB sequences persist in proper motion member stars. **c–f**, Comparisons of elemental abundances between the Ca-w and the Ca-s groups in M 22. Solid lines denote the mean values, and dashed lines denote standard deviations of each group. The Ca-s group has higher α-element (Si, Ca and Ti) and iron abundances, which must be supplied by numerous type II supernova explosions. **g, h**, Black 'plus' signs denote stars in NGC 1851; blue filled diamonds and red filled circles denote RGB stars studied with high-resolution spectroscopy in the Ca-w and the Ca-s groups, respectively[21,22]. **i–l**, As **c–f** but for NGC 1851. In the figure, we adopt the standard spectroscopic notation that for the element X with respect to hydrogen, $[\mathrm{X/H}] \equiv \log_{10}(N_X/N_H)_{\mathrm{star}} - \log_{10}(N_X/N_H)_{\mathrm{Sun}}$ (here N indicates number of atoms).

©2009 Macmillan Publishers Limited. All rights reserved

show that NGC 2808 shows at least two discrete RGB sequences with a large spread in calcium abundance. Similarly, M 5 has a very broad *hk* index in the RGB sequence and NGC 6752 shows discrete RGB sequences. It is interesting to note that all the globular clusters with signs of multiple stellar populations have relatively extended horizontal branches, while the globular clusters with normal horizontal branches (for example, NGC 6397 in Fig. 2 and Supplementary Fig. 13) show no spread or split in RGB. This is consistent with the suggestion that the extended horizontal branch is a signal of the presence of multiple stellar populations in globular clusters[25].

The overwhelming problem of the chemical enrichment by type II supernovae in globular clusters is that their ejecta are considered to be too energetic to be retained by less massive systems like typical Galactic globular clusters (≤a few times 10^5 solar masses)[11]. Our results therefore suggest that M 22, NGC 1851 and other globular clusters with a split RGB were much more massive in the past, unless the current understanding of supernovae explosions is greatly in error. Perhaps these globular clusters were once nuclei of dwarf-galaxy-like fragments and then accreted and dissolved in the Milky Way, as is widely accepted for ω Cen[1,2,26]. Recent calculations suggest that a massive (≥a few times 10^6 solar masses) star cluster embedded in a proto-dwarf galaxy could accrete gas from its host dwarf galaxy, which may cause the formation of second generation stars, producing multiple stellar populations[27]. Note that this scenario is also suggesting that the globular clusters with multiple stellar populations would be the remaining cores of the proto-galactic building blocks. This idea is supported by the recent investigations of the extended horizontal branch globular clusters (that is, globular clusters with signatures of multiple stellar populations), which have shown that extended horizontal branch globular clusters are clearly distinct from the normal globular clusters in orbital kinematics and mass[25]. Extensive photometric surveys for fainter stars in these globular clusters, as well as spectroscopic surveys for stars in double RGB sequences, would undoubtedly help to shed more light onto the discovery reported here.

Received 10 June; accepted 1 October 2009.

1. Lee, Y.-W. *et al.* Multiple stellar populations in the globular cluster ω Centauri as tracers of a merger event. *Nature* **402,** 55–57 (1999).
2. Bekki, K. & Freeman, K. C. Formation of ω Centauri from an ancient nucleated dwarf galaxy in the young Galactic disc. *Mon. Not. R. Astron. Soc.* **346,** L11–L15 (2003).
3. Freeman, K. C. in *The Globular Cluster-Galaxy Connection* (eds Smith, G. H & Brodie, J. P.) 608–614 (ASP Conf. Ser., Vol. 48, Astronomical Society of the Pacific, 1993).
4. Diemand, J., Kuhlen, M. & Madau, P. Formation and evolution of galaxy dark matter halos and their substructure. *Astrophys. J.* **667,** 859–877 (2007).
5. Johnson, C. I. *et al.* A large sample study of red giants in the globular cluster Omega Centauri (NGC 5139). *Astrophys. J.* **698,** 2048–2065 (2009).
6. Marino, A. F. *et al.* A double stellar generation in the globular cluster NGC 6656 (M22). Two stellar groups with different iron and s-process element abundance. *Astron. Astrophys.* **505,** 1099–1113 (2009).
7. Carretta, E. *et al.* Properties of second generation stars in globular clusters. Preprint at 〈http://arXiv.org/abs/0811.3591v1〉 (2008).
8. Georgiev, I. Y. *et al.* Globular cluster systems in nearby dwarf galaxies II. Nuclear star clusters and their relation to massive Galactic globular clusters. *Mon. Not. R. Astron. Soc.* **396,** 1075–1085 (2009).
9. Ventura, P. D. *et al.* Predictions for self-pollution in globular cluster stars. *Astrophys. J.* **550,** L65–L69 (2001).
10. Timmes, F. X., Woosley, S. E. & Weaver, T. A. Galactic chemical evolution: hydrogen through zinc. *Astrophys. J.* **98** (Suppl.), 617–658 (1995).
11. Baumgardt, H., Kroupa, P. & Parmentier, G. The influence of residual gas expulsion on the evolution of the Galactic globular cluster system and the origin of the Population II halo. *Mon. Not. R. Astron. Soc.* **384,** 1231–1241 (2008).
12. Anthony-Twarog, B. J. *et al.* Ca II H and K filter photometry on the uvby system. I-The standard system. *Astron. J.* **101,** 1902–1914 (1991).
13. Zinn, R. The globular cluster system of the Galaxy. I. The metal abundances and reddening of 79 globular clusters from integrated light measurements. *Astrophys. J.* **42** (Suppl.), 19–40 (1980).
14. Zinn, R. & West, M. J. The globular cluster system of the Galaxy. III. Measurements of radial velocity and metallicity for 60 clusters and a compilation of metallicities for 121 clusters. *Astrophys. J.* **55** (Suppl.), 45–66 (1984).
15. Stetson, P. B. DAOPHOT: A computer program for crowded-field stellar photometry. *Publ. Astron. Soc. Pacif.* **99,** 191–222 (1987).
16. Stetson, P. B. The center of the core-cusp globular cluster M15: CFHT and HST observations, ALLFRAME reductions. *Publ. Astron. Soc. Pacif.* **106,** 250–280 (1994).
17. Norris, J. & Freeman, K. C. The chemical inhomogeneity of M22. *Astrophys. J.* **266,** 130–143 (1983).
18. Richter, P., Hilker, M. & Richtler, T. Strömgren photometry in globular clusters: M55 & M22. *Astron. Astrophys.* **350,** 476–484 (1999).
19. Anthony-Twarog, B. J., Twarog, B. A. & Craig, J. CN and Ca abundance variations among the giants in M22. *Publ. Astron. Soc. Pacif.* **107,** 32–48 (1995).
20. Decressin, T., Charbonnel, C. & Meynet, G. Origin of the abundance patterns in Galactic globular clusters: constraints on dynamical and chemical properties of globular clusters. *Astron. Astrophys. J.* **475,** 859–873 (2007).
21. Yong, D. & Grundahl, F. An abundance analysis of bright giants in the globular cluster NGC1851. *Astrophys. J.* **672,** L29–L32 (2008).
22. Lee, J.-W. *et al.* Chemical inhomogeneity in red giant branch stars and RR Lyrae variables in NGC1851: two subpopulations in red giant branch. *Astrophys. J.* **695,** L78–L82 (2009).
23. Cudworth, K. M. Proper motions, membership, and photometry in the globular cluster M22. *Astron. J.* **92,** 348–357 (1986).
24. Brown, J. A. & Walllerstein, G. High-resolution CCD spectra of stars in globular clusters. VII. Abundances of 16 elements in 47 Tuc, M4, and M22. *Astron. J.* **104,** 1818–1830 (1992).
25. Lee, Y.-W., Gim, H. B. & Casetti-Dinescu, D. Kinematic decoupling of globular clusters with the extended horizontal branch. *Astrophys. J.* **661,** L49–L52 (2007).
26. Piotto, G. *et al.* Metallicities on the double main sequence of ω Centauri imply large helium enhancement. *Astrophys. J.* **621,** 777–784 (2005).
27. Pflamm-Altenburg, J. & Kroupa, P. Recurrent gas accretion by massive star clusters, multiple stellar populations and mass threshold for spherical stellar systems. *Mon. Not. R. Astron. Soc.* **397,** 488–494 (2009).

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

Acknowledgements J.-W.L. thanks A. Walker for providing the CTIO Ca filter transmission curve, D. Yong for NGC 1851 spectroscopic data before publication, and A. Yushchenko for discussions on spectrum synthesis. Support for this work was provided by the National Research Foundation of Korea to the Astrophysical Research Center for the Structure and Evolution of the Cosmos (ARCSEC). This work was based on observations made with the CTIO 1.0-m telescope, which is operated by the SMARTS consortium.

Author Contributions J.-W.L. performed observations, data analysis, interpretation, model simulations and writing of the manuscript; Y.-W.K. participated in observation planning; and J.L. performed part of the observations and data analysis. Y.-W. L. performed interpretation and writing of 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. Correspondence and requests for materials should be addressed to J.-W.L. (jaewoolee@sejong.ac.kr) or Y.-W.L. (ywlee2@yonsei.ac.kr).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 462|26 November 2009|doi:10.1038/nature08581

The cluster Terzan 5 as a remnant of a primordial building block of the Galactic bulge

F. R. Ferraro[1], E. Dalessandro[1], A. Mucciarelli[1], G. Beccari[2], R. M. Rich[3], L. Origlia[4], B. Lanzoni[1], R. T. Rood[5], E. Valenti[6,7], M. Bellazzini[4], S. M. Ransom[8] & G. Cocozza[4]

Globular star clusters are compact and massive stellar systems old enough to have witnessed the entire history of our Galaxy, the Milky Way. Although recent results[1–3] suggest that their formation may have been more complex than previously thought, they still are the best approximation to a stellar population formed over a relatively short timescale (less than 1 Gyr) and with virtually no dispersion in the iron content. Indeed, only one cluster-like system (ω Centauri) in the Galactic halo is known to have multiple stellar populations with a significant spread in iron abundance and age[4,5]. Similar findings in the Galactic bulge have been hampered by the obscuration arising from thick and varying layers of interstellar dust. Here we report that Terzan 5, a globular-cluster-like system in the Galactic bulge, has two stellar populations with different iron contents and ages. Terzan 5 could be the surviving remnant of one of the primordial building blocks that are thought to merge and form galaxy bulges.

We have recently obtained a set of high-resolution images of Terzan 5 in the K and J bands by using MAD[6], a Multi-Conjugate Adaptive Optics demonstrator instrument installed at the Very Large Telescope (VLT) of the European Southern Observatory (ESO). MAD operates at near-infrared wavelengths, thus revealing the only component of stellar radiation that can efficiently cross the thick clouds of dust obscuring the Galactic bulge. It is able to perform exceptionally good and uniform adaptive optics correction over its entire field of view ($1' \times 1'$), thus compensating for the degradation effects to the astronomical images induced by the Earth's atmosphere. In particular, we have obtained a set of K-band (2.2 μm) images of Terzan 5 close to the diffraction limit (Fig. 1). The sharpness and uniformity of the images yields very high quality photometry, resulting in an accurate (K, $J-K$) colour–magnitude diagram (CMD) even for the very central region of the cluster, and leading to a surprising discovery. We have detected two well-defined red horizontal branch clumps, separated in luminosity: a bright horizontal branch (BHB) at $K = 12.85$ mag and a faint horizontal branch (FHB) at $K = 13.15$ mag, the latter having a bluer ($J-K$) colour (Fig. 2).

We have carefully considered whether the double horizontal branch could be spurious. It is neither due to instrumental effects (Fig. 2), nor to differential reddening[7,8] (as the two horizontal branch clumps in the CMD are separated in a direction which is essentially orthogonal to the reddening vector), nor to field contamination (while field stars are expected to be almost uniformly distributed over the MAD field of view, the radial distributions of the stars belonging to the two horizontal branch clumps are significantly concentrated towards the cluster centre, and are inconsistent with a uniform distribution at more than the 5σ level; see Fig. 3a and Supplementary Information). We have also found that the radial distributions of the two horizontal branch populations are different (Fig. 3a): according to a Kolmogorov–Smirnov test, the BHB population is significantly (at $>3.5\sigma$ level) more centrally concentrated than that of the FHB. The stars belonging to the BHB are also substantially more numerous than those of the FHB near the cluster centre (that is, at distances $r < 20''$), becoming progressively more rare at larger radii (Fig. 3b).

Once alerted to the existence of the double horizontal branch, we have also identified the feature in optical observations obtained with the Advanced Camera for Surveys (ACS) on board the Hubble Space Telescope (HST; see Supplementary Fig. 1a). Although the strong

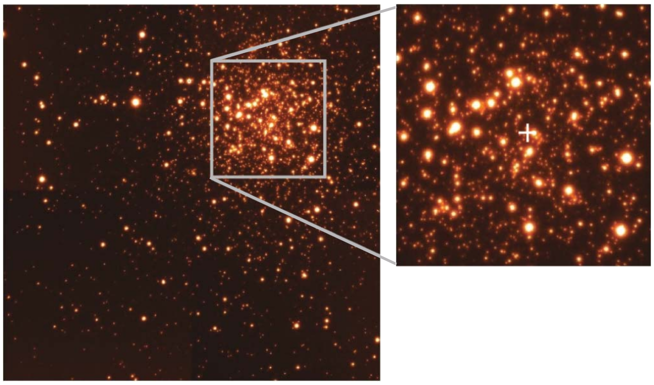


Figure 1 | MAD image of Terzan 5 in the K band. Observations were performed at the ESO-VLT (Paranal, Chile) on August 2008, through J and K filters. Exposure times were about two minutes in each filter. Shown is the best image obtained in the K band (the image size is $1' \times 1'$, north is up, east is left). The measured full-width at half-maximum (FWHM) of stars is $0.1''$, the Strehl ratio ranges between 15% and 24% over the entire field of view. The quality of the J image is slightly worse (FWHM $\approx 0.24''$ and Strehl ratio $<10\%$), but still much better than normally obtained with ground-based observations. A small ($16'' \times 16''$) portion of the K image sampling the very central region of Terzan 5 is shown magnified. The cluster centre of gravity (marked with the white cross) has been determined by averaging the positions of the resolved stars and following the same procedure adopted in previous studies[25]. It is located at right ascension $\alpha = 17$ h 48 min 4.85 s, declination $\delta = -24° \, 46' \, 44.6''$, which is $\sim 3''$ southeast from the centre listed in the most commonly adopted globular cluster catalogue[14], but in good agreement (within the errors $\Delta\alpha \approx \Delta\delta \approx 0.5''$) with the determination obtained from HST-NICMOS observations[9]. The barycentres of the two horizontal branch populations are coincident with the gravity centre within the errors.

[1]Department of Astronomy, University of Bologna, Via Ranzani, 1, 40127 Bologna, Italy. [2]ESA, Space Science Department, Keplerlaan 1, 2200 AG Noordwijk, The Netherlands. [3]Department of Physics and Astronomy, Math-Sciences 8979, UCLA, Los Angeles, California 90095-1562, USA. [4]INAF-Osservatorio Astronomico di Bologna, Via Ranzani, 1, 40127 Bologna, Italy. [5]Astronomy Department, University of Virginia, PO Box 400325, Charlottesville, Virginia 22904, USA. [6]European Southern Observatory, Alonso de Cordova 3107, Vitacura, Casilla 19001, Santiago, Chile. [7]Pontificia Universidad Catolica de Chile, Departamento de Astronomia, Avda Vicuña Mackenna 4860, 782-0436 Macul, Santiago, Chile. [8]National Radio Astronomy Observatory, Charlottesville, Virginia 22903, USA.

©2009 Macmillan Publishers Limited. All rights reserved

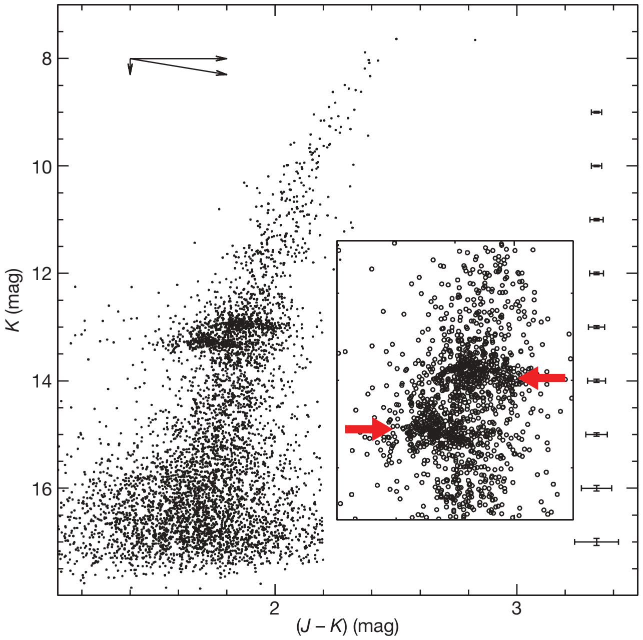


Figure 2 | The two horizontal branch clumps of Terzan 5. Main panel, MAD (*K*, *J* − *K*) CMD of the central region of Terzan 5. Inset, magnified view of the horizontal branch region, with the two horizontal branch clumps marked with red arrows. Terzan 5 is heavily obscured by thick clouds of dust (this effect is commonly called 'reddening') intervening between the system and the observer, in a way that strongly depends on the direction of the line of sight ('differential reddening')[7,8]. The effect of reddening on the *K* magnitude and the *J* − *K* colour is indicated by the reddening vector plotted in the main panel. Several tests have been performed on the images and the catalogue to exclude any possible spurious effect from the instrument or the reduction procedure. Stars in the two clumps do not show any peculiar spatial distribution on the detector. Moreover, the two clumps are not spuriously produced by the variation in size and shape of the point spread function, or the local level of the background. Error bars (1 s.e.m) are plotted at different magnitude levels. The contamination from Galactic bulge field stars in this CMD is negligible. In the 1 arcmin^2 field of view of MAD, we estimate (Supplementary Information) that 11 and 8 field stars should contaminate the faint and bright horizontal branch selection boxes (while we count 299 FHB stars and 310 BHB stars in the entire MAD sample).

differential reddening broadens the colour extension of the horizontal branch clumps by ~1 mag, the optical (*I*, *V* − *I*) CMD still shows a clear bimodal distribution of horizontal branch stars in the direction orthogonal to the reddening vector (Supplementary Fig. 1b). A hint of a double horizontal branch clump was already visible in a previously published CMD obtained with HST-NICMOS[9–11], although the shorter colour baseline provided by the J- and H-band observations did not clearly separate the two clumps.

Hence, we conclude that the existence of the two horizontal branch clumps is a real feature, and the differing radial distributions may indicate different physical origins of the two populations. In particular, a combination of different metallicity and age, with the population in the BHB clump being more metal-rich and younger than that in the FHB clump, could in principle reproduce the observed features (Supplementary Fig. 2). The only direct information previously available on the metal content of individual stars in Terzan 5 was from four bright giants near the tip of the red giant branch (RGB), giving an average iron-to-hydrogen abundance ratio [Fe/H] = −0.2 with a negligible dispersion[12]. Hence, we quickly secured medium-resolution near-infrared spectra of 6 horizontal branch stars (3 in each clump) at the Keck Telescope[13]. The derived radial velocities for the two groups of stars (−85 km s^{-1} in both cases) are fully consistent with the previous measures[12] and the systemic velocity of Terzan 5 quoted in the currently adopted globular cluster catalogue[14]. This confirms that all of the observed stars are cluster members and suggests that there is no significant kinematical difference between the two populations (this is also confirmed by proper motion studies; see Supplementary Information). Furthermore, we have found that the iron content of the stars in the two clumps differs by a factor of 3 (~0.5 dex): the FHB stars have [Fe/H] = −0.2, while the BHB stars have [Fe/H] = +0.3 (Fig. 4a).

To date, apart from a significant spread in the abundance patterns of a few light elements (such as Na and O)[1], the chemical composition of all globular clusters in the Galaxy is known to be extremely uniform in terms of iron content, with the only exception being ω Centauri[4,5] in the Galactic halo. Hence, Terzan 5 is the first stellar aggregate discovered in the Galactic bulge that has globular-cluster-like properties but also the signatures of a much more complex star formation history.

To further investigate this issue, we have performed a differential reddening correction[15] on the optical ACS catalogue and combined it with the near-infrared data, thus obtaining the (*K*, *V* − *K*) CMD shown in Fig. 4b. The presence of two distinct populations with a double horizontal branch and (possibly) two separate RGBs can be seen in this CMD. The RGB of the most metal-rich population appears to be more bent (as expected, because of the line blanketing due to a higher metal content). The observed features can be reproduced with

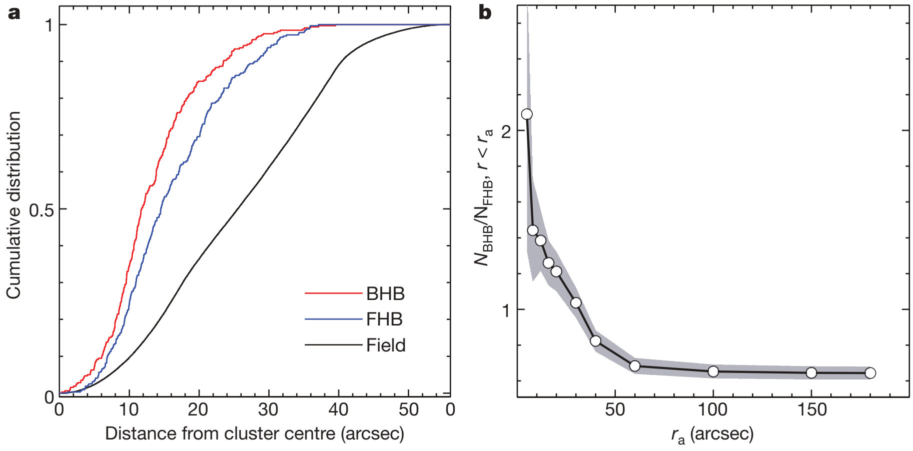


Figure 3 | Radial distribution of the two horizontal branch populations in Terzan 5. **a**, Cumulative radial distribution of the observed BHB stars (red line) and the FHB population (blue line), compared to that of field stars (solid black line), as a function of the projected distance from the cluster centre of gravity. The field distribution has been obtained from a synthetic sample of 100,000 points uniformly distributed in *X* and *Y* over the MAD field of view. **b**, Ratio between the number of observed BHB and FHB stars computed over areas of increasing radius, r_a. Points with $r_a < 30''$ refer to the MAD sample, those corresponding to larger radii have been computed by also using the ACS data. The grey area around the black line represents the 1σ uncertainty region. BHB stars are substantially more numerous than FHB stars in the cluster centre and they rapidly vanish at $r_a > 50''$.

©2009 Macmillan Publishers Limited. All rights reserved

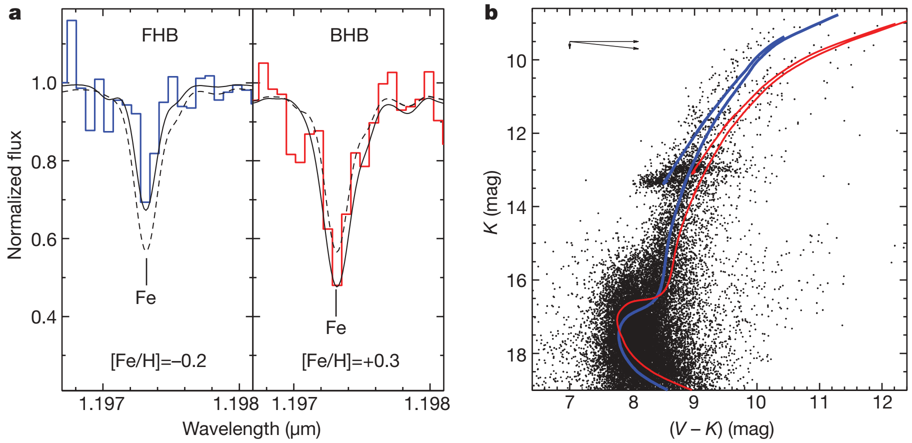


Figure 4 | Iron abundance and ages of the two populations. **a**, Combined J-band spectra near the 1.1973 μm iron line for three FHB (left) and three BHB (right) stars, as obtained with NIRSPEC at Keck II on 2 July 2009 (coloured lines). The measured equivalent widths of the lines and suitable spectral synthesis[12] yield iron abundances $[\mathrm{Fe/H}] \approx -0.2 \pm 0.1$ and $[\mathrm{Fe/H}] \approx +0.3 \pm 0.1$, respectively. The black solid lines correspond to the best-fit synthetic spectra obtained for temperatures and gravities derived from evolutionary models reproducing the observed colours of the horizontal branch stars: $T_{\mathrm{eff}} = 5{,}000$ K and $\log g = 2.5$ for the FHB stars, $T_{\mathrm{eff}} = 4{,}500$ K and $\log g = 2.0$ for the BHB stars. For sake of comparison, we also plot (as black dashed lines) the synthetic spectra obtained by adopting the same atmospheric parameters, but $[\mathrm{Fe/H}] = +0.3$ for the FHB and $[\mathrm{Fe/H}] = -0.2$ for the BHB. From the measured spectra, we also derived the stellar radial velocities and found an average value of $-85\ \mathrm{km\,s^{-1}}$ ($\sigma = 9\ \mathrm{km\,s^{-1}}$) and $-85\ \mathrm{km\,s^{-1}}$ ($\sigma = 10\ \mathrm{km\,s^{-1}}$) for the FHB and BHB stars, respectively (the typical uncertainty on the individual measure is of the order of $3\ \mathrm{km\,s^{-1}}$). These values are fully consistent with the previously measured radial velocities of four giants ($V_{\mathrm{r}} = -93 \pm 2\ \mathrm{km\,s^{-1}}$)[12] and the value ($V_{\mathrm{r}} = -94 \pm 15\ \mathrm{km\,s^{-1}}$) listed for Terzan 5 in the currently adopted globular cluster catalogue[14]. This observational fact confirms that the horizontal branch stars for which we have secured spectra are cluster members, and suggests that there is no significant kinematical difference between the two populations. **b**, (K, $V - K$) CMD of Terzan 5 obtained by combining VLT-MAD and HST-ACS data corrected for differential reddening. Two isochrones[26] with $[\mathrm{Fe/H}] = -0.2$ (heavy element mass fraction $Z = 0.01$, and helium mass fraction $Y = 0.26$) and $t = 12$ Gyr (blue line), and with $[\mathrm{Fe/H}] = +0.3$ ($Z = 0.03$, $Y = 0.29$) and $t = 6$ Gyr (red line) are overplotted on the data by adopting a colour excess[8] $E(B - V) = 2.38 \pm 0.05$ and a distance[8] $d = 5.9 \pm 0.5$ kpc. Note that the CMD cannot be reproduced by two isochrones with the measured metallicities and the same age. Owing to the large scatter at the turn-off level, we estimate that the uncertainty on the younger component age is about 2 Gyr.

two populations characterized by the observed metallicities and two different ages: $t = 12$ Gyr for the FHB and a significantly younger age ($t = 6$ Gyr) for the BHB.

Using the number of horizontal branch stars found in the combined MAD and ACS samples (see Supplementary Information for details), we estimate that the cluster harbours about 800 FHB stars and 500 BHB stars in total. This is even larger than the global horizontal branch population of 47 Tucanae[16], thus suggesting that Terzan 5 is more massive than previously thought (Supplementary Information).

The evidence for two distinct stellar populations and for a very large total mass suggests that Terzan 5 has experienced a quite troubled formation history. It might be the merger-product of two independent stellar aggregates[17]. Although such a possibility seems to be unlikely for globular clusters belonging to the Galactic halo, the chance of capturing a completely independent stellar system should be significantly larger if the orbits are confined within the Galactic bulge. In this scenario, however, it is not easy to explain why the metal-rich population is more centrally concentrated than the metal-poor one. Moreover, globular clusters younger than 10 Gyr are very rare in our Galaxy[18]. Rather, Terzan 5 could be a complex ω Centauri-like system[4,5] or the nuclear remnant of a disrupted galaxy, similar to the M 54–Sagittarius system[19,20] or the Carina dwarf spheroidal[21] in the metal-rich regime. The remnant of a disrupted dwarf galaxy would naturally present a larger central concentration of the metal-rich (and younger) population[22], as commonly observed in the satellites of the Milky Way and M 31. On the other hand, the strict similarity in iron abundance between Terzan 5 and the Galactic bulge population is fully compatible with the hypothesis that the (partial) disruption of its progenitor has contributed to the formation of the Galactic bulge[23].

Possible relics of the hierarchical assembly of the Galactic halo have been recently identified at high Galactic latitudes[24]. Terzan 5 may be the first example of the sub-structures that contributed to form the Galactic bulge. Indeed, our discovery could be the observational confirmation that galactic spheroids originate from the merging of pre-formed, internally evolved stellar systems, and that other similar objects might be hidden in the heavily obscured central region of the Galaxy.

Received 20 August; accepted 8 October 2009.

1. Gratton, R., Sneden, C. & Carretta, E. Abundance variations within globular clusters. *Annu. Rev. Astron. Astrophys.* **42,** 385–440 (2004).
2. Piotto, G. in *The Ages of Stars* (ed. Montmerle, T.) 233–244 (IAU Symp. 258, Cambridge Univ. Press, 2009).
3. Lee, J.-W., Kang, Y.-W., Lee, J. & Lee, Y.-W. Enrichment by supernovae in globular clusters with multiple populations. *Nature* doi:10.1038/nature08565 (this issue).
4. Norris, J. E. & Da Costa, G. S. The giant branch of Omega Centauri. IV. Abundance patterns based on echelle spectra of 40 red giants. *Astrophys. J.* **447,** 680–705 (1995).
5. Sollima, A. *et al.* Metallicities, relative ages, and kinematics of stellar populations in ω Centauri. *Astrophys. J.* **634,** 332–343 (2005).
6. Marchetti, E. *et al.* On-sky testing of the multi-conjugate adaptive optics demonstrator. *The Messenger* 129, 8–13 (2007); available at ⟨http://www.eso.org/sci/publications/messenger/archive/no.129-sep07/messenger-no129-8.pdf⟩.
7. Ortolani, S., Barbuy, B. & Bica, E. NTT VI photometry of the metal-rich and obscured bulge globular cluster Terzan 5. *Astron. Astrophys.* **308,** 733–737 (1996).
8. Valenti, E., Ferraro, F. R. & Origlia, L. Near-infrared properties of 24 globular clusters in the Galactic Bulge. *Astron. J.* **133,** 1287–1301 (2007).
9. Cohn, H. N., Lugger, P. M., Grindlay, J. E. & Edmonds, P. D. Hubble Space Telescope/NICMOS observations of Terzan 5: stellar content and structure of the core. *Astrophys. J.* **571,** 818–829 (2002).
10. Ortolani, S. *et al.* HST NICMOS photometry of the reddened bulge globular clusters NGC 6528, Terzan 5, Liller 1, UKS 1 and Terzan 4. *Astron. Astrophys.* **376,** 878–884 (2001).
11. Ortolani, S., Barbuy, B., Bica, E., Zoccali, M. & Renzini, A. Distances of the bulge globular clusters Terzan 5, Liller 1, UKS 1, and Terzan 4 based on HST NICMOS. *Astron. Astrophys.* **470,** 1043–1049 (2007).
12. Origlia, L. & Rich, R. M. High-resolution infrared spectra of bulge globular clusters: the extreme chemical abundances of Terzan 4 and Terzan 5. *Astron. J.* **127,** 3422–3430 (2004).
13. McLean, I. S. *et al.* Design and development of NIRSPEC: a near-infrared echelle spectrograph for the Keck II telescope. *Proc. SPIE* **3354,** 566–578 (1998).
14. Harris, W. E. A catalog of parameters for globular clusters in the Milky Way. *Astron. J.* **112,** 1487–1488 (1996).

©2009 Macmillan Publishers Limited. All rights reserved

15. Piotto, G. *et al.* Hubble Space Telescope observations of Galactic globular cluster cores. II. NGC 6273 and the problem of horizontal-branch gaps. *Astron. J.* **118,** 1727–1737 (1999).
16. Beccari, G., Ferraro, F. R., Lanzoni, B. & Bellazzini, M. A population of binaries in the asymptotic giant branch of 47 Tucanae? *Astrophys. J.* **652,** L121–L124 (2006).
17. Icke, V. & Alcaino, G. Is Omega Centauri a merger? *Astron. Astrophys.* **204,** 115–116 (1988).
18. Marín-Franch, A. *et al.* The ACS survey of galactic globular clusters. VII. Relative ages. *Astrophys. J.* **694,** 1498–1516 (2009).
19. Ibata, R. A., Gilmore, G. & Irwin, M. J. A dwarf satellite galaxy in Sagittarius. *Nature* **370,** 194–196 (1994).
20. Bellazzini, M. *et al.* The nucleus of the Sagittarius Dsph galaxy and M54: a window on the process of galaxy nucleation. *Astron. J.* **136,** 1147–1170 (2008).
21. Hurley-Keller, D. & Mateo, M. in *Astrophysical Ages and Times Scales* (eds von Hippel, T., Simpson, C. & Manset, N.) 322–324 (ASP Conf. Ser. Vol. 245, Astron. Soc. Pacif., 2001).
22. Harbeck, D. *et al.* Population gradients in local group dwarf spheroidal galaxies. *Astron. J.* **122,** 3092–3105 (2001).
23. Immeli, A., Samland, M., Gerhard, O. & Westera, P. Gas physics, disk fragmentation, and bulge formation in young galaxies. *Astron. Astrophys.* **413,** 547–561 (2004).
24. Belokurov, V. *et al.* Cats and dogs, hair and a hero: a quintet of new Milky Way companions. *Astrophys. J.* **654,** 897–906 (2007).
25. Lanzoni, B. *et al.* The surface density profile of NGC 6388: a good candidate for harboring an intermediate-mass black hole. *Astrophys. J.* **668,** L139–L142 (2007).
26. Pietrinferni, A., Cassisi, S., Salaris, M. & Castelli, F. A large stellar evolution database for population synthesis studies. I. Scaled solar models and isochrones. *Astrophys. J.* **612,** 168–190 (2004).

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

Acknowledgements We thank the MAD team at ESO and in particular P. D'Amico for performing the observations with MAD. This research was supported by ''Progetti Strategici di Ateneo 2006'' (University of Bologna), ''Progetti di Ricerca di Interesse Nazionale 2007 and 2008'' (Istituto Nazionale di Astrofisica), Agenzia Spaziale Italiana, and the Ministero dell'Istruzione, dell'Universitá e della Ricerca. We also acknowledge support from the ESTEC Faculty Visiting Scientist Programme. R.M.R. is supported by the NSF and STScI; and R.T.R. is partially supported by STScI. This research has made use of the ESO/ST-ECF Science Archive facility, which is a joint collaboration of the ESO and the Space Telescope – European Coordinating Facility. Part of the data presented here 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 NASA. The Observatory was made possible by the financial support of the W. M. Keck Foundation.

Author Contributions F.R.F. designed the study and coordinated the activity. E.D., A.M., G.B., E.V. and G.C. analysed the photometric dataset. R.M.R. and L.O. secured and analysed the Keck spectra. A.M. designed the reddening correction routine. M.B. performed radial distributions tests. E.D. and M.B performed the proper motion analysis. F.R.F., B.L. and L.O. wrote the paper. S.M.R., R.M.R., R.T.R. and M.B. critically contributed to the paper presentation. All authors contributed to discussion of the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.R.F. (francesco.ferraro3@unibo.it).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 462|26 November 2009|doi:10.1038/nature08566

Two-dimensional normal-state quantum oscillations in a superconducting heterostructure

Y. Kozuka[1]*, M. Kim[1]*, C. Bell[1,2], B. G. Kim[1,3], Y. Hikita[1] & H. Y. Hwang[1,2]

Semiconductor heterostructures provide an ideal platform for studying high-mobility, low-density electrons in reduced dimensions[1–4]. The realization of superconductivity in heavily doped diamond[5], silicon[6], silicon carbide[7] and germanium[8] suggests that Cooper pairs eventually may be directly incorporated in semiconductor heterostructures[9], but these newly discovered superconductors are currently limited by their extremely large electronic disorder. Similarly, the electron mean free path in low-dimensional superconducting thin films is usually limited by interface scattering, in single-crystal or polycrystalline samples, or atomic-scale disorder, in amorphous materials, confining these examples to the extreme 'dirty limit'[10]. Here we report the fabrication of a high-quality superconducting layer within a thin-film heterostructure based on $SrTiO_3$ (the first known superconducting semiconductor[11]). By selectively doping a narrow region of $SrTiO_3$ with the electron-donor niobium, we form a superconductor that is two-dimensional, as probed by the anisotropy of the upper critical magnetic field. Unlike in previous examples, however, the electron mobility is high enough that the normal-state resistance exhibits Shubnikov–de Haas oscillations that scale with the perpendicular field, indicating two-dimensional states. These results suggest that delta-doped $SrTiO_3$ provides a model system in which to explore the quantum transport and interplay[12] of both superconducting and normal electrons. They also demonstrate that high-quality complex oxide heterostructures can maintain electron coherence on the macroscopic scales probed by transport, as well as on the microscopic scales demonstrated previously[13].

The technique of delta-doping has been discussed in detail in the semiconductor literature[14]. A key point of this method is the reduction of the dopant-layer thickness to below the other characteristic lengths in the system, such as the electronic mean free path, l_{mfp}. In this limit, the electron-wavefunction weight in the nearby undoped regions can be significant, leading to a reduction in electron scattering by ionized impurities and an enhanced mobility relative to a uniformly doped thin film[15]. In the case of $SrTiO_3$ (STO), its extremely large dielectric constant at low temperatures[16–18] also allows effective impurity screening, leading to further improvements in the transport properties. STO can also be electron-doped to become a superconductor with the lowest known carrier density of any material[19]. For these reasons, this material is increasingly being used to investigate novel phenomena in field-effect devices[20–22]. However, all of these devices rely on electron confinement close to some form of interface, where such lattice screening is not fully effective. Furthermore, high-mobility doped STO thin films have until now been unavailable for the study of transport in more complex heterostructures. Recently, we developed a reliable technique to make high-quality niobium-doped $SrTiO_3$ (NSTO) thin films with transport properties similar to the highest-quality single crystals.

In the work reported here, our sample consisted of a thin-film heterostructure with 100-nm-thick undoped STO layers above and below a NSTO layer that was nominally 5.5 nm thick, as sketched in Fig. 1a. The motivation behind this design was the desire to remove the effects of surface depletion[23], as well as any electron scattering arising at the surface and, possibly, at the film–substrate interface. Thus, we have a narrow conducting channel with no physical interface controlling the wavefunction of the electrons, as it is only the fixed-charge potential of the dopants that confines them. As shown in Fig. 1b, the low-field Hall mobility at $T = 2$ K was 1,100 $\mathrm{cm^2\,V^{-1}\,s^{-1}}$ and the carrier density was $N_s = 4.7 \times 10^{13}\,\mathrm{cm^{-2}}$. If we simply assume that charge was uniformly distributed throughout the 5.5-nm NSTO layer, this gives a three-dimensional (3D) carrier density of $N_{3D} = 8.5 \times 10^{19}\,\mathrm{cm^{-3}}$. This value is within the range for which NSTO is superconducting[11], which we confirmed by dilution-refrigerator measurements (Fig. 1c) showing a clear superconducting transition with midpoint at $T_c = 370$ mK and a 10–90% width of 18 mK.

To investigate the nature of the superconductivity in this structure further, we measured the superconducting upper critical field, H_{c2}, at various angles, θ, between the sample plane and the magnetic field (Fig. 2a), finding a strong anisotropy between the parallel critical field ($\theta = 0°$), $\mu_0 H_{c2}^{\parallel}(T) = 1.87$ T, and the perpendicular critical field ($\theta = 90°$), $\mu_0 H_{c2}^{\perp}(T) = 0.061$ T, where μ_0 is the vacuum permeability.

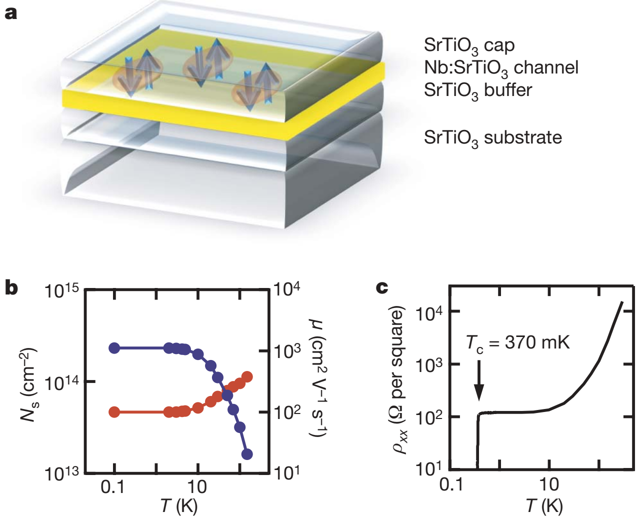


Figure 1 | Sample structure and transport characterization. a, A sketch of the delta-doped NSTO layer sandwiched between insulating STO buffer and cap layers on an STO substrate. Cooper pairs forming the superconducting layer in the delta-doped layer are shown schematically. Layer thicknesses are not to scale. **b,** Low-field sheet carrier density, N_s (red), and electron Hall mobility, μ (blue), versus temperature. **c,** Sheet resistance, ρ_{xx}, versus temperature, showing a clear superconducting transition at 370 mK.

[1]Department of Advanced Materials Science, University of Tokyo, Kashiwa, Chiba 277-8561, Japan. [2]Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan. [3]Department of Physics, Pusan National University, Busan 609-735, South Korea.

*These authors contributed equally to this work

©2009 Macmillan Publishers Limited. All rights reserved

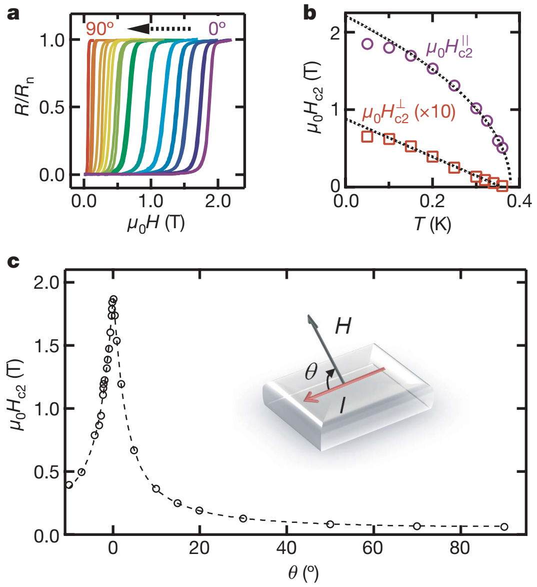


Figure 2 | Two-dimensional superconducting characteristics. **a**, Measurement of superconducting upper critical field, H_{c2}, using resistance versus magnetic field, H, for various angles between the sample plane and the magnetic field at $T = 50$ mK. R_n, normal-state resistance. **b**, H_{c2} versus temperature for H applied perpendicular to the sample plane (squares; field data multiplied by ten for clarity) and parallel to the plane (circles). Dashed lines are fits to linearized Ginzburg–Landau theory. **c**, Full angular dependence of H_{c2} ($H_{c2}(\theta)$), where the dashed line is a fit to equation (1). The sketch defines the angle between the sample plane and the applied magnetic field. I, current.

This is a consequence of the thickness of the superconducting layer being much thinner than the Ginzburg–Landau coherence length, ξ_{GL}, meaning that superconductivity in this type-II material cannot be suppressed by vortex entry for in-plane fields. From measurements of the temperature dependence of $H_{c2}^{\perp}$, we can directly extract ξ_{GL} using the linearized Ginzburg–Landau form

$$H_{c2}^{\perp}(T)=\frac{\Phi_0}{2\pi\xi_{GL}(0)^2}\left(1-\frac{T}{T_c}\right)$$

where Φ_0 is the flux quantum and $\xi_{GL}(0)$ is the extrapolation of ξ_{GL} to $T = 0$ K. This fit (Fig. 2b) gives $\xi_{GL}(0) = 61.2 \pm 1.4$ nm. For a two-dimensional (2D) superconductor

$$H_{c2}^{||}(T)=\frac{\Phi_0\sqrt{12}}{2\pi\xi_{GL}(0)d}\left(1-\frac{T}{T_c}\right)^{1/2}$$

where d is the superconducting thickness. This characteristic square-root dependence corresponds accurately with our data. From this fit, we extract a superconducting thickness of $d = 8.4 \pm 0.1$ nm. This value is in reasonable correspondence with the growth thickness if we consider that the wavefunction spreading into the undoped STO will tend to increase the effective thickness of the delta layer[14] (Supplementary Information). Furthermore, a careful fitting of the full $H_{c2}(\theta)$ data using the formula first derived in ref. 24, that is,

$$\left|\frac{H_{c2}(\theta)\sin\theta}{H_{c2}^{\perp}}\right|+\left(\frac{H_{c2}(\theta)\cos\theta}{H_{c2}^{||}}\right)^2=1 \qquad (1)$$

shows good agreement, as illustrated in Fig. 2c.

Next we turn to the normal-state properties above the critical field at which superconductivity is suppressed. For $\mu_0 H = 5$–14 T, we observed Shubnikov–de Haas quantum oscillations in the transverse geometry, superimposed on a background of positive magnetoresistance (Fig. 3a) and periodic in $1/H$. After the magnetoresistance background has been subtracted, measurements for various values of θ show a clear scaling of the peak positions with the reciprocal perpendicular component of the magnetic field, that is, with $1/H_{\perp} = 1/H\sin\theta$, as is evident from a comparison of Fig. 3b and Fig. 3c. This feature is critical, and demonstrates that the Shubnikov–de Haas oscillations are due to orbits around a cylindrically shaped Fermi surface, indicative of quantum transport in 2D systems. Thus, for this sample, we find 2D superconductivity at low magnetic fields and Shubnikov–de Haas oscillations due to a 2D Fermi surface topology at high magnetic fields. This is an unexpected result, for here a 2D Fermi surface has been engineered from a purely 3D superconducting oxide host. As a result, this structure is free from the finite perpendicular coupling and dispersion that is always present in naturally layered, quasi-2D bulk superconductors. Although many 2D superconductors have been formed from 3D materials using thin films, they have generally been characterized by very low values of l_{mfp}. For example, in amorphous bismuth, the electrons are scattered on an atomic scale, resulting in single-electron transport that is completely incoherent and diffusive[10]. Nevertheless, these systems are superconducting, albeit in the dirty limit, reflecting the robustness of superconductivity to disorder. By contrast, the Shubnikov–de Haas oscillations we demonstrate in our artificial 2D superconductor are made possible by the high crystalline coherence and lack of interface or surface scattering.

From a Fourier transform of the Shubnikov–de Haas oscillations in terms of $1/\mu_0 H$ in the range $5\,\mathrm{T} < \mu_0 H < 14\,\mathrm{T}$ (Fig. 4a), it is clear that more than one frequency is observed in the oscillations. Nominal peak assignments by Lorentzian fitting give three primary components (Fig. 4a). These correspond to respective sheet carrier densities, N_{SdH}, of $1.8 \times 10^{12}\,\mathrm{cm}^{-2}$, $3.6 \times 10^{12}\,\mathrm{cm}^{-2}$ and $4.8 \times 10^{12}\,\mathrm{cm}^{-2}$ (representing 0.14, 0.27 and 0.37% of the Brillouin zone, respectively), calculated using the free carrier approximation assuming spin degeneracy. Thus, N_s, the sheet carrier density derived from the Hall data, and N_{SdH} show significant disagreement. This suggests that in addition to the intrinsic conduction band structure of STO, which is composed of three non-degenerate pockets at the zone centre[25], there

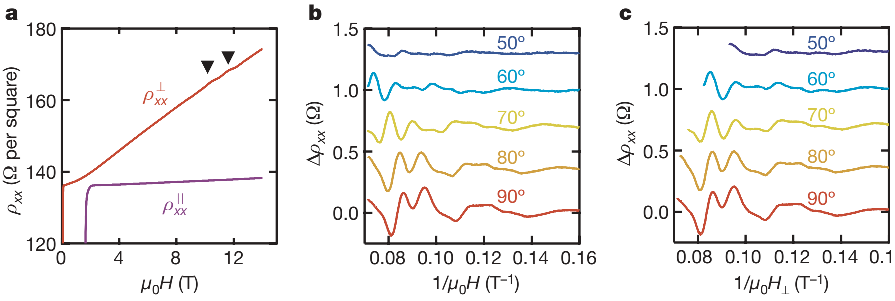


Figure 3 | Two-dimensional quantum oscillations in the normal state. **a**, Longitudinal resistivity in the perpendicular ($\rho_{xx}^{\perp}$) and parallel ($\rho_{xx}^{||}$) geometry, from $\mu_0 H = 0$–14 T at $T = 100$ mK. Shubnikov–de Haas oscillations are visible (arrowheads) in $\rho_{xx}^{\perp}(H)$. The sudden increase in ρ_{xx} at low fields is due to the superconducting upper critical field being crossed. **b**, **c**, Amplitude of the Shubnikov–de Haas oscillations, $\Delta\rho_{xx}$, after background subtraction, for various angles at $T = 100$ mK, versus the reciprocal total magnetic field (**b**) and the reciprocal perpendicular magnetic field component ($1/\mu_0 H_{\perp}$; **c**).

©2009 Macmillan Publishers Limited. All rights reserved

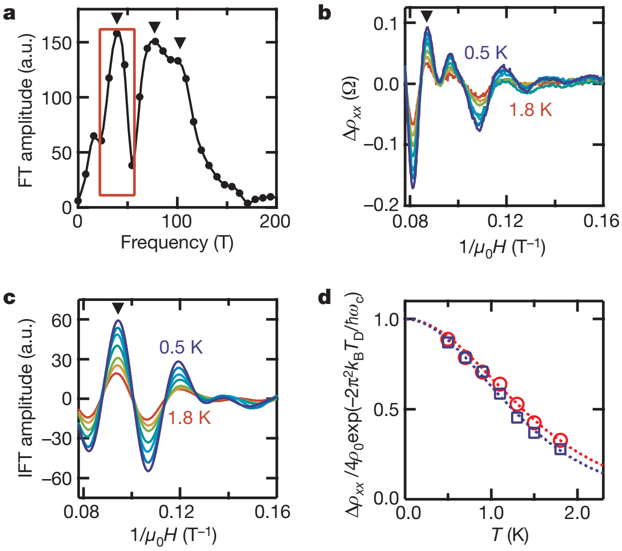


Figure 4 | Carrier effective mass. a, Fourier transform (FT) of the Shubnikov–de Haas oscillations at $T = 500$ mK. Lorentzian fitting gives primary peaks corresponding to the frequencies 37.5, 74.1 and 99.4 T. a.u., arbitrary units. **b,** Temperature dependence of oscillation amplitude for $\theta = 90^\circ$. **c,** Inverse Fourier transform (IFT) amplitude for the boxed area in **a** around 37.5 T. **d,** Temperature dependence of the scaled Shubnikov–de Haas (circles) and IFT (squares) amplitudes for peaks indicated by arrowheads in **b** and **c.** Fits to equation (2) give effective masses of $(1.24 \pm 0.02)m_0$ and $(1.26 \pm 0.03)m_0$ for $\Delta\rho_{xx}$ and IFT data, respectively.

is a complex electronic structure originating from multi-subband occupancy. We must therefore be cautious in drawing strong conclusions from analysis based on a single band picture: detailed band structure calculations are necessary to shed light on the true energy-level diagram of our delta-doped system, taking into account the nonlinear response of the STO permittivity[16]. An approximate model indicates that three light-electron subbands are occupied, as well as four heavy and two spin–orbit-split subbands (Supplementary Information). Accordingly, we may conclude that the discrepancy between N_s and N_{SdH} arises because only selected subbands in some of the electron pockets contribute to the Shubnikov–de Haas oscillations, as a result of different effective masses and/or scattering times. Nevertheless, the Shubnikov–de Haas oscillations in Fig. 3c clearly show that a high-mobility 2D electron gas can be created in a superconducting oxide heterostructure. For the lowest Shubnikov–de Haas frequency (37.5 T; Fig. 4a), the Landau-level index is 2.2 at the highest measurement field, 14 T, approaching the quantum limit.

To estimate the effective mass, m^*, and the Dingle temperature, T_D, of the 2D electron gas, we measured the temperature dependence of the Shubnikov–de Haas oscillations (Fig. 4b) and analysed it in two different ways. First we chose the oscillation peak at $\mu_0H = 11.4$ T ($1/\mu_0H = 0.087$ T^{-1}) and fitted the Shubnikov–de Haas amplitude using the form[2]

$$\Delta\rho_{xx} = 4\rho_0 \exp\left(-2\pi^2 k_B T_D/\hbar\omega_c\right)\frac{2\pi^2 k_B T/\hbar\omega_c}{\sinh\left(2\pi^2 k_B T/\hbar\omega_c\right)} \qquad (2)$$

where ρ_0 is the non-oscillatory component of the resistivity, ω_c is the cyclotron frequency ($e\mu_0H/m^*$, where e is the elementary charge), k_B is Boltzmann's constant and $\hbar$ is Planck's constant divided by 2π. The fit shown in Fig. 4d (red circles) leads to an effective mass of $m^* = (1.24 \pm 0.02)m_0$ (m_0 being the bare electron mass) and a Dingle temperature of $T_D = 5.58 \pm 0.07$ K, corresponding to a quantum scattering time of $\tau_q = \hbar/2\pi k_B T_D = (2.18 \pm 0.03) \times 10^{-13}$ s. However, the existence of multiple frequency components in the Shubnikov–de Haas oscillations indicates that this simple analysis neglects multiple contributions to a given peak. Accordingly, we carried out the same analysis for the temperature dependence of the Shubnikov–de Haas oscillations in Fig. 4c, which were constructed from the inverse Fourier transform of the first Fourier peak, centred at 37.5 T (Fig. 4a). In this way (Fig. 4d, blue squares), we find that $m^* = (1.26 \pm 0.03)m_0$, $T_D = 4.48 \pm 0.2$ K and $\tau_q = (2.71 \pm 0.12) \times 10^{-13}$ s. The two approaches thus provide similar values and suggest that all Shubnikov–de Haas components have similar effective masses.

The values we calculate for m^* are comparable to the light effective mass for STO found in band structure calculations[25] and bulk experiments[26], lending some support to the subband assignments discussed above, although for a quantitative comparison a more sophisticated calculation would required. This intact effective mass, even under 2D confinement, may be a consequence of the unique design of our heterostructure, which is wholly composed of STO, avoiding apparent electronic and structural barriers[27]. Using $m^* = 1.24m_0$, we obtain a Drude scattering time of $\tau = (7.82 \pm 0.13) \times 10^{-13}$ s from the low-field Hall mobility (Fig. 1b), which is in reasonable agreement with the value of τ_q found above. We estimate the ratio between l_{mfp} and the Bardeen–Cooper–Schrieffer coherence length, ξ_{BCS}, to be $1.76\pi k_B T_c \tau/\hbar \approx 0.21$, that is, intermediate between the clean ($\xi_{BCS} \ll l_{mfp}$) and dirty ($\xi_{BCS} \gg l_{mfp}$) limits. Given the Fermi energy that we calculate (Supplementary Information), an approximate value of $l_{mfp} \approx 97$ nm can be estimated for the lowest-energy subband.

In conclusion, we have demonstrated that delta-doped SrTiO$_3$ thin-film heterostructures can open new avenues connecting the previously disparate worlds of high-mobility semiconductors and low-dimensional superconductors. As we approach the limit at which Landau quantization becomes relevant, a potentially new and fascinating world opens before us[12]. Whether these novel phases proposed using idealized models, such as re-entrant superconductivity, can be observed will depend sensitively on the band structure in the delta-doped region, the *g* factor and Zeeman splitting, as well as further improvements in mobility. We also note that heterostructures such as these offer the possibility of engineering low-density quantum superlattices, which are artificial analogues of the layered high-temperature superconductors in which recent observations of Fermi surface oscillations have provoked intense interest[28]. With precise layer control in our heterostructures, it may be possible to tune the coupling between 2D superconducting layers, which has been suggested to be of central importance in understanding the high-temperature superconducting copper oxides[29,30].

METHODS SUMMARY

We grew the sample by pulsed laser ablation in an atmosphere of less than 10^{-8} torr of oxygen at 1,200 °C. Single-crystal STO and NSTO (1 atomic per cent) targets were used together with an STO (100) substrate. After growth, the sample was annealed *in situ* at 900 °C under an oxygen partial pressure of 10^{-2} torr for 30 min to refill oxygen vacancies in the STO layers. We ultrasonically wire-bonded the sample with aluminium wire and then made magnetotransport measurements in an Oxford Instruments Kelvinox MX 400 dilution refrigerator with a base temperature of 10 mK, as calibrated using a ^{60}Co nuclear-orientation thermometer, and a pumped ^{3}He cryostat. *In situ* angular rotation could be performed with a relative accuracy of better than 0.05°, with $\theta = 0^\circ$ defined by the minimum of the Hall voltage. The sample was measured with a 16-Hz a.c. current bias of 100 nA, which was far below the superconducting critical current of 35.9 μA at $T = 50$ mK as measured by d.c. methods.

Received 19 June; accepted 7 October 2009.

1. v. Klitzing, K., Dorda, G. & Pepper, M. New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance. *Phys. Rev. Lett.* **45,** 494–497 (1980).
2. Ando, T., Fowler, A. B. & Stern, F. Electronic properties of two-dimensional systems. *Rev. Mod. Phys.* **54,** 437–672 (1982).
3. Tsui, D. C., Stormer, H. L. & Gossard, A. C. Two-dimensional magnetotransport in the extreme quantum limit. *Phys. Rev. Lett.* **48,** 1559–1562 (1982).
4. Abrahams, E., Kravchenko, S. V. & Sarachik, M. P. Metallic behavior and related phenomena in two dimensions. *Rev. Mod. Phys.* **73,** 251–266 (2001).
5. Ekimov, E. A. *et al.* Superconductivity in diamond. *Nature* **428,** 542–545 (2004).
6. Bustarret, E. *et al.* Superconductivity in doped cubic silicon. *Nature* **444,** 465–468 (2006).
7. Ren, Z.-A. *et al.* Superconductivity in boron-doped SiC. *J. Phys. Soc. Jpn* **76,** 103710 (2007).

©2009 Macmillan Publishers Limited. All rights reserved

8. Herrmannsdörfer, T. *et al.* Superconducting state in a gallium-doped germanium layer at low temperatures. *Phys. Rev. Lett.* **102,** 217003 (2009).
9. Blase, X., Bustarret, E., Chapelier, C., Klein, T. & Marcenat, C. Superconducting group-IV semiconductors. *Nature Mater.* **8,** 375–382 (2009).
10. Goldman, A. M. & Marković, N. Superconductor–insulator transitions in the two-dimensional limit. *Phys. Today* **226,** 39–44 (1998).
11. Schooley, J. F., Hosler, W. R. & Cohen, M. L. Superconductivity in semiconducting $SrTiO_3$. *Phys. Rev. Lett.* **12,** 474–475 (1964).
12. Rasolt, M. & Tešanović, Z. Theoretical aspects of superconductivity in very high magnetic fields. *Rev. Mod. Phys.* **64,** 709–754 (1992).
13. Ohtomo, A., Muller, D. A., Grazul, J. L. & Hwang, H. Y. Artificial charge-modulation in atomic-scale perovskite titanate superlattices. *Nature* **419,** 378–380 (2002).
14. Schubert, E. F. Delta doping of III–V compound semiconductors: fundamentals and device applications. *J. Vac. Sci. Technol. A* **8,** 2980–2996 (1990).
15. Schubert, E. F., Cunningham, J. E. & Tsang, W. T. Electron-mobility enhancement and electron-concentration enhancement in δ-doped n-GaAs at $T = 300$ K. *Solid State Commun.* **63,** 591–594 (1987).
16. Saifi, M. A. & Cross, L. E. Dielectric properties of strontium titanate at low temperature. *Phys. Rev. B* **2,** 677–684 (1970).
17. Sakudo, T. & Unoki, H. Dielectric properties of $SrTiO_3$ at low temperatures. *Phys. Rev. Lett.* **26,** 851–853 (1971).
18. Müller, K. A. & Burkard, H. $SrTiO_3$: an intrinsic quantum paraelectric below 4 K. *Phys. Rev. B* **19,** 3593–3602 (1979).
19. Hulm, J. K., Ashkin, M., Deis, D. W. & Jones, C. K. Superconductivity in semiconductors and semimetals. *Prog. Low Temp. Phys.* **6,** 205–242 (1970).
20. Nakamura, H. *et al.* Low temperature metallic state induced by electrostatic carrier doping of $SrTiO_3$. *Appl. Phys. Lett.* **89,** 133504 (2006).
21. Ueno, K. *et al.* Electric-field-induced superconductivity in an insulator. *Nature Mater.* **7,** 855–858 (2008).
22. Caviglia, A. D. *et al.* Electric field control of the $LaAlO_3/SrTiO_3$ interface ground state. *Nature* **456,** 624–627 (2008).
23. Ohtomo, A. & Hwang, H. Y. Surface depletion in doped $SrTiO_3$ thin films. *Appl. Phys. Lett.* **84,** 1716–1718 (2004).
24. Tinkham, M. Effect of fluxoid quantization on transitions of superconducting films. *Phys. Rev.* **129,** 2413–2422 (1963).
25. Mattheiss, L. F. Effect of the 110°K phase transition on the $SrTiO_3$ conduction bands. *Phys. Rev. B* **6,** 4740–4753 (1972).
26. Uwe, H., Yoshizaki, R., Sakudo, T., Izumi, A. & Uzumaki, T. Conduction band structure of $SrTiO_3$. *Jpn. J. Appl. Phys.* **24** (suppl. 24–2), 335–337 (1985).
27. Yang, M. J. *et al.* Enhancement of cyclotron mass in semiconductor quantum wells. *Phys. Rev. B* **47,** 1691–1694 (1993).
28. Doiron-Leyraud, N. *et al.* Quantum oscillations and the Fermi surface in an underdoped high-T_c superconductor. *Nature* **447,** 565–568 (2007).
29. Anderson, P. W. Interlayer tunneling mechanism for high-T_c superconductivity: comparison with *c* axis infrared experiments. *Science* **268,** 1154–1155 (1995).
30. Chakravarty, S., Kee, H.-Y. & Völker, K. An explanation for a universality of transition temperatures in families of copper oxide superconductors. *Nature* **428,** 53–55 (2004).

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

Acknowledgements We thank M. R. Beasley, A. M. Goldman, A. F. Hebard, A. Kapitulnik, P. B. Littlewood, Y. Liu, A. H. MacDonald and H. Takagi for discussions, and M. Lippmaa for technical assistance.

Author Contributions Y.K. and M.K. performed sample fabrication, measurements and data analysis. C.B., B.G.K., Y.H. and H.Y.H. assisted with the planning, measurements and analysis of the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.Y.H. (hyhwang@k.u-tokyo.ac.jp).

©2009 Macmillan Publishers Limited. All rights reserved

Electrical creation of spin polarization in silicon at room temperature

Saroj P. Dash[1], Sandeep Sharma[1], Ram S. Patel[1], Michel P. de Jong[1] & Ron Jansen[1]

The control and manipulation of the electron spin in semiconductors is central to spintronics[1,2], which aims to represent digital information using spin orientation rather than electron charge. Such spin-based technologies may have a profound impact on nanoelectronics, data storage, and logic and computer architectures. Recently it has become possible to induce and detect spin polarization in otherwise non-magnetic semiconductors (gallium arsenide and silicon) using all-electrical structures[3–9], but so far only at temperatures below 150 K and in n-type materials, which limits further development. Here we demonstrate room-temperature electrical injection of spin polarization into n-type and p-type silicon from a ferromagnetic tunnel contact, spin manipulation using the Hanle effect and the electrical detection of the induced spin accumulation. A spin splitting as large as 2.9 meV is created in n-type silicon, corresponding to an electron spin polarization of 4.6%. The extracted spin lifetime is greater than 140 ps for conduction electrons in heavily doped n-type silicon at 300 K and greater than 270 ps for holes in heavily doped p-type silicon at the same temperature. The spin diffusion length is greater than 230 nm for electrons and 310 nm for holes in the corresponding materials. These results open the way to the implementation of spin functionality in complementary silicon devices and electronic circuits operating at ambient temperature, and to the exploration of their prospects and the fundamental rules that govern their behaviour.

Inducing spin polarization in a semiconductor can be done efficiently and at reasonable current levels by electrical transfer of spins from a ferromagnetic metal across a thin tunnel barrier, as established using optical detection methods for GaAs[10,11] and Si at low temperature[12]. Spin polarization in n-type semiconductors has been detected in all-electrical devices[3–9] at low temperature (5–50 K; in a few cases up to 150 K). Electrical spin detection is often done in a lateral non-local geometry[3,5–7], where the non-local voltage representing the spin polarization in the semiconductor is typically of the order of 10 μV. A second scheme[8,9] uses a single contact for both injection and detection, in a three-terminal geometry (Fig. 1a). We use the latter, single-interface geometry to extract the spin polarization and spin accumulation induced in the semiconductor, the spin lifetime and the variation with temperature, T, and bias voltage, V.

The experiment has three significant features. The first is the electrical injection of a spin-polarized tunnel current from the ferromagnet into the Si, producing an imbalance in the electron population in the Si conduction band or in the hole population in the valence band (see Fig. 1b for n-type Si). This is described by different electrochemical potentials, $\mu^{\uparrow}$ and $\mu^{\downarrow}$, for the up and down spin directions, respectively, and a spin accumulation, $\Delta\mu = \mu^{\uparrow} - \mu^{\downarrow}$. The orientation of the spin polarization is determined by the magnetization direction of the ferromagnet, which is parallel to the interface (that is, in-plane). The spin accumulation is greatest directly underneath the contact and decays with increasing distance from the interface with a certain spin diffusion length, L_{SD}. The second feature is the controlled reduction of the spin accumulation by means of the Hanle effect (Fig. 2a) in an applied magnetic field, B, perpendicular to the carrier spins in the Si. This causes precession of the spins at the Larmor frequency, $\omega_{\mathrm{L}} = g\mu_{\mathrm{B}}B/\hbar$, where g is the Landé g-factor, μ_{B} is the Bohr magneton and $\hbar$ is Planck's constant divided by 2π. As a result, the spin accumulation decays as a function of B with an approximately Lorentzian line shape given by $\Delta\mu(B) = \Delta\mu(0)/(1 + (\omega_{\mathrm{L}}\tau)^2)$, where τ is the spin lifetime (see Supplementary Information for further discussion of the line shape). The third feature of the experiment is the electrical detection of the spin accumulation. This is done using the same tunnel interface, keeping the tunnel current, I, constant and recording the voltage, V, across the contact as B is changed ($V = V_{\mathrm{Si}} - V_{\mathrm{FM}}$, where V_{Si} and V_{FM} are, respectively, the potentials of the Si and the ferromagnetic electrode). For a linear response, the resulting voltage change, ΔV, is equal[13,14] to TSP × $\Delta\mu/2$, where TSP is the known[15,16] tunnel spin polarization of the ferromagnet–insulator interface.

These three features are simultaneously required for a voltage signal to be observed. Hence, the room-temperature (300 K) data shown in Fig. 2b, c demonstrate electrical injection of spin polarization into (n-type) silicon from a ferromagnetic tunnel contact, the

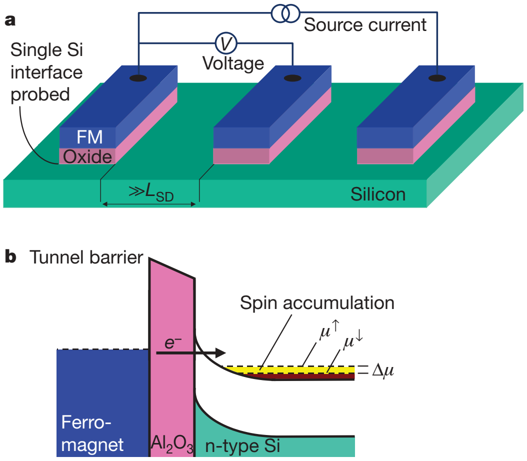


Figure 1 | Device geometry and energy diagram of magnetic contact with n-Si. a, Three-terminal device for injection and detection of spin polarization in Si under a single contact (left) consisting of an oxide insulator and a ferromagnetic-metal electrode (FM; blue). Contacts used to source current (right) and detect the voltage (middle) are placed away from the active interface by more than several spin diffusion lengths (L_{SD}). Each contact has an area of 100 × 200 μm^2. **b,** Energy band profile of the junction, depicting the ferromagnet, the Al_2O_3 barrier and the n-type Si conduction and valence bands bending up towards the oxide, forming a depletion region in the Si that acts as a second part of the tunnel barrier.

[1] MESA+ Institute for Nanotechnology, University of Twente, 7500 AE Enschede, The Netherlands.

©2009 Macmillan Publishers Limited. All rights reserved

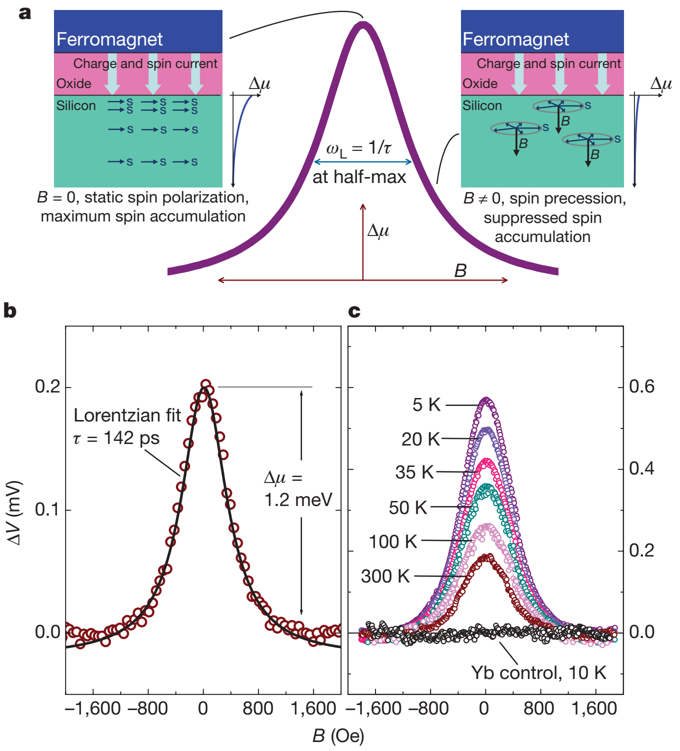


Figure 2 | Electrical injection and detection of a large spin accumulation in n-type Si at 300 K. a, Hanle effect, producing a decay of the net spin accumulation, $\Delta\mu$, due to spin precession in a magnetic field, B, perpendicular to the electron spins (s) in the Si. At constant current, a voltage change, ΔV, across the junction results. **b,** Detected ΔV across an n-Si–Al_2O_3–$Ni_{80}Fe_{20}$(5 nm)–Co(20 nm) tunnel junction at $T = 300$ K, as a function of magnetic field perpendicular to the interface. Data are taken with a constant source current of 734 μA, corresponding to $V = +172$ mV at $B = 0$. The solid line is a Lorentzian fit with $\tau = 142$ ps. **c,** Detected ΔV for various temperatures, as indicated, for the same junction (for all curves, $V = +172$ mV at $B = 0$; the source current varied from 250 μA (5 K) to 734 μA (300 K)). Also shown (black symbols; +172 mV, 730 μA) is data at 10 K for a control device with 2 nm of Yb inserted between the Al_2O_3 and the $Ni_{80}Fe_{20}$ in an otherwise identical junction. Measurement accuracy is represented by the size of the data symbols used.

Hanle precession of the electron spins in the silicon and the electrical detection of the spin accumulation. For constant tunnel current across an n-Si–Al_2O_3–$Ni_{80}Fe_{20}$ junction, we observe that the voltage decreases with increasing applied magnetic field as spin precession gradually reduces $\Delta\mu$ to zero. The signal is reasonably described by a Lorentzian line shape (solid line in Fig. 2b). The slight deviation at the highest B values is discussed in Supplementary Information. Similar data (Fig. 2c) were obtained over the full range of temperatures investigated, with ΔV being larger at low values of T.

Several arguments can be made to exclude the possibility of artefacts contributing to the signal. The resistances of the ferromagnetic metal and the Si between the two voltage probes contribute to the voltage, but they are at least two orders of magnitude smaller than the resistance of the tunnel barrier. The Si showed no significant magnetoresistance due to Lorentz deflection of the electrons by the applied magnetic field, which moreover would have produced a voltage increase as the magnetic field increased. Nevertheless, we performed a decisive test using a control device with 2 nm of non-magnetic Yb inserted between Al_2O_3 and $Ni_{80}Fe_{20}$ in an otherwise identical junction (Fig. 2c, black symbols). This is known[17] to suppress the spin polarization of the injected tunnel current such that $\Delta\mu = 0$, which is what we observed. A similar null result for the Yb control device was obtained over the full range of T and V values investigated. This unambiguously proves that the observed signals are bona fide and represent spin accumulation induced by injection of a spin-polarized tunnel current.

Perhaps the most noteworthy feature is the clear and large spin accumulation observed at room temperature. The magnitude of the spin accumulation at the tunnel interface is obtained from $\Delta V = \text{TSP} \times \Delta\mu/2$, using the known[15,16] TSP value, of 0.3, for Al_2O_3–$Ni_{80}Fe_{20}$ at 300 K. We then obtain $\Delta\mu = 1.2$ meV at 300 K, which is large. From the half-width of the Hanle curve (for which $\omega_L = 1/\tau$), we obtain the spin lifetime $\tau = 142$ ps for our heavily doped n-Si with a measured electron density of 1.8×10^{19} cm^{-3} at 300 K. Although there is no transport data available for comparison, electron spin resonance data[18,19] and recent theory[20] give electron spin lifetimes of about 10 ns at 300 K for low-doped n-Si in which the Elliott–Yafet mechanism due to phonon scattering is dominant. Impurity scattering by the high density of donors in our samples is expected to reduce the spin lifetime. To first order, the spin relaxation time due to the Elliott–Yafet mechanism is given by $\tau_k/4\langle b^2\rangle$, where τ_k is the momentum relaxation time and $\langle b^2\rangle$ is the spin-mixing probability arising from the spin–orbit coupling of the electronic states ($\langle b^2\rangle$ is about 4×10^{-6} for conduction-band electrons in Si at 300 K (ref. 20)). With the value of τ_k derived from the measured mobility (118 cm^2 V^{-1} s^{-1}), this predicts a spin lifetime of about 1 ns, consistent with electron spin resonance data[21,22] for heavily doped n-Si. Our measured value is smaller, suggesting that the spin lifetime is reduced in the proximity of the oxide interface and the ferromagnetic metal electrode. We note that, strictly speaking, we should consider the extracted spin lifetime of 142 ps as a lower bound (Supplementary Information).

We also obtain the spin diffusion length $L_{SD} = \sqrt{D\tau}$ in the Si, where D is the diffusion constant ($D = 3.7$ cm^2 s^{-1} at 300 K as derived from the measured electron mobility). With $\tau = 142$ ps, we then obtain $L_{SD} = 230$ nm at room temperature for our heavily doped n-type Si. Such values are sufficient to transfer spin information over the typical length ($L < 100$ nm) of the channels of modern silicon transistors with only a modest decay of the spin accumulation.

Comparable data was reproducibly obtained from several devices prepared in different runs. Therefore, we can now systematically investigate the factors that control the spin accumulation. Let us first concentrate on n-type Si and examine the influence of the tunnel barrier, which has two parts: the Al_2O_3 tunnel barrier and the Schottky tunnel barrier in the Si due to carrier depletion near the oxide interface (Fig. 1b). The latter is 0.7–0.8 eV high and about 5 nm wide for the Si doping concentration used, making it transparent to tunnelling electrons. We examine whether the spin accumulation is influenced by the presence of this Schottky tunnel barrier by removing it and the associated depletion region by exposing the Si to a flux of Cs before preparation of the Al_2O_3 and the ferromagnetic electrode (Methods Summary). The Cs is known[23] to create states in the Si bandgap close to the conduction-band minimum. For Si–Al_2O_3–ferromagnet structures, this results in an almost flat band condition, as illustrated in the inset of Fig. 3, with a Schottky barrier height of less than 0.2 eV. When the Schottky tunnel barrier is suppressed with Cs, a clear Hanle signal is still observed (Fig. 3a, b). We find that at 300 K, the spin accumulations with and without Cs are of the same order of magnitude, and that the width of the Hanle curve is not changed. Both observations show that the spin accumulation at room temperature is robust and not drastically influenced by the Schottky tunnel barrier in the Si.

From the above result, we conclude that the large value of $\Delta\mu$ at 300 K represents the true spin accumulation in the Si. However, a different behaviour appears below 200 K. For junctions with Cs (no Schottky tunnel barrier), the spin signal changes only weakly with T (Fig. 3c) and the value of τ extracted from the width of the Hanle curve increases at low values of T (Fig. 3d). In sharp contrast, the junctions without Cs show an anomalous enhancement of the spin signal below 200 K, and a peculiar variation of τ, which does not increase at low T values. This anomalous behaviour below 200 K is probably due to two-step tunnelling through localized states at the oxide–semiconductor interface. This was recently proposed[9] to explain the unexpected large

©2009 Macmillan Publishers Limited. All rights reserved

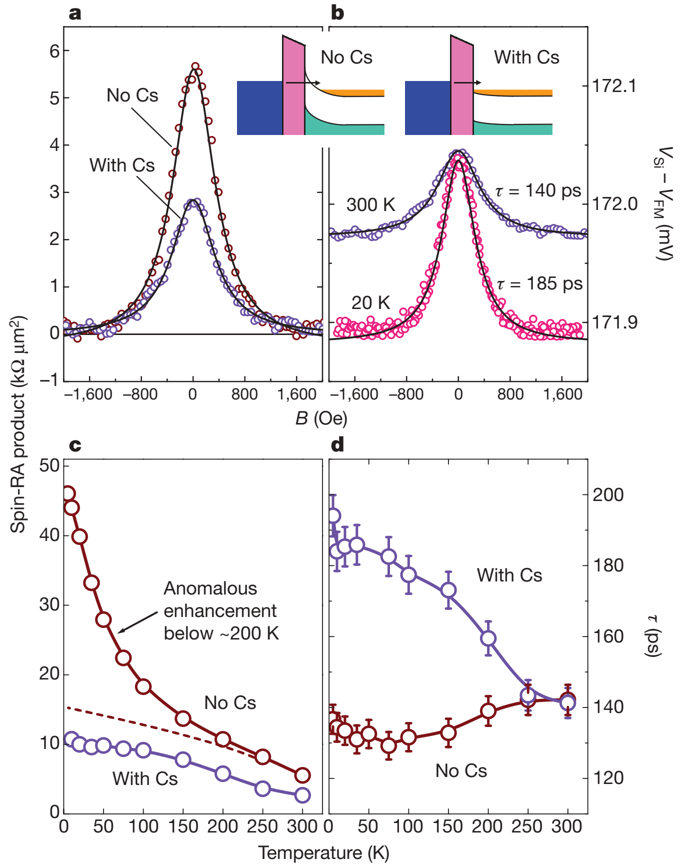


Figure 3 | Spin accumulation in n-type devices with the depletion region of the Si removed by Cs. **a,** Hanle signals at 300 K for junctions without Cs (same data as in Fig. 2) and with Cs, displayed as the spin-RA product. Data are taken with constant source currents of 511 µA and 734 µA for the junctions with and, respectively, without Cs, corresponding to about $V = +172$ mV at $B = 0$. **b,** Hanle curves with Cs at $T = 300$ K and $T = 20$ K. Solid lines are fits to Lorentzians with the τ values as indicated. Inset, energy-band profiles of the junctions with and without Cs. **c,** Spin-RA product versus T with and without Cs. The dashed line projects the expected signal without anomalous enhancement. **d,** The τ values extracted from Lorentzian fits versus T. Measurement accuracy is represented by the size of the data symbols used. Error bars define the range of τ values for which a reasonable fit of the Hanle curve is obtained.

spin signals observed in GaAs–Al_2O_3–Co structures at low temperature. Compared with the semiconductor bulk, such localized states occupy a small volume and, for the same spin-injection current, support a larger spin accumulation as long as they are sufficiently decoupled from the conduction-band states in the Si bulk. This is the case for junctions without Cs, where a Schottky tunnel barrier separates the interface from the bulk. When the Schottky tunnel barrier is removed using Cs, the direct coupling between the interface and the bulk equalizes their spin accumulations, and the enhancement disappears. Hence, our experiments provide direct evidence for the importance of the proposed two-step tunnelling mechanism below 200 K.

The absence of any anomalous signal enhancement for the junctions with Cs (in which there is an Al_2O_3 tunnel barrier only) is evidence that in this case the true spin accumulation in the Si is obtained over the full temperature range. The spin signal, presented in Fig. 3c as the product of spin resistance, $\Delta V/I$, and area (the 'spin-RA product'), should vary with T as $\tau \times \mathrm{TSP}^2$, because $\Delta\mu$ scales with the TSP of the injected current and with τ, and another factor of TSP arises from the detection of the spin accumulation (from $\Delta V = \mathrm{TSP} \times \Delta\mu/2$). Using the values of τ extracted from the width of the Hanle curve, and that fact that $\mathrm{TSP} \propto (1 - \alpha T^{3/2})$ with $\alpha = (3–5) \times 10^{-5}\ \mathrm{K}^{-3/2}$ as previously determined[24], at low temperature we can expect the signal for the junctions with Cs to increase by a factor of 2.5. This is not too different from the factor of four observed. The increase in the extracted τ values, from 140 ps at 300 K to about 190 ps at low temperature, is reasonable for a spin relaxation time[20,21]. The conventional formula, $\tau_k/4\langle b^2\rangle$, for spin relaxation due to the Elliott–Yafet mechanism predicts a modest increase at low temperature. The measured mobility (which is directly proportional to τ_k) changes by less than 5%, whereas $\langle b^2\rangle$ was calculated[20] to decrease by 30–50% at low temperature. The observed increase in τ, of 35%, is consistent with this.

An important question is how large $\Delta\mu$ can be and how it varies with applied bias voltage (or current). We find (Fig. 4) that below 200 K, the spin-RA product is anomalously large owing to the contribution of two-step tunnelling through interface states, as discussed. Above 200 K, this contribution is negligible, and the data at 300 K is believed to represent the intrinsic behaviour. The spin-RA product at 300 K is asymmetric with respect to bias polarity, decreasing significantly for $V < 0$ for extraction of electrons from the Si, but depending only weakly on V for injection of electrons into the Si at $V > 0$ (note that the spin detection efficiency also varies with V). A constant spin-RA product (300 K and $V > 0$) implies that the induced spin accumulation scales linearly with current, reaching a maximum of $\Delta\mu \approx 2.9$ meV ($\Delta V = 0.43$ mV and $\mathrm{TSP} = 0.3$) for the largest current (+1.5 mA). Assuming a parabolic conduction band and a Fermi–Dirac distribution for each spin, this translates into densities of $0.94 \times 10^{19}\ \mathrm{cm}^{-3}$ and $0.86 \times 10^{19}\ \mathrm{cm}^{-3}$ for majority and, respectively, minority spin electrons at room temperature and a sizeable electron spin polarization of 4.6% in the n-type Si.

Next we describe spin polarization in p-type Si at room temperature. The polarization is created in the valence band and the electronic carriers are holes. Results are shown in Fig. 5 for boron-doped p-type Si with a measured hole density of $4.8 \times 10^{18}\ \mathrm{cm}^{-3}$ at 300 K. A clear Hanle signal is observed (Fig. 5a), demonstrating electrically induced spin polarization of holes in the valence band of p-type silicon, the spin precession of the holes and the electrical detection of the spin accumulation of holes. From the width of the Hanle curve, we extract a value of $\tau = 270$ ps for the hole spin lifetime at 300 K, which is larger than that for electrons in n-type Si (Fig. 2). Comparing with the conduction band, a stronger spin–orbit coupling strength in the Si valence band, and hence a smaller value of τ, might be expected. This is apparently compensated for by the density of acceptor impurities in the p-type sample being less than the donor impurity density in the n-type samples. We have used the free-electron g-factor, $g = 2$, also for valence band holes, in the absence of unique and accurate data[25]. If the g-factor for holes is different, the value of τ has to be adjusted correspondingly. For $\tau = 270$ ps and the measured hole mobility of $117\ \mathrm{cm}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1}$ ($D = 3.6\ \mathrm{cm}^2\ \mathrm{s}^{-1}$), we obtain a hole

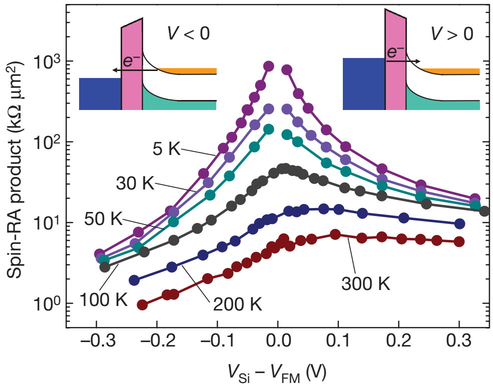


Figure 4 | Variation of spin signals with applied bias voltage in n-type Si devices. Spin-RA product as a function of applied bias voltage, V, at different temperatures, as indicated, for the same junction as in Fig. 2. For $V > 0$ and $V < 0$, spin-polarized electrons are injected into and, respectively, extracted from the Si conduction band, as sketched in the insets. Measurement accuracy is represented by the size of the data symbols used.

©2009 Macmillan Publishers Limited. All rights reserved

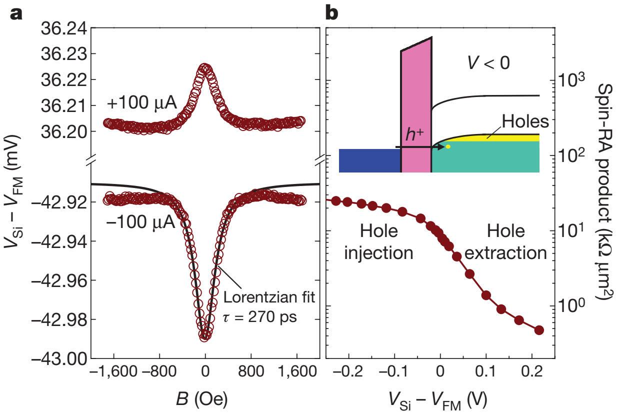


Figure 5 | Spin accumulation of holes in p-type Si at 300 K. a, Detected ΔV across a p-Si–Al_2O_3–$Ni_{80}Fe_{20}$ tunnel junction at $T = 300$ K, as a function of applied magnetic field, B. Data for the two curves are taken with a constant current of either −100 μA or +100 μA, as indicated. The solid line is a Lorentzian fit with $\tau = 270$ ps. **b,** Spin-RA product versus applied bias voltage at 300 K. Inset, energy-band diagram for $V < 0$, in which spin-polarized holes (h^+) tunnel from the ferromagnetic metal into the valence band of the Si, where they are added to the pre-existing holes (yellow). This is equivalent to electrons tunnelling from filled states (green) in the Si valence band into empty states in the ferromagnetic metal. Measurement accuracy is represented by the size of the data symbols used.

spin diffusion length of $L_{SD} = 310$ nm at room temperature for our p-type Si. Figure 5b shows the variation of the spin-RA product with V for p-type devices. Just as for n-type Si, this product is nearly constant as a function of bias voltage for the polarity in which (hole) carriers are injected into the Si ($V < 0$ in this case), and exhibits a faster decay when the spin accumulation is created by extracting (hole) carriers from the Si ($V > 0$).

An elementary estimation of the steady state value of $\Delta\mu$, balancing the net amount of injected spins with an equal amount of spin flips in the Si per unit time[13,14], predicts a $\Delta\mu$ value about two orders of magnitude smaller than that observed. This may in part be due to a possible underestimation of the extracted spin lifetime (Supplementary information). However, we propose that another likely factor is the lateral inhomogeneity of the tunnel current. This is well known to exist in tunnel junctions as a result of the thickness and composition variations of the barrier. The real (local) tunnel current density that determines the spin accumulation may then be significantly larger than the average current density calculated from the geometric contact area.

The electrical creation and detection of a large and robust spin accumulation in Si at room temperature is a useful advance given the prevalence of Si in semiconductor technology. The scaling of $\Delta\mu$ with current density implies that even larger values should be feasible with optimized low-resistance contacts. Contact materials with larger TSP values can be used, and a larger spin lifetime may be obtained for Si with a lower doping density and/or optimized interfaces. This and other characteristics, and the fundamental rules that govern the behaviour of spin in Si devices at room temperature, can now be further explored.

METHODS SUMMARY

We fabricated the Si/Al_2O_3/ferromagnetic metal contacts on Si (100) substrates as previously described[15]. The n-type silicon-on-insulator wafer has a 5-μm thick active Si layer with As doping and a resistivity of 3 mΩ cm at 300 K. The p-type silicon-on-insulator wafer has a 3-μm thick active Si layer with B doping and a resistivity of 11 mΩ cm at 300 K. After surface treatment by hydrofluoric acid to remove oxide, the substrate was introduced into the load-lock chamber in which, if desired, it was exposed to Cs using a Cs alkali-metal dispenser[26] (SAES Getters). The current through the dispenser was increased in steps to 6 A in 18 min and kept at 6 A for another 15 min, and the pressure was constant at 10^{-7} mbar. After its transfer into the ultrahigh-vacuum chamber, we prepared the tunnel barrier by electron-beam deposition of Al_2O_3 (with nominal thicknesses of 0.5 nm for n-type Si and 0.7 nm for p-type Si) from an Al_2O_3 single-crystal source, followed by plasma oxidation for 2.5 min and electron-beam deposition of the ferromagnetic-metal top electrode.

Received 2 August; accepted 8 October 2009.

1. Žutić, I., Fabian, J. & Das Sarma, S. Spintronics: fundamentals and applications. *Rev. Mod. Phys.* **76,** 323–410 (2004).
2. Chappert, C., Fert, A. & Nguyen van Dau, F. The emergence of spin electronics in data storage. *Nature Mater.* **6,** 813–823 (2007).
3. Lou, X. *et al.* Electrical detection of spin transport in lateral ferromagnet-semiconductor devices. *Nature Phys.* **3,** 197–202 (2007).
4. Appelbaum, I., Huang, B. & Monsma, D. J. Electronic measurement and control of spin transport in silicon. *Nature* **447,** 295–298 (2007).
5. van 't Erve, O. M. J. *et al.* Electrical injection and detection of spin-polarized carriers in silicon in a lateral transport geometry. *Appl. Phys. Lett.* **91,** 212109 (2007).
6. Ando, Y. *et al.* Electrical injection and detection of spin-polarized electrons in silicon through Fe_3Si/Si Schottky tunnel barrier. *Appl. Phys. Lett.* **94,** 182105 (2009).
7. Ciorga, M. *et al.* Electrical spin injection and detection in lateral all-semiconductor devices. *Phys. Rev. B* **79,** 165321 (2009).
8. Lou, X. *et al.* Electrical detection of spin accumulation at a ferromagnet–semiconductor interface. *Phys. Rev. Lett.* **96,** 176603 (2006).
9. Tran, M. *et al.* Enhancement of the spin accumulation at the interface between a spin-polarized tunnel junction and a semiconductor. *Phys. Rev. Lett.* **102,** 036601 (2009).
10. Hanbicki, A. T., Jonker, B. T., Itskos, G., Kioseoglou, G. & Petrou, A. Efficient electrical spin injection from a magnetic metal/tunnel barrier contact into a semiconductor. *Appl. Phys. Lett.* **80,** 1240–1242 (2002).
11. Motsnyi, V. F. *et al.* Electrical spin injection in a ferromagnet/tunnel barrier/semiconductor heterostructure. *Appl. Phys. Lett.* **81,** 265–267 (2002).
12. Jonker, B. T., Kioseoglou, G., Hanbicki, A. T., Li, C. H. & Thompson, P. E. Electrical spin-injection into silicon from a ferromagnetic metal/tunnel barrier contact. *Nature Phys.* **3,** 542–546 (2007).
13. Fert, A. & Jaffrès, H. Conditions for efficient spin injection from a ferromagnetic metal into a semiconductor. *Phys. Rev. B* **64,** 184420 (2001).
14. Osipov, V. V. & Bratkovsky, A. M. Spin accumulation in degenerate semiconductors near modified Schottky contact with ferromagnets: spin injection and extraction. *Phys. Rev. B* **72,** 115322 (2005).
15. Min, B. C., Motohashi, K., Lodder, J. C. & Jansen, R. Tunable spin-tunnel contacts to silicon using low-work-function ferromagnets. *Nature Mater.* **5,** 817–822 (2006).
16. Park, B. G., Banerjee, T., Lodder, J. C. & Jansen, R. Tunnel spin polarization versus energy for clean and doped Al_2O_3 barriers. *Phys. Rev. Lett.* **99,** 217206 (2007).
17. Patel, R. S., Dash, S. P., de Jong, M. P. & Jansen, R. Magnetic tunnel contacts to silicon with low-work-function ytterbium nanolayers. *J. Appl. Phys.* **106,** 016107 (2009).
18. Lepine, D. J. Spin resonance of localized and delocalized electrons in phosphorus-doped silicon between 20 and 300 K. *Phys. Rev. B* **2,** 2429–2439 (1970).
19. Fabian, J., Matos-Abiague, A., Ertler, C., Stano, P. & Žutić, I. Semiconductor spintronics. *Acta Phys. Slov.* **57,** 565–907 (2007).
20. Cheng, J. L., Wu, M. W. & Fabian, J. Theory of the spin relaxation of conduction electrons in silicon. Preprint at ⟨http://arxiv1.library.cornell.edu/abs/0906.4054⟩ (2009).
21. Kodera, H. Effect of doping on the electron spin resonance in phosphorus doped silicon. II. *J. Phys. Soc. Jpn* **21,** 1040–1045 (1966).
22. Anderberg, J. M., Einevoll, G. T., Vier, D. C., Schultz, S. & Sham, L. J. Probing the Schottky barrier with conduction electron spin resonance. *Phys. Rev. B* **55,** 13745–13751 (1997).
23. Biagi, R. *et al.* Photoemission investigation of alkali-metal-induced two-dimensional electron gas at the Si(111)(1×1):H surface. *Phys. Rev. B* **67,** 155325 (2003).
24. Shang, C. H., Nowak, J., Jansen, R. & Moodera, J. S. Temperature dependence of magnetoresistance and surface magnetization in ferromagnetic tunnel junctions. *Phys. Rev. B* **58,** R2917–R2920 (1998).
25. Feher, G., Hensel, J. C. & Gere, E. A. Paramagnetic resonance absorption from acceptors in silicon. *Phys. Rev. Lett.* **5,** 309–311 (1960).
26. Succi, M., Canino, R. & Ferrario, B. Atomic-absorption evaporation flow-rate measurements of alkali metal dispensers. *Vacuum* **35,** 579–582 (1985).

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

Acknowledgements This work was financially supported by the NWO-VIDI programme and the Netherlands Foundation for Fundamental Research on Matter.

Author Contributions S.P.D. fabricated most of the devices and carried out most of the measurements. S.S. and M.P.d.J. contributed to the device fabrication and some of the measurements. R.S.P. contributed to the Yb control experiment. All co-authors contributed important insights and ideas. R.J. supervised and coordinated the research. R.J. and S.P.D. wrote the paper, with help from all co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.J. (ron.jansen@el.utwente.nl).

©2009 Macmillan Publishers Limited. All rights reserved

Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies

A. Biastoch[1], C. W. Böning[1], F. U. Schwarzkopf[1] & J. R. E. Lutjeharms[2]

The transport of warm and salty Indian Ocean waters into the Atlantic Ocean—the Agulhas leakage—has a crucial role in the global oceanic circulation[1] and thus the evolution of future climate. At present these waters provide the main source of heat and salt for the surface branch of the Atlantic meridional overturning circulation (MOC)[2]. There is evidence from past glacial-to-interglacial variations in foraminiferal assemblages[3] and model studies[4] that the amount of Agulhas leakage and its corresponding effect on the MOC has been subject to substantial change, potentially linked to latitudinal shifts in the Southern Hemisphere westerlies[5]. A progressive poleward migration of the westerlies has been observed during the past two to three decades and linked to anthropogenic forcing[6], but because of the sparse observational records it has not been possible to determine whether there has been a concomitant response of Agulhas leakage. Here we present the results of a high-resolution ocean general circulation model[7,8] to show that the transport of Indian Ocean waters into the South Atlantic via the Agulhas leakage has increased during the past decades in response to the change in wind forcing. The increased leakage has contributed to the observed salinification[9] of South Atlantic thermocline waters. Both model and historic measurements off South America suggest that the additional Indian Ocean waters have begun to invade the North Atlantic, with potential implications for the future evolution of the MOC.

The Agulhas leakage is the result of a complex, highly nonlinear interplay between the strong western boundary current (WBC) along the South African coast, the Agulhas Current[10], and vigorous mesoscale activity arising in its source regions[11] and south of Africa where the bulk of the Agulhas Current waters are retroflected back into the Indian Ocean. As part of the retroflection process, the intermittent formation of intense oceanic eddy structures—Agulhas rings—carries warm and salty Indian Ocean water into the South Atlantic. The leakage can affect the MOC in two ways. (1) The mesoscale activity in the retroflection regime induces wave processes in the South Atlantic that dynamically modulate the MOC on decadal timescales[7]. (2) On longer timescales, it has been demonstrated in idealized studies[12] that the northward advection of salinity anomalies from the Agulhas regime influences deep-water formation in the northern North Atlantic.

Owing to its mean latitudinal position south of Africa, the zero line of the wind stress curl permits an interoceanic connection of the subtropical gyres of the South Indian and the Atlantic Ocean, a 'supergyre'[13] (Fig. 1a). Studies of atmospheric observations and reanalyses have noted a poleward intensification of the westerly winds during the last decades[6], a trend that is projected to continue during the twenty-first century[14]. How have these changes to the wind field affected the Agulhas system[10], in particular interoceanic transport? We used a high-resolution (1/10°) model of the greater Agulhas region (green box in Fig. 2) that has been demonstrated to realistically simulate the complicated circulation around South Africa[7,8,15]. The model is nested into a global ocean/sea-ice model which by itself would significantly over-estimate the Agulhas leakage owing to its coarse (1/2°) resolution[8]. In addition to the reference experiment (AG01-R) which provides a hindcast simulation subject to the atmospheric forcing variability of the last decades[16], two sensitivity experiments are considered: AG01-C is driven by a repeated-year forcing, so it inherently explicitly excludes any inter-annual (and anthropogenic) forcing trend; AG01-S[8] uses a variant of the high-resolution nesting domain, exploring the sensitivity of the leakage behaviour to an omission of the mesoscale activity in the upstream regions of the Agulhas Current.

The model hindcast shows that during the past decades the supergyre has extended poleward by about 2° of latitude, a direct consequence of the poleward shift of the westerlies (Fig. 1b). Such behaviour is not produced by the sensitivity experiment under

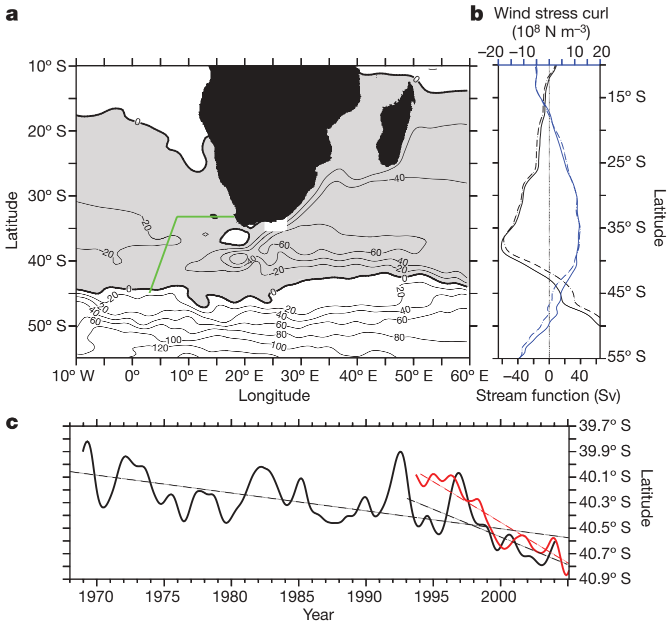


Figure 1 | Large-scale circulation changes south of Africa. a, Time-mean (1995–2004) horizontal streamfunction in the Agulhas region (contours marked in Sverdrups) in AG01-R, with grey shading denoting anticyclonic circulation. The GoodHope section used for the quantification of Agulhas leakage is marked by the green line. **b**, Latitudinal dependence of zonal averages (20°–60° E) of the streamfunction (black) and zonally averaged wind stress curl over the Indian Ocean (20°–110° E, blue) for periods 1965–1974 (dashed) and 1995–2004 (solid). **c**, Latitude of zero sea surface height in AG01-R (0°–40° E zonal average, black) and a corresponding constant sea surface height line in satellite data (Aviso, red). Dashed lines indicate linear trends over full time range and over the past decade.

[1]Leibniz-Institut für Meereswissenschaften, Düsternbrooker Weg 20, 24105 Kiel, Germany. [2]Department of Oceanography, University of Cape Town, 7700 Rondebosch, South Africa.

©2009 Macmillan Publishers Limited. All rights reserved

repeated-year forcing (AG01-C), confirming that we can neglect possible effects of spurious model trends and attribute the shift to the decadal changes in the external forcing. An observational test of the change in the geometry of the super-gyre is provided by satellite altimeter data, available since the beginning of the 1990s: the observed sea surface height pattern shows a southward migration, similar to that of the model simulation (Fig. 1c), with a clear increase in the past decade.

As a consequence of the trend in the atmospheric forcing, and consistent with studies of regional ocean observations[17], the waters in the southwest Indian Ocean exhibit warming and salinification tendencies (Fig. 2; Supplementary Fig. 1). The model simulation captures the observed trends[18], and indicates the regional changes as part of a larger pattern that zonally extends into the South Atlantic. The warming/salinification can be explained by a southward shift (Supplementary Fig. 2) of the boundary between the subtropical gyre circulation and the Antarctic Circumpolar Current, and thus considered to be part of the hemispheric-scale poleward migration of this frontal zone[19]. (Note that the warming pattern is absent in the sensitivity experiment under repeated-year forcing, AG01-C.)

Another, potentially even more important, consequence of the forcing trend occurs in the Agulhas leakage (Fig. 3). For a rigorous determination of Agulhas leakage we trace the amount of water originating in the Agulhas Current at 32° S and arriving at the GoodHope section (see ref. 20 and green line in Fig. 1a) using a Lagrangian tracking technique[8,21,22]. Partially masked by a strong year-to-year variability, there has been a significant trend of 1.2 Sv ($1\ \text{Sv} = 1 \times 10^6\ \text{m}^3\ \text{s}^{-1}$) per decade, resulting in a total increase of more than 5 Sv over the course of the integration.

Of potential relevance to ocean monitoring efforts[20,23] is the manifestation of circulation changes in the WBC system east of Africa. The model simulation demonstrates that it is not possible to infer the WBC changes from linear vorticity dynamics, as attempted in previous calculations[6,18] (Fig. 4a): there is no simple relation between the Sverdrup transport variability calculated from the wind stress curl over the Indian Ocean and the actual transport variability, presumably because of the strong topographic effects shielding the Agulhas region from the east[24] and the inherent nonlinearities of eddy–mean flow interaction in the WBC[15]. The importance of the WBC nonlinearities for the generation of inter-annual to decadal transport variability is elucidated by the sensitivity experiments (Fig. 4b). AG01-C, although forced without inter-annual variability, exhibits low-frequency transport variations of similar intensity (but different phase). In contrast, AG01-S, a sensitivity experiment with a smaller nesting domain excluding Mozambique eddies[11], produces, although under identical forcing, a temporal variation differing from the reference experiment. The sequence demonstrates that inter-annual transport variations in the Agulhas regime are largely decoupled from the large-scale wind field and governed primarily by the internal dynamics in the WBC regime. Only on longer, decadal timescales does the effect of these eddies appear to fade, and transport changes with (AG01-R) and without (AG01-S) Mozambique eddies begin to exhibit a clear correlation ($r = 0.95$, Supplementary Fig. 3). The model results have implications for ongoing ocean-monitoring efforts at the western boundary: they suggest that observations of a few years' duration are of limited value as an index of inter-annual gyre-scale transport changes in the southern Indian Ocean.

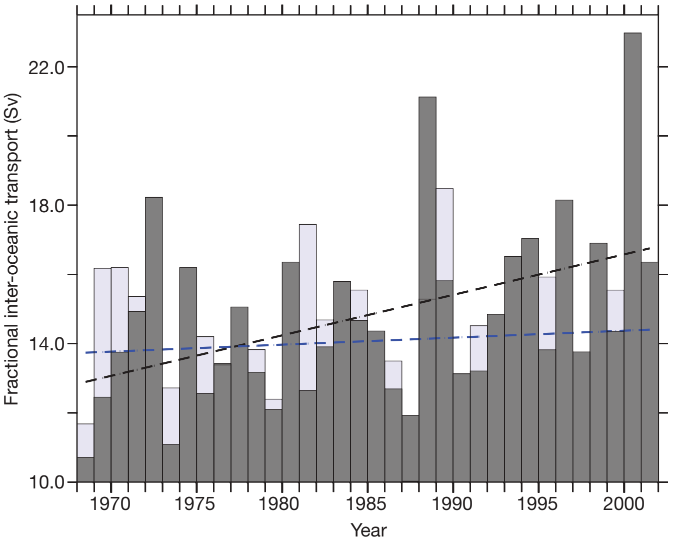


Figure 3 | Increase of Agulhas leakage. Inter-oceanic transport as obtained by float releases within the southward-flowing Agulhas Current at 32° S: fractional transports (in Sverdrups) across the GoodHope line in the Cape Basin (green line in Fig. 1a) in the reference experiment (AG01-R, grey bars) and the repeated-year (AG01-C, light-blue bars) experiment. The dashed lines mark the linear trend of 1.2 Sv per decade in AG01-R (black) and 0.2 Sv in AG01-C (blue).

As suggested by theoretical arguments[25], an increase (decrease) in the strength of the Agulhas Current should translate into a decrease (increase) of the leakage; a corresponding behaviour was noted for the present model[22]. It also holds for the long-term trend in which the increasing transport from the Indian to the Atlantic Ocean (1.2 Sv per decade) is linked to a multi-decadal decrease of the Agulhas Current transport (−1.4 Sv per decade). The observed warming south of Africa is therefore an expression of the large-scale frontal changes, rather than an advective effect of the Agulhas Current[18] (Supplementary Fig. 4). The simulation suggests that this 10% reduction of the WBC can be

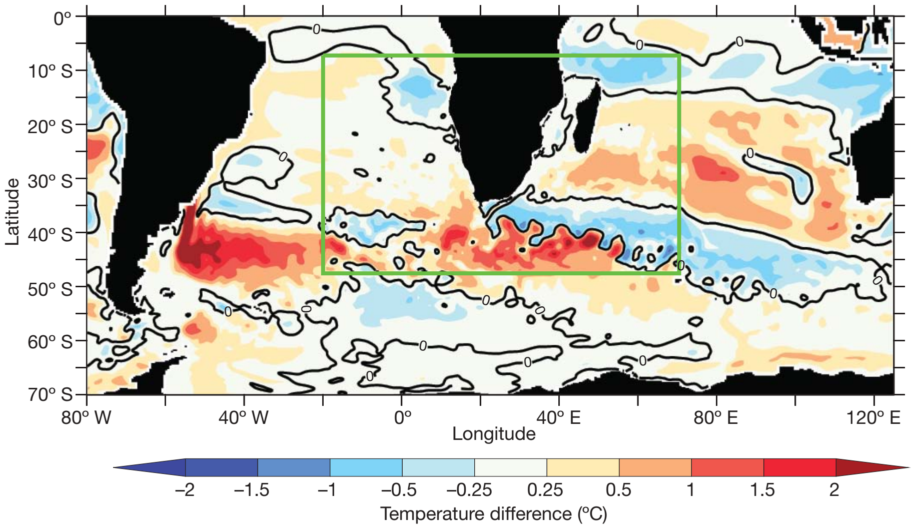


Figure 2 | Thermocline changes in the Southern Hemisphere. The colour scale shows the difference (2000–2004 minus 1968–1972) in the upper ocean (0–200 m) temperature (in °C), simulated in AG01-R. The green box denotes the boundaries of the high-resolution nest. The zero contour line (black) separates cooling from warming areas.

©2009 Macmillan Publishers Limited. All rights reserved

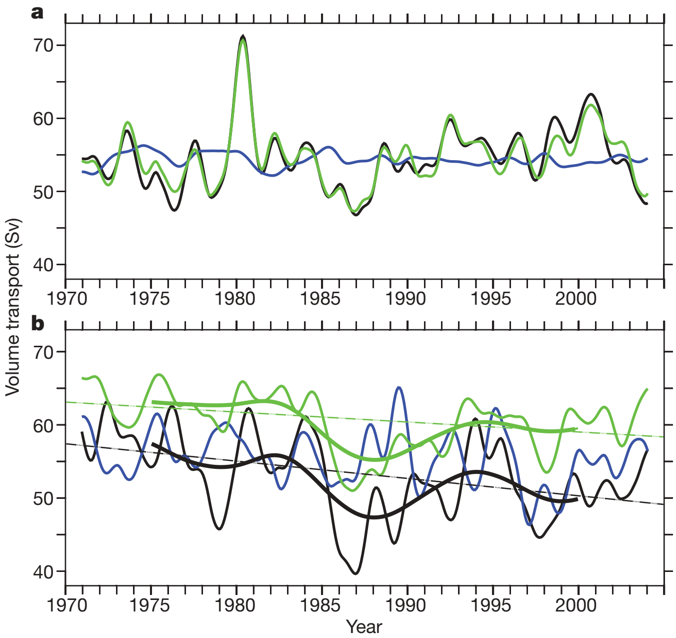


Figure 4 | Inter-annual to decadal variability in the Agulhas Current. **a**, Inter-annual variability of the Sverdrup transport at 32° S, calculated from the atmospheric wind stress (CORE[16]), for AG01-R (black), AG01-C (blue) and AG01-S (green). **b**, Agulhas Current transport at 32° S, calculated across the WBC and its first recirculation (line colours as in **a**), and filtered with a 23-month (thin curves) and 121-month Hanning window (bold). Long-term linear trend (dashed) of the strength of the WBC. (For alternative calculations of the WBC transport, see Supplementary Fig. 3.)

attributed to a combination of decreased transport through the Mozambique Channel (Supplementary Fig. 5), which is sensitive to shifts in the wind fields[26], and a reduced recirculation in the Southwest Indian Ocean subgyre[10].

What are the consequences of the increased Agulhas leakage for the large-scale circulation in the Atlantic Ocean? The present model series does not permit the effect of the leakage trend on the MOC transport to be isolated because the model includes other potential effects of changes in the westerlies, which may affect the MOC via changes in the Antarctic Circumpolar Current and its associated deep- and intermediate-water formation processes in the Southern Ocean[27]. However, by using the Lagrangian tracking technique we can follow and inspect the properties of the waters that enter the South Atlantic across the GoodHope section (Fig. 5). Consistent with the comparison of Agulhas leakage in the stand-alone base model and the nested model[8], the net volume transport towards the North Atlantic did not change from the 1970s to the 2000s; the increased leakage instead led to an enhanced horizontal super-gyre. However, there has been a striking trend in the freshwater transport: the inter-hemispheric export of salt originating from Agulhas leakage has increased by 25% over the course of the integration.

In the South Atlantic, almost all of the inter-hemispheric, northward volume transport (that is, the upper-layer branch of the MOC) is channelled through the North Brazil Current[1]. Corresponding to the increase in the northward salt transport, the model simulation reveals an increasing trend in the salinity of the North Brazil Current core. We have assessed this behaviour by performing an analysis of salinity profiles collected in historic ocean data archives (Fig. 5 inset). The observed records show a salinity increase in the North Brazil Current during the 1970s and 1980s, corresponding to the observed salinification of the subtropical thermocline waters during the past decade, which has been attributed to changes in the hydrological cycle[9]. The model simulation suggests that the majority of this salinification can be traced to the increased invasion of Indian Ocean waters as a result of wind-driven changes.

The suite of model experiments highlights a far-reaching consequence of the anthropogenic shifts in the Southern Hemisphere westerlies—the salinity increase described above—which is projected to continue and accelerate during the twenty-first century[14]. An increased import of salty water into the northward branch of the inter-hemispheric MOC could eventually become a significant factor for the freshwater budget of the deep-water formation regions in the subpolar North Atlantic[12]. To what degree it could help to stabilize a potentially declining 'Gulf Stream system' caused by subarctic freshening in a warming climate[28] needs to be investigated using coupled

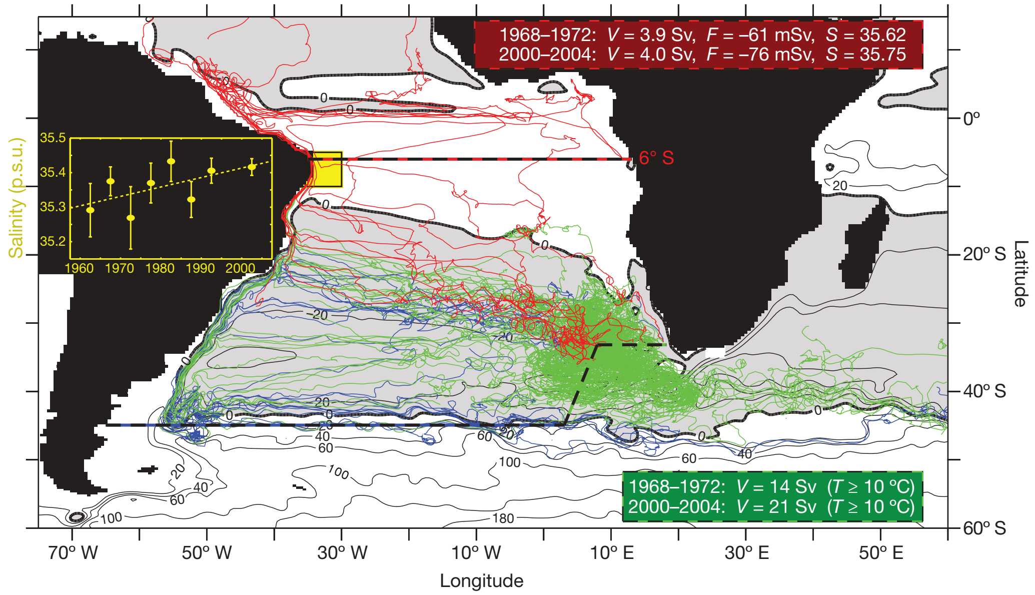

1968–1972: V = 3.9 Sv, F = –61 mSv, S = 35.62
2000–2004: V = 4.0 Sv, F = –76 mSv, S = 35.75
1968–1972: V = 14 Sv (T ≥ 10 °C)
2000–2004: V = 21 Sv (T ≥ 10 °C)
6° S
Salinity (p.s.u.)
35.5
35.4
35.3
35.2
1960
1970
1980
1990
2000
0
20
40
60
100
180
0°
20° S
40° S
60° S
Latitude
70° W
50° W
30° W
10° W
10° E
30° E
50° E
Longitude

Figure 5 | Pathways of the Agulhas leakage into the North Atlantic. Example trajectories of virtual floats released with temperature $T \geq 10$ °C along the GoodHope section and leaving the South Atlantic towards the Indian Ocean (green), the Southern Ocean (blue) or the North Atlantic (red). For statistical numbers see Supplementary Table 1. Volume (V) and freshwater (F) transports (negative numbers indicate salt advection to the north) are shown for the full 6° S (red-black dashed line) and GoodHope (black dashed) sections for the indicated periods, and the mean salinity (S) of floats leaving the domain in the core of the North Brazil Current (depth range, 100–600 m). The inset shows an analysis of historic salinity profiles (in practical salinity units, p.s.u.) averaged over the same depth range off the east coast of South America (yellow box). The error bars depict 2σ errors (95% confidence intervals). The shading and the contour lines depict the streamfunction as in Fig. 1a.

©2009 Macmillan Publishers Limited. All rights reserved

ocean–atmosphere models and dedicated sensitivity studies. Our model results emphasize the need to capture realistically the mesoscale processes of the Agulhas leakage regime in climate model projections of future MOC evolution, and the importance of ocean-monitoring programmes in key areas for inter-ocean and inter-hemispheric transports.

METHODS SUMMARY

For a realistic model simulation of the Agulhas Current system the horizontal resolution is a critical factor. Here we use a high-resolution (1/10°) model (green box in Fig. 2) nested[29] into a coarser (1/2°) global ocean/sea-ice model based on the NEMO code (v2.3)[30], developed by the DRAKKAR collaboration. Previous studies demonstrated the success of this set-up in reproducing the salient circulation features of the Agulhas system[8,15] and its dynamic impact on the Atlantic MOC[7]. In addition to the reference experiment (AG01-R) in which surface forcing fields[16] were applied over the period 1958–2004, two sensitivity experiments have been performed. AG01-C, without inter-annual variability in the forcing fields, allows us, by comparison with AG01-R, to isolate the internal variability and a possible spurious model drift from the changes by the forcing fields. In AG01-S, the high-resolution nest terminates at 27° S, thus excluding mesoscale eddies in the source regions of the Agulhas Current[8].

A Lagrangian method[21] was used to quantify the Agulhas leakage. Virtual floats, each seeded as a fraction of the Agulhas Current transport at 32° S, were advected by the time-dependent three-dimensional flow field. Summing up their individual transports at the GoodHope section[20] resulted in a total number for the leakage transport. Another float integration elucidated the longer-term paths of the leakage towards the North Atlantic by seeding at the GoodHope section and advecting with the velocities of two different pentads (five-year periods) from the model. The model fields are compared with gridded time series of sea surface height (http://www.aviso.oceanobs.com) based on satellite altimeter data, produced by Ssalto/Duacs and distributed by Aviso, with support from CNES. The mean dynamic topography Rio05 was produced by the CLS Space Oceanography Division. To assess temporal changes in salinity fields in the North Brazil Current near 6° S, we used an extensive compilation of historical observations by BLUElink (ftp://ftp.marine.csiro.au/pub/omas/BOA) and CARS (http://www.marine.csiro.au/~dunn/cars2006).

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

Received 15 July; accepted 16 September 2009.

1. Gordon, A. L. Interocean exchange of thermocline water. *J. Geophys. Res.* **91,** 5037–5046 (1986).
2. Friocourt, Y., Drijfhout, S., Blanke, B. & Speich, S. Water mass export from Drake Passage to the Atlantic, Indian, and Pacific oceans: a Lagrangian model analysis. *J. Phys. Oceanogr.* **35,** 1206–1222 (2005).
3. Peeters, F. J. C. *et al.* Vigorous exchange between Indian and Atlantic Ocean at the end of the last five glacial periods. *Nature* **400,** 661–665 (2004).
4. Knorr, G. & Lohmann, G. Southern Ocean origin for the resumption of Atlantic thermohaline circulation during deglaciation. *Nature* **424,** 532–536 (2003).
5. Bard, E. & Rickaby, R. E. M. Migration of the subtropical front as a modulator of glacial climate. *Nature* **460,** 380–383 (2009).
6. Cai, W. Antarctic ozone depletion causes an intensification of the Southern Ocean super-gyre circulation. *Geophys. Res. Lett.* **33,** L03712 (2006).
7. Biastoch, A., Böning, C. W. & Lutjeharms, J. R. E. Agulhas leakage dynamics affects decadal variability in atlantic overturning circulation. *Nature* **456,** 489–492 (2008).
8. Biastoch, A., Lutjeharms, J. R. E., Böning, C. W. & Scheinert, M. Mesoscale perturbations control inter-ocean exchange south of Africa. *Geophys. Res. Lett.* **35,** L20602 (2008).
9. Curry, R. & Mauritzen, C. Dilution of the northern North Atlantic Ocean in recent decades. *Science* **308,** 1772–1774 (2005).
10. Lutjeharms, J. R. E. *The Agulhas Current* (Springer, 2006).
11. De Ruijter, W. P. M., Ridderinkhof, H., Lutjeharms, J. R. E., Schouten, M. W. & Veth, C. Observations of the flow in the Mozambique Channel. *Geophys. Res. Lett.* **29,** 140–141 (2002).
12. Weijer, W., de Ruijter, W. P. M., Sterl, A. & Drijfhout, S. S. Response of the Atlantic overturning circulation to South Atlantic sources of buoyancy. *Glob. Planet. Change* **34,** 293–311 (2002).
13. Speich, S., Blanke, B. & Cai, W. Atlantic meridional overturning circulation and the Southern Hemisphere supergyre. *Geophys. Res. Lett.* **34,** L23614 (2007).
14. Sen Gupta, A. *et al.* Projected changes to the Southern Hemisphere ocean and sea-ice in the IPCC AR4 climate models. *J. Clim.* **22,** 3047–3078 (2009).
15. Biastoch, A., Beal, L. M., Casal, T. G. D. & Lutjeharms, J. R. E. Variability and coherence of the Agulhas Undercurrent in a high-resolution ocean general circulation model. *J. Phys. Oceanogr.* **39,** 2417–2435 (2009).
16. Large, W. G. & Yeager, S. G. *Diurnal to Decadal Global Forcing for Ocean and Sea-Ice Models: the Data Sets and Flux Climatologies*. (NCAR Technical Note NCAR/TN-460+STR, 2004).
17. Alory, G., Wijffels, S. & Meyers, G. Observed temperature trends in the Indian Ocean over 1960–1999 and associated mechanisms. *Geophys. Res. Lett.* **34,** L02606 (2007).
18. Rouault, M., Penven, P. & Pohl, B. Warming in the Agulhas Current system since the 1980s. *Geophys. Res. Lett.* **36,** L12602 (2009).
19. Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. & Schwarzkopf, F. V. The response of the Antarctic Circumpolar Current to recent climate change. *Nature Geosci.* **1,** 864–869 (2008).
20. Swart, S. *et al.* Transport and variability of the Antarctic Circumpolar Current south of Africa. *J. Geophys. Res.* **113,** C09014 (2008).
21. Blanke, B., Arhan, M., Madec, G. & Roche, S. Warm water paths in the equatorial Atlantic as diagnosed with a general circulation model. *J. Phys. Oceanogr.* **29,** 2753–2768 (1999).
22. Van Sebille, E., Biastoch, A., van Leeuwen, P. J. & de Ruijter, W. P. M. A weaker Agulhas Current leads to more Agulhas leakage. *Geophys. Res. Lett.* **36,** L03601 (2009).
23. Ridderinkhof, H. & De Ruijter, W. P. M. Moored current observations in the Mozambique Channel. *Deep Sea Res. II* **50,** 1933–1955 (2003).
24. Matano, R. P., Beier, E. J. & Strub, P. T. Large-scale forcing of the Agulhas variability: the seasonal cycle. *J. Phys. Oceanogr.* **32,** 1228–1241 (2002).
25. Ou, H. W. & de Ruijter, W. P. M. Separation of an inertial boundary current from a curved coastline. *J. Phys. Oceanogr.* **16,** 280–289 (1986).
26. Biastoch, A., Reason, C. J. C., Lutjeharms, J. R. E. & Boebel, O. The importance of flow in the Mozambique Channel to seasonality in the greater Agulhas Current system. *Geophys. Res. Lett.* **26,** 3321–3324 (1999).
27. Oke, P. & England, M. Oceanic response to changes in the latitude of the Southern Hemisphere subpolar westerly winds. *J. Clim.* **17,** 1040–1054 (2004).
28. Meehl, G. *et al. Global Climate Projections* 747–846 (Cambridge Univ. Press, 2007).
29. Debreu, L., Vouland, C. & Blayo, E. AGRIF: Adaptive grid refinement in Fortran. *Comput. Geosci.* **34,** 8–13 (2008).
30. Madec, G. *NEMO Ocean Engine*. Technical Report 27 (Note du Pôle de Modélisation, Institut Pierre Simon Laplace, 2006).

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

Acknowledgements The integrations of the experiments have been performed at the Höchstleistungsrechenzentrum Stuttgart (HLRS) and the Computing Centre at Kiel University. We thank the NEMO, AGRIF and Ariane System Teams for their technical support. The analysis was supported by the DFG projects BO 907/2-2 and SFB 754 (TP A2, http://www.sfb754.de). J.R.E.L. received support from the International Bureau of the BMBF (SUA 07/004) and from the South African National Research Foundation.

Author Contributions A.B. and C.W.B. conceived the experimental concept. A.B. implemented and conducted the experiments, and carried out the analysis. F.U.S. performed the observational analysis. All authors discussed the results and jointly wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.B. (abiastoch@ifm-geomar.de).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

For a realistic model simulation of the greater Agulhas Current system the horizontal resolution is essential. The nested model configurations used here build on the ocean/sea-ice numerical framework of the 'Nucleus for European Modelling of the Ocean' (NEMO, v2.3)[30] and 'Adaptive Grid Refinement in Fortran' (AGRIF)[29], developed in the framework of the DRAKKAR collaboration.

The base model uses the global ORCA05 configuration, a tripolar, quasi-isotropic grid with a nominal resolution of 1/2°; its cell size of 45–50 km in the Agulhas region is not resolving the mesoscale. In the vertical dimension 46 levels (with ten levels in the upper 100 m, and 250-m resolution at the deepest levels) are used, whereby the bottom cells are allowed to be partially filled. This improved representation of topographic slopes, in combination with a refined, energy- and enstrophy-conserving advection scheme, has led to marked improvements in the global circulation features[31]. Additional subgrid-scale mixing parameterizations include a representation of mixed layer dynamics by a 1.5-level turbulent kinetic energy closure, a bi-Laplacian viscosity, and an iso-neutral Laplacian scheme. For tracer advection a total variance dissipation scheme[32]—a second-order, two-step monotonic scheme with moderate numerical diffusion—is used.

The model is driven at the surface by a consistent data set, CORE[16], which is a combination of the NCEP/NCAR atmospheric reanalysis[33] and independent observations used to correct known biases, and to globally balance the heat and freshwater budgets. Turbulent fluxes are computed via bulk formulae, allowing some feedback of the ocean on the atmospheric fluxes[16,34]. Data are prescribed at six-hourly (wind speed, humidity and atmospheric temperature), daily (short- and long-wave radiation) and monthly (rain and snow) resolution, with inter-annual variability over the time range 1958–2004. To avoid an artificial model drift due to an excess of freshwater[34], the CORE precipitation field has been replaced north of 30° N by observational values[35]; in addition common practice[34] has been followed by damping sea surface salinity towards monthly-mean climatological values with a piston velocity of 50 m per 300 days (about a one-month timescale) poleward of 70° N and 50° S. Equatorward of these latitudes (and in the high-resolution nest) a very weak damping (more than a one-year timescale) was used, leaving the evolution within the Agulhas area almost unaffected. Several studies have demonstrated the fidelity of this ORCA05 set-up in simulating the salient features of the Atlantic MOC[36,37].

The experiments were initialized from rest using temperatures and salinities from a global climatology[38,39] and integrated over 20 years using the repeated-year version of the CORE forcing data. After that time all prognostic model fields were interpolated onto the 1/10° nest in the greater Agulhas region (20° W–70° E, 47°S–7° S, green box in Fig. 2). The fivefold refinement of the original ORCA05 grid, with an average grid cell of 9.5 km at 30° S, resolves the baroclinic Rossby radius of ∼30 km in this regime. Apart from some resolution-dependent scaling the same parameterizations have been used. Both base and nested model were then integrated over the full period of 1958–2004. The nesting approach acts in two ways so that not only is the regional model continually receiving information from the outside ocean, but also it feeds the effect of the mesoscale dynamics in the leakage regime back to the base model at all timescales. This feature enabled us to isolate the effect of the Agulhas dynamics onto the Atlantic MOC[7].

Previous studies demonstrated the success of the model set-up in reproducing all the important circulation features of the greater Agulhas Current system, including a realistic WBC structure off Africa[15], the upstream perturbations arising from the Mozambique Channel[11], and their interplay with the shedding of Agulhas rings[8]. In addition to the reference experiment (AG01-R) using the inter-annually varying CORE forcing fields, two sensitivity experiments have been performed: AG01-C, without inter-annual variability in the forcing fields, allows us to isolate the internal variability and a possible spurious model drift from the changes due to the forcing fields by comparison with AG01-R. In AG01-S the high-resolution nest terminates at 27° S, thus excluding the mesoscale eddies in the source regions of the Agulhas Current from the model solution[8].

Similar to a previous analysis[8] a Lagrangian method[21] was used, in which a large number (typically 10^5–10^6) of virtual floats was seeded continuously over a year into the Agulhas Current at 32° S. Each float represents a fraction (∼0.1 Sv) of the total volume transport, is advected by the time-dependent flow field and is counted when crossing the GoodHope section[20] (green line in Fig. 1a) in the Atlantic within three years of release. The algorithm uses all velocity components and analytically calculates a three-dimensional streamfunction for any given five-daily averages, thereby avoiding spurious diffusion. A second float integration elucidated the longer-term paths of the Agulhas leakage towards the North Atlantic by seeding fractional transports (∼0.01 Sv) crossing the GoodHope section with temperatures exceeding 10 °C and advecting those with the five-daily base model flow fields for two different pentads (1968–1972 and 2000–2004). About a third of the floats do not reach one of the control sections within five years (especially the ones looping in the subtropical gyre); so to draw out the path differences, the float integration was elongated by repeatedly cycling through these pentads until ∼95% of the floats had crossed one of the sections (Fig. 5). Comparison with float experiments over ten-year periods and different repetitions led to similar conclusions for the volume and freshwater transports arriving at 6° S (Supplementary Table 1).

The model fields were compared with gridded time series of sea surface height (http://www.aviso.oceanobs.com) based on satellite altimeter data, produced by Ssalto/Duacs and distributed by Aviso, with support from CNES. The mean dynamic topography Rio05 was produced by the CLS Space Oceanography Division.

To assess temporal changes in salinity fields in the North Brazil Current near 6° S, we used an extensive compilation of historical shipboard hydrographic cast and high-quality buoy data provided by the BLUElink Ocean Archive (BOA, ftp://ftp.marine.csiro.au/pub/omas/BOA). Our model-data comparison focused on the region 40° W–30° W, 5° S–10° S. The irregularly distributed BOA profiles constitute the basis for the high-resolution, gridded 'CSIRO Atlas of Regional Seas' (CARS, http://www.marine.csiro.au/∼dunn/cars2006)[40] climatology. Following previous studies[19] we minimized the potential spatial aliasing arising from the temporally varying distributions of sampling points by subtracting the climatological mean from the individual BOA profiles. The resulting salinity values were averaged over the individual pentads and the depth range of the North Brazil Current (100–600 m). Error bars were drawn using 95% confidence intervals, given by doubling the population standard deviation divided by the square root of the number of degrees of freedom.

31. Barnier, B. *et al.* Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy permitting resolution. *Ocean Dyn.* **56,** 543–567 (2006).
32. Zalesak, S. T. Fully multidimensional flux corrected transport algorithms for fluids. *J. Comput. Phys.* **31,** 335–362 (1979).
33. Kalnay, E. *et al.* The NCEP/NCAR 40-years reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
34. Griffies, S. *et al.* Coordinated ocean-ice reference experiments (COREs). *Ocean Model.* **26,** 1–46 (2009).
35. Béranger, K., Barnier, B., Gulev, S. & Crépon, M. Comparing 20 years of precipitation estimates from different sources over the world ocean. *Ocean Dyn.* **56,** 104–138 (2006).
36. Latif, M. *et al.* Is the thermohaline circulation changing? *J. Clim.* **19,** 4631–4637 (2006).
37. Biastoch, A., Böning, C. W., Getzlaff, J., Molines, J.-M. & Madec, G. Causes of interannual-decadal variability in the meridional overturning circulation of the mid-latitude North Atlantic Ocean. *J. Clim.* **21,** 6599–6615 (2008).
38. Conkright, M. *et al. World Ocean Database 2001* Vol. 1 *Introduction* 1–167 (NOAA Atlas NESDIS 42, US Government Printing Office 13, 2002).
39. Steele, M., Morfley, R. & Ermold, W. PHC: A global ocean hydrography with a high-quality Arctic Ocean. *J. Clim.* **14,** 2079–2087 (2001).
40. Ridgway, K., Dunn, J. & Wilkin, J. Ocean interpolation by four-dimensional weighted least squares: application to the waters around Australasia. *J. Atmos. Ocean. Technol.* **19,** 1357–1375 (2002).

©2009 Macmillan Publishers Limited. All rights reserved

Convective upwelling in the mantle beneath the Gulf of California

Yun Wang[1], Donald W. Forsyth[1] & Brian Savage[2]

In the past six million years, Baja California has rifted obliquely apart from North America, opening up the Gulf of California[1]. Between transform faults, seafloor spreading and rifting is well established in several basins. Other than hotspot-dominated Iceland, the Gulf of California is the only part of the world's seafloor-spreading system that has been surrounded by enough seismometers to provide horizontal resolution of upper-mantle structure at a scale of 100 kilometres over a distance great enough to include several spreading segments. Such resolution is needed to address the long-standing debate about the relative importance of dynamic and passive upwelling in the shallow mantle beneath spreading centres. Here we use Rayleigh-wave tomography to image the shear velocity in the upper 200 kilometres or so of the mantle. Low shear velocities similar to those beneath the East Pacific Rise oceanic spreading centre underlie the entire length of the Gulf, but there are three concentrated locations of anomalously low velocities spaced about 250 kilometres apart. These anomalies are 40 to 90 kilometres beneath the surface, at which depths petrological studies indicate that extensive melting of passively upwelling mantle should begin[2,3]. We interpret these seismic velocity anomalies as indicating that partial melting triggers dynamic upwelling driven by either the buoyancy of retained melt or by the reduced density of depleted mantle.

Upwelling in the mantle beneath spreading centres is thought to be largely a passive process in which mantle flow is driven by viscous coupling to the separating plates. The uniformity of composition of mid-ocean ridge basalts (MORBs) derived from partial melting of the upwelling mantle regardless of location of the spreading centre, even when a ridge is about to be subducted beneath a continental margin, and the lack of any deep, underlying structure below the asthenosphere[4–6], are strong indications that upwelling is not primarily driven dynamically by buoyant, hot mantle. After pressure-release melting begins in the shallow mantle, however, there can be a component of dynamic upwelling driven either by the retention of buoyant melt or by depletion of the mantle matrix following removal of melt to form the oceanic crust[7–10]. This dynamic component could concentrate the upwelling of the solid mantle into a relatively narrow sheet beneath a linear ridge or into a series of diapir-like upwelling centres. The importance of the dynamic component and the spacing between any upwelling centres depend critically on the poorly known viscosity structure of the oceanic mantle[7–10].

A seismic refraction experiment along the northern East Pacific Rise found a series of low-velocity anomalies in the uppermost few kilometres of the mantle[11]. The off-axis location of at least one of the anomalies and regular spacing suggests a dynamic component, but because only the very shallowest mantle is imaged, it is not clear whether it represents a process of upward melt migration or deeper upwelling of the solid mantle matrix. A long-term deployment of broadband seismometers around the Gulf of California[12] (Fig. 1) provides the opportunity to probe deeper into the mantle to the expected depths of solid upwelling and melt production. In our study, we used vertical-component, fundamental-mode, Rayleigh-wave seismograms from 93 earthquakes recorded at 25 broadband stations (Supplementary Fig. 1) to generate a three-dimensional tomographic image of shear velocity of the upper mantle to depths

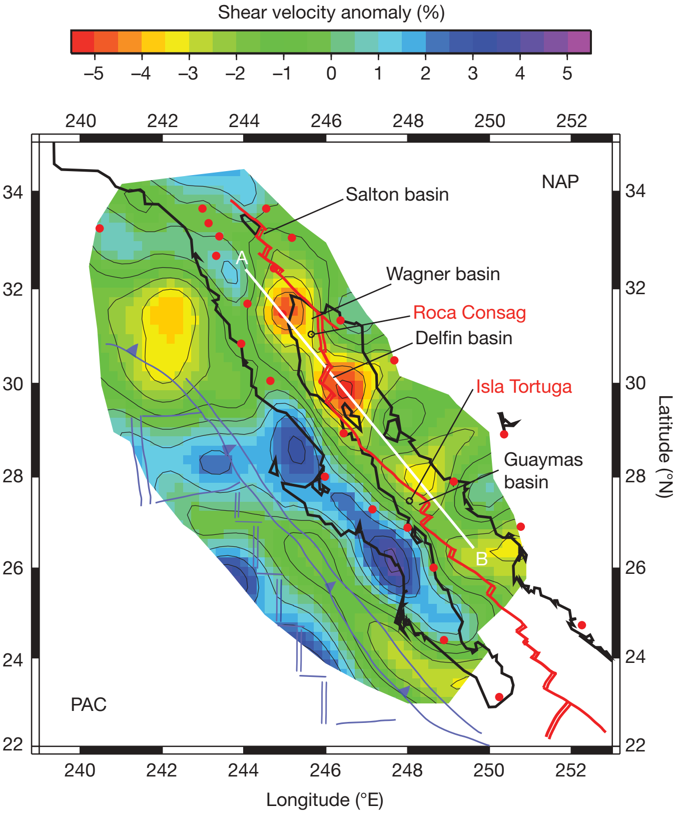


Figure 1 | Shear velocity anomalies averaged over a depth of 50 to 90 km beneath the Gulf of California and Baja California region. Negative anomalies correspond to slow regions. The contour interval is 0.5%. The coastline is indicated by the heavy black line. Red lines are the current plate boundary, with double lines indicating rifts or spreading centres and single lines indicating transform faults. Blue lines indicate the fossil spreading centre, trench and strike-slip fault west of Baja California. Red dots are broadband seismic stations employed in this study. White line AB is the location of the profile shown in Fig. 3. PAC and NAP stand for Pacific Plate and North America Plate, respectively. High velocities beneath southern Baja California are probably caused by a remnant piece of subducted slab still attached to the fossil Magdelena microplate seaward of Baja California[28].

[1]Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA. [2]Department of Geosciences, University of Rhode Island, Kingston, Rhode Island 02881, USA.

©2009 Macmillan Publishers Limited. All rights reserved

of about 210 km. The sources have very good azimuthal distribution, which is needed for good tomographic resolution and to avoid bias from azimuthal anisotropy. Clock corrections were needed for three stations whose timing was not properly synchronized (clock corrections are critical to the reliable resolution of structure). In an earlier study of this area using the two-station method, it was noted that the apparent velocity along reversed paths between two stations sometimes did not agree[13]; we interpret this discrepancy as being caused by clock errors and find corrections from primary, compressional (P)-wave delays and surface waves that agree well with each other (Supplementary Figs 2 and 3).

Shear velocity anomalies averaged over a depth range of 50–90 km are mapped in Fig. 1. This depth range is where extensive melting of passively upwelling asthenosphere is expected to begin[2,3]. The most pronounced velocity variations in this depth range are a series of low-velocity anomalies distributed along the Gulf of California, spaced about 250 km apart. These anomalies reach a maximum amplitude of 6% to 7% lower than the average velocities at about 60 km. Low velocities underlie nearly all of the active rift and transform plate boundary, but are lowest near the Wagner, Delfin and Guaymas basins. These three low-velocity centres are near, but do not lie directly beneath the major rifting centres in the basins, as might be expected for completely passive upwelling and melting. There is a recently formed island with MORB tholeiitic character[14], Isla Tortuga, lying at the edge of the Guaymas basin anomaly (Fig. 1). Another small Quaternary island[15], Roca Consag, lies between the Wagner anomaly and the rifting centre, but the Las Tinajas volcanic field directly overlying the Wagner anomaly on land represents typical subduction-related, calc-alkaline volcanic activity that ceased in the early Miocene epoch[16]. We suggest that most of the melt produced within these low-velocity centres may migrate horizontally back to spreading centres, perhaps along the sloping base of the thermal lithosphere.

The vertical resolution length for shear velocity structure at a target depth of 60 km is about 50 km, that is, the average shear velocity over the depth range 40–90 km can be resolved with useful uncertainty (there is always a trade-off possible between resolution length and model variance). Although perhaps not statistically significant, the Wagner and Guaymas anomalies are concentrated more in the upper part of this depth range whereas the Delfin anomaly may be strongest in the lower part. The Wagner and Guaymas anomalies do not extend deeper than about 100 km. We compare the average shear velocities beneath these centres to the averages beneath points midway between the centres in Fig. 2. The average shear wave velocity in the depth range of 40–90 km beneath the centres is about 4.01 km s^{-1}, equivalent to that found in the same depth range beneath the fast-spreading East Pacific Rise[4,6,17] or beneath volcanically active areas east of the Sierra Nevada in California[18]. The average velocity of the anomalous centres is less than the average beneath points in between the centres at the 95% confidence level, although the in-between average is also quite low ($\sim$4.12 km s^{-1}) and may still require the presence of melt.

The effect of melt on seismic velocities is still a controversial topic, because it depends on the poorly known melt fraction, the shapes of the melt pockets, and on whether the dominant attenuation mechanism is grain boundary relaxation[19] or melt squirt[20]. Shear (S)-wave velocities in the mantle lower than $\sim$4.3 km s^{-1} are difficult to achieve with the effects of temperature and composition alone unless the attenuation quality factor Q is 30 or less[21,22]. Such low values of Q are not found in the mantle beneath the southern East Pacific Rise[23] where S velocities are similarly low, suggesting that the velocities are affected by attenuation outside the seismic frequency band, most probably by the melt squirt mechanism[24]. Although the number and distribution of stations are not sufficient to resolve Q adequately within the Gulf, the similarity to the East Pacific Rise prompts us to suggest that the low velocities correspond to a small melt fraction and that the three low-velocity centres indicate higher melt fractions than in the surrounding areas.

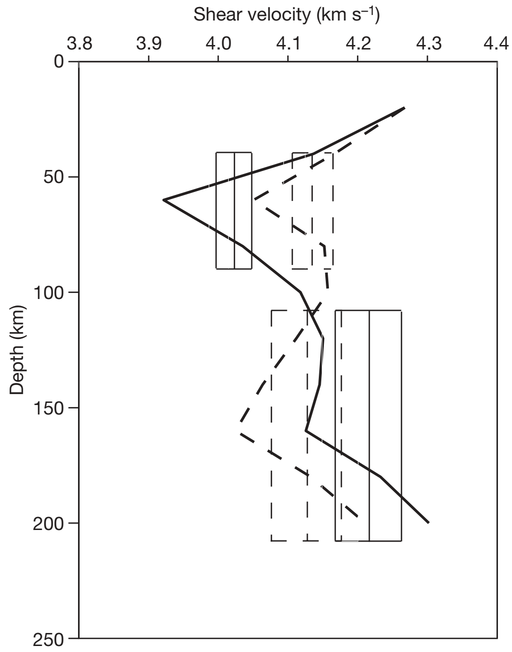


Figure 2 | Average vertical shear wave velocity profiles. The average vertical shear wave velocity profiles through the centres of low velocity at 60 km (solid line) and at points along profile AB (Fig. 1) midway between the centres (dashed line) are compared. Boxes represent average velocities over resolvable depth ranges plus or minus one standard deviation. The average velocities are significantly different in the depth range 40–90 km at the 95% confidence level, but differences in the depth range 110–210 km are not as significant.

The average velocity from 110 km deep to 210 km deep, which is less than 4.2 km s^{-1}, is also quite low beneath the Gulf of California (Fig. 2), suggesting perhaps incipient melting in the presence of a small amount of dissolved water continuing significantly deeper than 110 km. Although only on the margins of statistical significance, the velocities beneath the 60-km-deep, low-velocity centres in the depth range of 110–210 km are actually faster than beneath the points in between the centres in this depth range. See Methods section for a discussion of the reliability of this observation.

In summary, we attribute these low-velocity anomalies to dynamic, buoyancy-driven upwelling and melting (Fig. 3) initially triggered by extension that began in the Gulf region about six million years ago. The role of melting is suggested by the very low shear velocities and the fact that the maximum anomaly is at a depth of about 60 km, at which petrological models indicate mantle melting should be at the greatest rate (melt volume per kilometre uplift)[3]. The spacing of the anomalies about 250 km apart and the offset of the centres of the anomalies from the current, nearby spreading centres suggest that shallow, melt-depletion buoyancy and melt-retention buoyancy may have organized the initially passively driven upwelling into regularly spaced cells. The 250-km spacing is consistent with the characteristic dynamic segmentation length predicted in numerical models of buoyant mantle instabilities triggered by rifting and melting[9]. The flow pattern is undoubtedly three-dimensional; downflow to balance the buoyancy-driven upwelling could be accommodated in a direction perpendicular to the strike of the Gulf or could reduce the rate of upwelling between the centres, as we have depicted in Fig. 3.

Although dynamic upwelling is induced primarily by the depletion or melt retention near the top of the melting column, upwelling is expected to extend deep into the asthenosphere[10]. It is primarily the viscosity structure of this underlying asthenosphere that controls the spacing between upwelling centres[9]. The low average seismic velocities from 110 km deep to 210 km deep suggests that there may be low viscosities and incipient melting in this range; in Fig. 3 we illustrate the bottom of the incipient melting region arbitrarily in the middle of

©2009 Macmillan Publishers Limited. All rights reserved

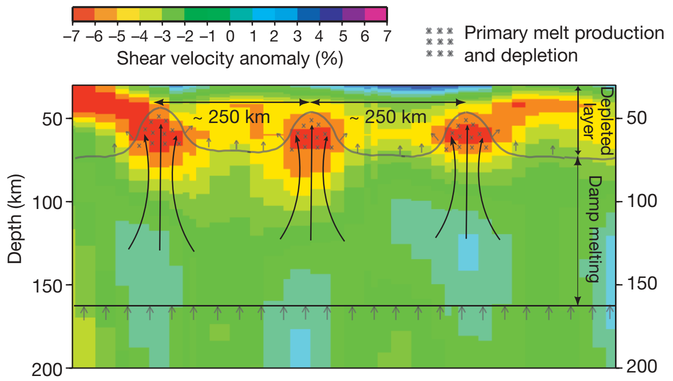


Figure 3 | Schematic interpretation of anomalous mantle velocities along profile AB. The most prominent anomalies are the low-velocity anomalies centred at depths of 60–70 km (from left to right, Wagner, Delfin and Guaymas basins), which we interpret as centres of enhanced melt concentration and upwelling. Arrows indicate directions of solid flow of the mantle matrix. Melting is assumed to begin at ~160 km in the presence of a small amount of water, leading to low S velocities (Fig. 2). Upwelling at this depth may be broadly and relatively uniformly distributed. At 60–70 km, the rate of melting is expected to increase substantially, but most of the melt is extracted, leaving a matrix that is depleted above this depth, but still containing a small melt fraction. The high-velocity anomaly near 30 km depth between the Delfin and Guaymas anomalies is due to cooling and thickening of the lithosphere with increasing age of the sea floor (see location of profile AB with respect to the spreading centres in Fig. 1).

this resolvable depth range, that is, at 160 km. Upwelling will only be obvious in terms of seismic velocity anomalies where there is a large lateral temperature contrast or when there is an effect of retained melt on the shear velocity. In the case of an overall passive system in which local, dynamic upwelling is induced by partial melting, no major anomalies are expected in the underlying asthenosphere except possibly through anisotropic effects associated with realignment of crystal fabric. A potential indication of such a realignment is the presence of higher velocities in the depth range 110–210 km than beneath the adjacent areas, which is consistent with the more vertical alignment of olivine a-axes (the fastest direction of seismic wave propagation in the mineral) expected within upwelling regions.

Two possible alternative interpretations should be mentioned. First, it is possible that the low-velocity centres are artefacts of the inversion. Resolution tests indicate that variations in station and path coverage would not break up a more uniform anomaly into separate anomalies (Supplementary Fig. 7), but coherent noise due to complexities in the incoming wave fields that are not adequately represented by our two-plane-wave approximation could be responsible. With the many sources we use from a wide variety of azimuths, we think this explanation is unlikely. Second, it is possible that the low-velocity centres do indeed correspond to higher melt concentrations, but they do not represent centres of upwelling. Instead, upwelling and melting could occur in a broad region around each centre of extension, but melt extraction beneath spreading centres could be more efficient than beneath transform faults or off-axis, leaving trapped pockets of melt that do not coincide with the regions of maximum melt production.

METHODS SUMMARY

We use a modified version of the two-plane wave method for Rayleigh-wave tomography[25,26] using finite-frequency kernels[27] to find three-dimensional shear velocity structure. Modifications are described in detail in the Methods.

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

Received 15 April; accepted 28 September 2009.

1. Stock, J. M. & Hodges, J. M. Pre-Pliocene extension around the Gulf of California and the transfer of Baja California to the Pacific Plate. *Tectonics* **8,** 99–115 (1989).
2. Langmuir, C. H., Klein, E. M. & Plank, T. in *Mantle Flow and Melt Generation at Mid-Ocean Ridges* (eds Morgan, P. J., Blackman, D. K., & Sinton, J. M.) 183–280 (Geophysical Monograph Series 71, American Geophysical Union, 1992).
3. Asimow, P. D., Hirschmann, M. M. & Stolper, E. M. Calculation of peridotite partial melting from thermodynamic models of minerals and melts. IV. Adiabatic decompression and the composition and mean properties of mid-ocean ridge basalts. *J. Petrol.* **42,** 963–998 (2001).
4. Nishimura, C. & Forsyth, D. W. The anisotropic structure of the upper mantle in the Pacific. *Geophys. J.* **96,** 203–229 (1989).
5. Hammond, W. C. & Toomey, D. R. Seismic velocity anisotropy and heterogeneity beneath the Mantle Electromagnetic and Tomography experiment (MELT) region of the East Pacific Rise from analysis of P and S body waves. *J. Geophys. Res.* **108,** 2176, doi:10.1029/2002JB001789 (2003).
6. Ritzwoller, M., Shapiro, N. & Zhong, S.-J. Cooling history of the Pacific lithosphere. *Earth Planet. Sci. Lett.* **226,** 69–84 (2004).
7. Buck, W. R. & Su, W. S. Focused mantle upwelling below mid-ocean ridges due to feedback between viscosity and melting. *Geophys. Res. Lett.* **16,** 641–644 (1989).
8. Parmentier, E. M. & Morgan, P. J. The spreading rate dependence of three-dimensional oceanic spreading center structure. *Nature* **348,** 325–328 (1990).
9. Magde, L. S. & Sparks, D. W. Three-dimensional mantle upwelling, melt generation, and melt migration beneath segmented slow spreading ridges. *J. Geophys. Res.* **106,** 20571–20583 (1997).
10. Choblet, G. & Parmentier, E. M. Mantle upwelling and melting beneath slow spreading centers: effects of variable rheology and melt productivity. *Earth Planet. Sci. Lett.* **184,** 589–604 (2001).
11. Toomey, D. R. *et al.* Skew of mantle upwelling beneath the East Pacific Rise governs segmentation. *Nature* **446,** 409–414 (2007).
12. Clayton, R. W. *et al.* The NARS-Baja seismic array in the Gulf of California rift zone. *MARGINS Newsl.* **13,** 1–4 (2004).
13. Zhang, X. *et al.* Surface wave tomography of the Gulf of California. *Geophys. Res. Lett.* **34,** L15305, doi:10.1029/2007GL030631 (2007).
14. Batiza, R. Geology, petrology and geochemistry of Isla Tortuga, a recently formed tholeiitic island in the Gulf of California. *Geol. Soc. Am. Bull.* **89,** 1309–1324 (1978).
15. Martin, A. *et al.* Recent volcanism in the northern Gulf of California and the Salton trough: why a preponderance of evolved magmas? *AGU Fall Meet.* abstr. T11A–1841 (2008).
16. Axen, G. Extensional segmentation of the main gulf escarpment, Mexico and United States. *Geology* **23,** 515–518 (1995).
17. Harmon, N., Forsyth, D. W. & Weeraratne, D. S. Thickening of young Pacific lithosphere from high resolution Rayleigh wave tomography: a test of the conductive cooling model. *Earth Planet. Sci. Lett.* **278,** 96–106, doi:10.1016/j.epsl.2008.11.025 (2008).
18. Yang, Y. & Forsyth, D. W. Rayleigh wave phase velocities, small-scale convection, and azimuthal anisotropy beneath southern California. *J. Geophys. Res.* **111,** B07306, doi:10.1029/2005JB004180 (2006).
19. Faul, U. H., FitzGerald, J. D. & Jackson, I. Shear-wave attenuation and dispersion in melt-bearing olivine polycrystals. II. Microstructural interpretation and seismological implications. *J. Geophys. Res.* **109,** doi:10.1029/2003B002407 (2004).
20. Hammond, W. C. & Humphreys, E. D. Upper mantle seismic wave velocity: effects of realistic partial melt geometries. *J. Geophys. Res.* **105,** 10975–10986 (2000).
21. Karato, S. & Jung, H. Water, partial melting and the origin of the seismic low velocity and high attenuation zone in the upper mantle. *Earth Planet. Sci. Lett.* **157,** 193–207 (1998).
22. Stixrude, L. & Lithgow-Bertelloni, C. Mineralogy and elasticity of the oceanic upper mantle: origin of the low-velocity zone. *J. Geophys. Res.* **110,** B03204, doi:10.1029/2004JB002965 (2005).
23. Yang, Y., Forsyth, D. W. & Weeraratne, D. S. Seismic attenuation near the East Pacific Rise and the origin of the low-velocity zone. *Earth Planet. Sci. Lett.* **258,** 260–268 (2007).
24. Hammond, W. C. & Humphreys, E. D. Upper mantle seismic wave attenuation: effects of realistic partial melt distribution. *J. Geophys. Res.* **105,** 10987–10999 (2000).
25. Forsyth, D. W. & Li, A. in *Seismic Earth: Array Analysis of Broadband Seismograms* (eds Levander, A. & Nolet, G.) 81–97 (Geophysical Monograph Series 157, American Geophysical Union, 2005).
26. Yang, Y. & Forsyth, D. W. Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophys. J. Int.* **166,** 1148–1160, doi:10.1111/j.1365-246X.2006.02972.x (2006).
27. Zhou, Y., Dahlen, F. A. & Nolet, G. 3-D sensitivity kernels for surface-wave observables. *Geophys. J. Int.* **158,** 142–168 (2004).
28. Zhang, X. *et al.* 3D shear velocity structure beneath the Gulf of California from Rayleigh wave dispersion. *Earth Planet. Sci. Lett.* **279,** 255–262 (2009).

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

Acknowledgements This research was supported by the MARGINS programme of the National Science Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.W. (yun_wang@brown.edu).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

To perform the tomographic inversion for three-dimensional distribution of shear velocities in the mantle, we first solve for the lateral variations in phase velocities for periods ranging from 22 to 143 s, then invert those phase velocity maps as a function of period for vertical variations in shear velocity. To find the lateral variations in phase velocity, we use a modified version of the two-plane wave method[14,15]. The effects of scattering due to lateral heterogeneities in the study area are accounted for by using finite-frequency response kernels for both amplitude and phase[16]. To account for distortion of the incoming wave field by heterogeneities outside the array, we approximate the field for each earthquake as the sum of two incoming plane waves of unknown amplitude, phase and propagation direction, amounting to six wave field parameters for each event. Because the area of study is so elongated, we subdivide the region into four sub-regions (Supplementary Fig. 4), with different wave field parameters in each sub-region.

Our emphasis is on mantle structure, but the variations in sediment thickness, water depth, and oceanic versus continental crust have a significant effect, particularly at shorter periods. We construct a three-dimensional regional model of the crust based on topography and previous seismological studies of crustal structure (Supplementary Fig. 6), then we predict phase velocities at each period throughout the region using this regional crustal model, assuming that mantle structure everywhere is uniform and the same as the average mantle structure found beneath southern California[17]. These predicted velocities serve as the starting model for the inversion at each period. We use a two-dimensional Gaussian smoothing function in the inversion with characteristic length of 80 km, which means in practice that features 80–100 km across can be resolved. In the subsequent inversion for shear velocity structure, we also use the regional crustal model as the starting model, but crustal velocity is allowed to change if required by the data. This procedure is designed to minimize the mapping of crustal features erroneously into the mantle.

The area of reliable velocities shown in Fig. 1 is based on the covariance and resolution analysis of the inversions and represents a subset of the total model region (Supplementary Fig. 4). We note that the existence of crossing rays and the width of the finite-frequency sensitivity kernels extends the resolved region outside the limits of the array of stations, particularly to the west (Supplementary Fig. 5). We include azimuthal anisotropy in the inversion, dividing the region into three areas for that purpose, but resolution of azimuthal anisotropy is poor owing to limited station distribution. Inclusion of anisotropic terms made no significant difference in the lateral velocity anomalies.

The inversion for phase velocity coefficients at the nodal points is underdetermined. We provide regularization by using a combination of minimum length and minimum curvature criteria for the least-squares solution, minimizing differences from the starting model. We accomplish this by treating the starting phase velocity coefficients as a priori information with a standard error of $\sim 0.25\,\mathrm{km\,s^{-1}}$ at most periods and some off-diagonal terms in the model covariance matrix, effectively damping the solution.

In this type of tomographic study, inversions of phase velocities for vertical velocity structure often have a tendency for velocity anomalies to reverse in sign with depth, for example, a deep, fast anomaly underlying a shallower, slow anomaly. Usually the deeper anomaly is not statistically significant (unless it represents some real structure). The primary causes of this problem are limited vertical resolving power and unequal resolution of phase velocity anomalies at different periods. The longest periods typically have fewer good-quality observations, broader sensitivity kernels, and larger errors in relative time of arrival (the same fractional phase accuracy produces larger time differences at longer periods), leading to lower resolution. If the same damping factor is applied, the amplitude of phase velocity anomalies at longer periods will be underestimated relative to shorter periods because the starting model will have increased relative weight and the anomalies will be effectively averaged over a larger region. We took two steps to ameliorate this problem. First, we relaxed the damping at longer periods to make the lateral resolution more uniform. This step has two effects: it increases the amplitude of the anomalies and increases the variance, which decreases the weight given to the longer period data in the inversion. Second, we changed the regularization in the point-by-point inversion of phase velocities for the shear velocity structure from a minimum length criterion to a criterion that penalizes correlated changes in layers remote from the target layer. To do this, we added off-diagonal terms in the a priori model covariance matrix. We determined these off-diagonal terms from the resolution matrix for a minimum length criterion. A row of the resolution matrix shows how the information for a target layer is intertwined with information about other layers. If there were terms remote from the target layer that were in the form of oscillations, we introduced them into a priori covariance matrix.

Despite these modifications that reduced the overall negative correlation of deep structure with shallower structure, the reversal in relative velocity illustrated in Fig. 2 remains along the spreading Gulf of California. Although not statistically significant at the 95% confidence level, and so perhaps not worthy of attention, we do not believe that this reversal is an artefact of the method, because in the period range of 90–110 s, the phase velocities are faster at the same spots as the shallow, low-velocity centres than they are elsewhere along the AB profile (Supplementary Fig. 5), thus forcing a velocity reversal in the depth range of 110–210 km.

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Aero-tactile integration in speech perception

Bryan Gick[1,2] & Donald Derrick[1]

Visual information from a speaker's face can enhance[1] or interfere with[2] accurate auditory perception. This integration of information across auditory and visual streams has been observed in functional imaging studies[3,4], and has typically been attributed to the frequency and robustness with which perceivers jointly encounter event-specific information from these two modalities[5]. Adding the tactile modality has long been considered a crucial next step in understanding multisensory integration. However, previous studies have found an influence of tactile input on speech perception only under limited circumstances, either where perceivers were aware of the task[6,7] or where they had received training to establish a cross-modal mapping[8–10]. Here we show that perceivers integrate naturalistic tactile information during auditory speech perception without previous training. Drawing on the observation that some speech sounds produce tiny bursts of aspiration (such as English 'p')[11], we applied slight, inaudible air puffs on participants' skin at one of two locations: the right hand or the neck. Syllables heard simultaneously with cutaneous air puffs were more likely to be heard as aspirated (for example, causing participants to mishear 'b' as 'p'). These results demonstrate that perceivers integrate event-relevant tactile information in auditory perception in much the same way as they do visual information.

Many languages use an expulsion of air, or 'aspiration', to convey basic lexical contrasts[12]. English speakers use this mechanism to distinguish aspirated sounds such as 'pa' and 'ta' from unaspirated sounds such as 'ba' and 'da'. All four human dermal mechanoreceptors[13], as well as hair-follicle mechanoreceptors[14], respond to air puffs. Aerodynamically, a puff is characterized as a short burst of turbulent airflow with a relatively higher initial pressure[15,16], typical of the transient pressure pattern produced in aspirated speech sounds[17].

We created auditory stimuli by recording eight repetitions of each of the syllables 'pa', 'ba', 'ta' and 'da' from a male native speaker of English, matching for duration (390–450 ms each), fundamental frequency (falling pitch from 90 Hz to 70 Hz) and intensity (normalized to 70 decibels ($10^{-5}\,\mathrm{W\,m^{-2}}$). Participants heard syllables in two separate blocks: one containing only labial consonants ('pa' and 'ba'), the other containing only alveolar consonants ('ta' and 'da'). The 16 unique tokens in each block were heard four times each—twice as auditory-only controls and twice paired with tactile stimuli. Auditory stimuli were accompanied by white noise played at a volume intended to reduce the overall accuracy of token identification and so generate significant ambiguity; actual accuracy is documented in Supplementary Tables 1–3.

We used a solenoid valve attached to an air compressor to synthesize small puffs of air designed to replicate the pressure profile (transient boundary condition), high frequency noise, low frequency 'pop' duration and temporal relation to vowel onset of natural speech aspiration.

In our first experiment, air puffs were applied cutaneously on the dorsal surface of the hand between the right thumb and forefinger through ¼-inch (0.635-cm) vinyl tubing at 6 pounds per square inch (p.s.i.; 6 p.s.i. $\approx 421.84\,\mathrm{g\,cm^{-2}}$) fixed at 8 cm from the skin surface. The back of the hand was chosen because it has high tactile sensitivity[18], and because it is a location where tactile stimulation including airflow has been observed to elicit non-specific activation of some second-stage auditory cortical neurons in macaques[19].

We considered that participants may have a good deal of previous experience with air puffs on the hand coupled with speech sounds—from concurrently hearing their own voice and feeling their own breath on their hands during speech. To determine whether the interaction would persist even at a body location lacking frequent self-experience, we designed a second experiment in which we applied air puffs to the centre of the neck at the suprasternal notch—a location where participants typically receive no direct airflow during their own speech production (though perceivers presumably do, at least on rare occasion, feel interlocutors' aspirated air on their skin). As with the hand experiment, air puffs were delivered through ¼-inch vinyl tubing at 6 p.s.i. fixed at 8 cm from the skin surface.

In addition to the hand and neck trials, an 'auditory-only' experiment was designed to ensure that delivery of the air puffs was inaudible to participants. In this trial, the ¼-inch tube was placed immediately beside the participants' right headphone at a distance of 5 cm and a pressure of 6 p.s.i., aimed tangentially forward so that airflow was not felt directly on the skin or hair.

A single stereo audio signal supplied both the auditory stimuli heard by participants and the activation signal to open the air valve. The right channel carried the spoken syllables to both ears through headphones worn by participants, while the left channel activated the solenoid by outputting 50-ms 10-kHz sine waves at the maximum amplitude of the computer's sound card ($\sim$1 V) through a voltage amplifier to a relay. The sine waves were time-aligned with the speech signal such that, after correction for system latency, air puffs exited the tube starting 50 ms before vowel onset and ending at the moment of vowel onset, thus simulating the timing of naturally produced English aspirated consonants.

Male and female participants were tested in all experiments. Before the experiment, participants were told that they might experience background noise and unexpected puffs of air. Participants were seated in a soundproof booth and asked to identify by pressing a button whether they heard 'pa' or 'ba' in the labial block, and 'ta' or 'da' in the alveolar block. Participants were then blindfolded and provided with auditory stimuli through sound-isolating headphones. The setup of equipment to deliver tactile stimuli was completed after the participants were blindfolded to conceal the body location of air puffs.

A mixed design repeated-measures analysis of variance was conducted with two consonant aspiration conditions (aspirated and unaspirated) by two airflow conditions (presence and absence) by two places of articulation (labial and alveolar) by three experiments (hand, neck and auditory-only). Results indicated weak main effects of aspiration ($F(1,63) = 5.426$, $P = 0.023$) (that is, perceivers identified unaspirated stops slightly more readily across all experiments) and place ($F(1,63) = 6.714$, $P = 0.012$) (that is, perceivers were slightly

[1] Department of Linguistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [2] Haskins Laboratories, New Haven, Connecticut 06511-6695, USA.

©2009 Macmillan Publishers Limited. All rights reserved

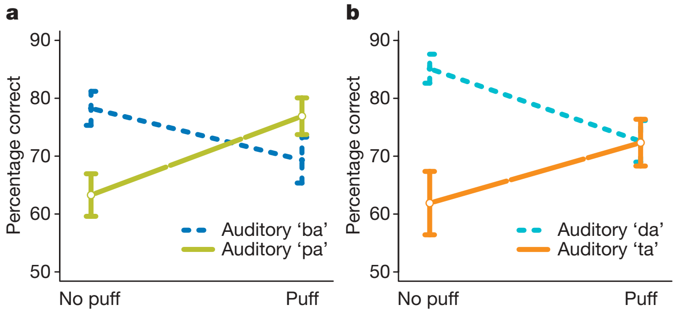


Figure 1 | Interaction graphs for the hand experiment with standard error bars. a, Labial; **b**, alveolar.

more accurate discerning alveolar versus labial stops), and strong main effects of aspiration × airflow ($F(1,63) = 26.095$, $P < 0.001$) (airflow caused perception of both unaspirated and aspirated stops as aspirated more often) and aspiration × airflow × experiment ($F(2,63) = 7.600$, $P = 0.001$) (that is, the effect of airflow applied to the neck and hand experiments, but not to the auditory-only experiment). There was no significant main effect of airflow, or of interaction between airflow and experiment (that is, application of airflow does not affect overall accuracy of perception of stimuli). No other significant effects were observed.

To identify whether there were significant interactions between aspiration and airflow in the hand and neck experiments, but not the auditory-only experiment, separate analyses of variance with repeated measures factors of aspiration (aspirated versus unaspirated) and air puffs (present versus absent) were conducted for both the alveolar and labial blocks of all experiments. Furthermore, to determine whether these interactions demonstrated augmentation of aspirated stop perception as well as interference with unaspirated stop perception, one-way repeated-measures analyses of variance comparing air puffs (present versus absent) were run separately for aspirated and unaspirated tokens.

Results for the hand experiment showed that the interaction of air puffs with the perception of aspiration was significant ($\alpha = 0.05$) for both the alveolar ($F(1,21) = 17.888$, $P < 0.001$, partial $\eta^2 = 46.0\%$) and labial ($F(1,21) = 14.785$, $P < 0.001$, partial $\eta^2 = 41.3\%$) blocks (Fig. 1). Further, the presence of an air puff enhanced correct identification of aspirated tokens ('pa' ($F(1,21) = 14.309$, $P = 0.001$, partial $\eta^2 = 40.5\%$) and 'ta' ($F(1,21) = 8.650$, $P = 0.008$, partial $\eta^2 = 29.2\%$)), and interfered with correct identification of unaspirated tokens ('ba' ($F(1,21) = 5.597$, $P = 0.028$, partial $\eta^2 = 21.0\%$) and 'da' ($F(1,21) = 16.979$, $P < 0.001$, partial $\eta^2 = 44.7\%$)).

Results for the neck experiment showed that the interaction of air puffs with the perception of aspiration was significant for both the alveolar ($F(1,21) = 5.486$, $P = 0.029$, partial $\eta^2 = 20.7\%$) and labial ($F(1,21) = 8.404$, $P = 0.009$, partial $\eta^2 = 28.6\%$) blocks (Fig. 2). Further, the presence of an air puff enhanced correct identification of aspirated tokens ('pa' ($F(1,21) = 7.140$, $P = 0.014$, partial $\eta^2 = 25.4\%$) and 'ta' ($F(1,21) = 6.020$, $P = 0.023$, partial $\eta^2 = 22.3\%$)) and showed a weak effect of interference with correct identification of unaspirated tokens ('ba' ($F(1,21) = 3.421$, $P = 0.078$, partial $\eta^2 = 14.0\%$) and 'da' ($F(1,21) = 1.291$, $P = 0.269$, partial $\eta^2 = 5.8\%$)).

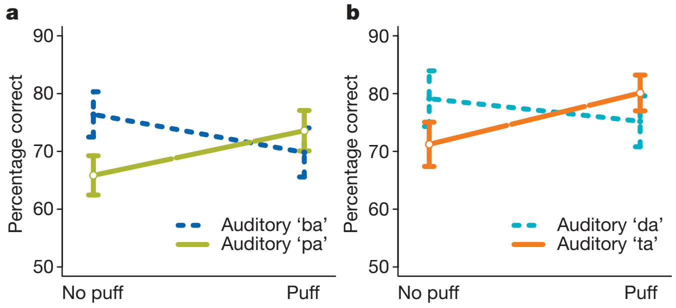


Figure 2 | Interaction graphs for the neck experiment with standard error bars. a, Labial; **b**, alveolar.

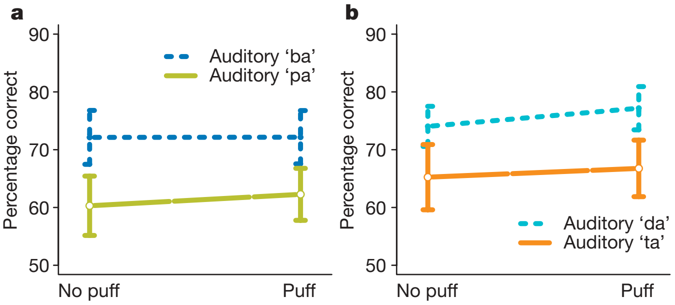


Figure 3 | Interaction graph for control experiment with standard error bars. a, Labial; **b**, alveolar.

No significant interaction between aspiration and air puffs was found for the auditory-only experiment (alveolar or labial block, $F(1,21) < 1$), confirming that participants could not hear the airflow or compressor activation (Fig. 3).

Our findings support the hypothesis that the human perceptual system integrates specific, event-relevant information across auditory and tactile modalities in much the same way as has been previously observed in auditory-visual coupling. This effect occurs in perceivers without previous training or awareness of the task, and at body locations where the effect is unlikely to be reinforced by frequent experience. These results complement recent work showing the involvement of the somatosensory system in speech perception[20], suggesting that the neural processing of speech is more broadly multimodal than previously believed. The methods used in this paper represent a model that will enable future functional imaging studies of passive audio-tactile and visuo-tactile integration, as well as behavioural studies of multi-sensory perception in previously untested populations, including infants and the blind. As these findings describe perceptual enhancement during passive perception, they imply possible future directions in audio and telecommunication applications and aids for the hearing impaired.

METHODS SUMMARY

Synthetic air puffs. The airflow device consisted of a 3-gallon (11.35-l) Jobmate oil-less air compressor connected to an IQ Valves on–off two-way solenoid valve (model W2-NC-L8PN-S078-MB-W6.0-V110) connected to a Campbell Hausfeld MP513810 air filter, which reduced the sound volume conducted through the ¼-inch vinyl tubing. The tubing was passed through a cable port into the soundproof room and mounted on a microphone boom-stand. The synthetic puff airflow was quickly turbulent upon leaving the tube, with an average turbulence duration of 84 ms, compared with 60 ms voice onset time for our speaker's average (mean) 'pa', and close to the range of voice onset time of 54–80 ms for English word-onset voiceless (aspirated) stops[12]. The output pressure of the synthesized puffs was adjusted so that impact was minimally perceptible by participants. As such, microphone recordings at 8 cm showed an average peak relative non-dimensional pressure of 0.023 for the synthetic puffs, compared with 0.096 for our speaker's average 'pa'.

Procedure. In total, we tested 66 participants, 22 for each of the experimental trials (hand and neck) and the auditory-only trial. Half received the labial ('pa', 'ba') block first, and half received the alveolar ('ta', 'da') block first. Within each block, participants heard 12 practice tokens (six with and six without air puffs) followed by 16 experimental tokens for each condition (aspirated versus unaspirated, puff versus no puff, randomized), totalling 64 experimental tokens per block. A custom-built computer program written in Java 1.6 recorded responses from a customized keypad and presented new tokens 1,500 ms after each response. Half of the participants pressed the left button to indicate an aspirated response, and half pressed the right button.

Received 1 September; accepted 9 October 2009.

1. Sumby, W. H. & Pollack, I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).

©2009 Macmillan Publishers Limited. All rights reserved

2. McGurk, H. & MacDonald, J. W. Hearing lips and seeing voices. *Nature* **264,** 746–748 (1976).
3. Calvert, G. A. *et al.* Activation of auditory cortex during silent lipreading. *Science* **276,** 593–596 (1997).
4. Calvert, G. A. & Campbell, R. Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* **15,** 57–70 (2003).
5. Diehl, R. L. & Kluender, K. R. On the objects of speech perception. *Ecol. Psychol.* **1,** 121–144 (1989).
6. Fowler, C. & Dekle, D. Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17,** 816–828 (1991).
7. Gick, B., Jóhannsdóttir, K. M., Gibraiel, D. & Mühlbauer, J. Tactile enhancement of auditory and visual speech perception in untrained perceivers. *J. Acoust. Soc. Am.* **123,** EL72–EL76 (2008).
8. Sparks, D. W., Kuhl, P. K., Edmonds, A. E. & Gray, G. P. Investigating the MESA (multipoint electrotactile speech aid): the transmission of segmental features of speech. *J. Acoust. Soc. Am.* **63,** 246–257 (1978).
9. Reed, C. M., Durlach, N. I., Braida, L. D. & Schultz, M. C. Analytic study of the Tadoma method: effects of hand position on segmental speech perception. *J. Speech Hear. Res.* **32,** 921–929 (1989).
10. Bernstein, L. E., Demorest, M. E., Coulter, D. C. & O' Connell, M. P. Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* **90,** 2971–2984 (1991).
11. Derrick, D., Anderson, P., Gick, B. & Green, S. Characteristics of air puffs produced in English 'pa': data and simulation. *J. Acoust. Soc. Am.* **125,** 2272–2281 (2009).
12. Lisker, L. & Abramson, A. S. A cross-language study of voicing in initial stops: acoustical measurements. *Word* **20,** 384–423 (1964).
13. Mizobuchi, K. *et al.* Single unit responses of human cutaneous mechanoreceptors to air-puff stimulation. *Clin. Neurophysiol.* **111,** 1577–1581 (2000).
14. Sabah, N. H. Controlled stimulation for hair follicle receptors. *J. Appl. Physiol.* **36,** 256–257 (1974).
15. Sangras, R., Kwon, O. C. & Faeth, G. M. Self-preserving properties of unsteady round nonbuoyant turbulent starting jets and puffs in still fluids. *J. Heat Transfer* **124,** 460–469 (2002).
16. Diez, F. J., Sangras, R., Kwon, O. C. & Faeth, G. M. Erratum: ''Self-preserving properties of unsteady round nonbuoyant turbulent starting jets and puffs in still fluids. *ASME J. Heat Transfer*, **124**, 460–469 (2002)''. *J. Heat Transfer* **125,** 204–205 (2003).
17. Stevens, K. N. *Acoustic Phonetics* Ch. 7 (MIT Press, 1998).
18. Weinstein, S. in *The Skin Sense* (ed. Kenshalo, D. R.) 195–222 (Thomas, 1968).
19. Fu, K.-M. G. *et al.* Auditory cortical neurons respond to somatosensory stimulation. *J. Neurosci.* **23,** 7510–7515 (2003).
20. Ito, T., Tiede, M. & Ostry, D. J. Somatosensory function in speech perception. *Proc. Natl Acad. Sci. USA* **106,** 1245–1248 (2009).

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

Acknowledgements This research was funded by a Discovery Grant from the Natural Sciences and Engineering Council of Canada to B.G., and by a National Institutes of Health Grant DC-02717 to Haskins Laboratories. We acknowledge the contributions of Y. Ikegami for initial piloting, R. Lee for piloting tapping protocols, L. McLeod for help setting up the University of British Columbia system, G. Ramsay and J. Wang for building switch mechanisms, S. Renardy for organizing protocols, D. H. Whalen for discussion and students of the University of British Columbia's Interdisciplinary Speech Research Laboratory for organizing participants and other assistance.

Author Contributions B.G. conceived and designed the experiment; D.D. designed and performed the data analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.G. (gick@interchange.ubc.ca).

©2009 Macmillan Publishers Limited. All rights reserved

Central control of fever and female body temperature by RANKL/RANK

Reiko Hanada[1], Andreas Leibbrandt[1], Toshikatsu Hanada[1], Shiho Kitaoka[2], Tomoyuki Furuyashiki[2], Hiroaki Fujihara[3], Jean Trichereau[1], Magdalena Paolino[1], Fatimunnisa Qadri[4], Ralph Plehm[4], Steffen Klaere[5], Vukoslav Komnenovic[1], Hiromitsu Mimata[6], Hironobu Yoshimatsu[6], Naoyuki Takahashi[7], Arndt von Haeseler[5], Michael Bader[4], Sara Sebnem Kilic[8], Yoichi Ueta[3], Christian Pifl[9], Shuh Narumiya[2] & Josef M. Penninger[1]

Receptor-activator of NF-κB ligand (TNFSF11, also known as RANKL, OPGL, TRANCE and ODF) and its tumour necrosis factor (TNF)-family receptor RANK are essential regulators of bone remodelling, lymph node organogenesis and formation of a lactating mammary gland[1–4]. RANKL and RANK are also expressed in the central nervous system[5,6]. However, the functional relevance of RANKL/RANK in the brain was entirely unknown. Here we report that RANKL and RANK have an essential role in the brain. In both mice and rats, central RANKL injections trigger severe fever. Using tissue-specific Nestin-Cre and GFAP-Cre *rank^{floxed}* deleter mice, the function of RANK in the fever response was genetically mapped to astrocytes. Importantly, Nestin-Cre and GFAP-Cre *rank^{floxed}* deleter mice are resistant to lipopolysaccharide-induced fever as well as fever in response to the key inflammatory cytokines IL-1β and TNFα. Mechanistically, RANKL activates brain regions involved in thermoregulation and induces fever via the COX2-PGE$_2$/EP3R pathway. Moreover, female Nestin-Cre and GFAP-Cre *rank^{floxed}* mice exhibit increased basal body temperatures, suggesting that RANKL and RANK control thermoregulation during normal female physiology. We also show that two children with RANK mutations exhibit impaired fever during pneumonia. These data identify an entirely novel and unexpected function for the key osteoclast differentiation factors RANKL/RANK in female thermoregulation and the central fever response in inflammation.

To test for a possible brain-specific function of the RANKL/RANK system, we performed stereotactic intracerebroventricular (i.c.v.) injections of recombinant RANKL into the lateral ventricle of rats (Supplementary Fig. 1a). Injection of RANKL did not alter serum alkaline phosphatase and serum Ca^{2+} levels (Supplementary Fig. 1b), suggesting that central RANKL administration has no overt effects on osteoclasts. Within minutes after i.c.v. injection, RANKL administration resulted in markedly reduced activity of all animals tested. Further analyses revealed that RANKL i.c.v. injected rats developed very high fever (Fig. 1a, Supplementary Fig. 1c). Heat inactivation of RANKL abolished the fever response (Fig. 1b, Supplementary Fig. 1d), excluding possible endotoxin contaminations. Similar to rats, i.c.v. injections of RANKL into mouse brains triggered hyperthermia (Fig. 1c). As expected from a febrile response[7], i.c.v. injections of RANKL also resulted in markedly reduced activity (Fig. 1d, Supplementary Fig. 2a). *In vivo* inhibition of RANKL with the natural decoy receptor osteoprotegerin alleviated the fever response in mice (Fig. 1e) and rats (Supplementary Fig. 2b). Moreover, i.c.v. injections of RANKL resulted in induction of adrenocorticotropic hormone (ACTH) in the serum and Uncoupling protein 1 (UCP1) in the brown adipose tissue (Supplementary Fig. 3), required to generate heat by non-shivering thermogenesis[8]. By contrast, intraperitoneal (i.p.) delivery even of high doses of RANKL did not result in any changes in body temperature nor in activity (Supplementary Fig. 2c–e). Thus, central nervous system administration of RANKL can trigger fever.

We next set out to map RANK expression in the brain. RANK protein was specifically expressed in the preoptic area (POA) and the Medial Septal nucleus (MSn) (Supplementary Figs 4 and 5). Immunostaining with neuronal (NeuN), astrocyte (GFAP), and microglia (Iba1) markers indicated that RANK is expressed on neurons and astrocytes in the POA/MSn region (Supplementary Figs 4b and 6a–c). Staining for RANKL protein revealed strong expression in the choroid plexus of the ventricles (Supplementary Fig. 7). In addition, *in situ* hybridization for RANKL revealed mRNA expression in the Lateral Septal nucleus (LSn) (Supplementary Figs 4b and 8). Thus, RANKL and RANK are expressed in the POA/MSn/LSn region, key brain regions involved in thermoregulation[7,9].

c-Fos expression has been previously used to map brain regions involved in fever[10,11]. In rat, i.c.v. injection of RANKL resulted in strong c-Fos activation in the POA, MSn, the ventromedial hypothalamus (VMH), the dorsomedial hypothalamus (DMH), and the periventricular nucleus (PVN) (Supplementary Fig. 9). The PVN, VMH, and the DMH regions relay central thermoregulation to stimulation of the sympathetic nervous system[9]. We also observed that i.c.v. RANKL triggers c-Fos activation in the suprachiasmatic nucleus (SCN) (Supplementary Fig. 9), the central regulator of circadian activity. Similar to rats, i.c.v. injections of RANKL into wild-type mice resulted in nuclear accumulation of c-Fos protein in the POA, MSn, PVN, VMH, DMH, and the SCN (not shown). Using c-Fos–GFP (green fluorescent protein) reporter mice[12], we detected increased transcriptional activation of the c-Fos promoter, in the POA, MSn, PVN, VMH, DMH, and the SCN following i.c.v. injection of RANKL (Supplementary Fig. 10). Thus, RANKL/RANK triggers a functional response in brain areas involved in the fever response.

To genetically confirm the role of the RANKL/RANK system in the central fever response, we generated a *rank^{floxed}* allele that would allow us to engineer tissue-specific RANK knockout mice. (For details see Supplementary Methods and Supplementary Fig. 11). Using this line, we generated Nestin-Cre *rank^{floxed}* mice to delete

[1] IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, 1030 Vienna, Austria. [2] Kyoto University Graduate School of Medicine, Department of Pharmacology, Kyoto 606-8501, Japan. [3] Department of Physiology, School of Medicine, University of Occupational and Environmental Health, Kitakyushu 807-8555, Japan. [4] Max Delbrueck Centre for Molecular Medicine, 13125 Berlin, Germany. [5] Center of Integrated Bioinformatics, Max F. Perutz Laboratories, 1030 Vienna, Austria. [6] Oita University Faculty of Medicine, Oita 879-5593, Japan. [7] Institute for Oral Science, Matsumoto Dental University, Nagano 399-0781, Japan. [8] Uludag University Medical Faculty, 16059 Bursa, Turkey. [9] Medical University of Vienna, Center for Brain Research, 1090 Vienna, Austria.

©2009 Macmillan Publishers Limited. All rights reserved

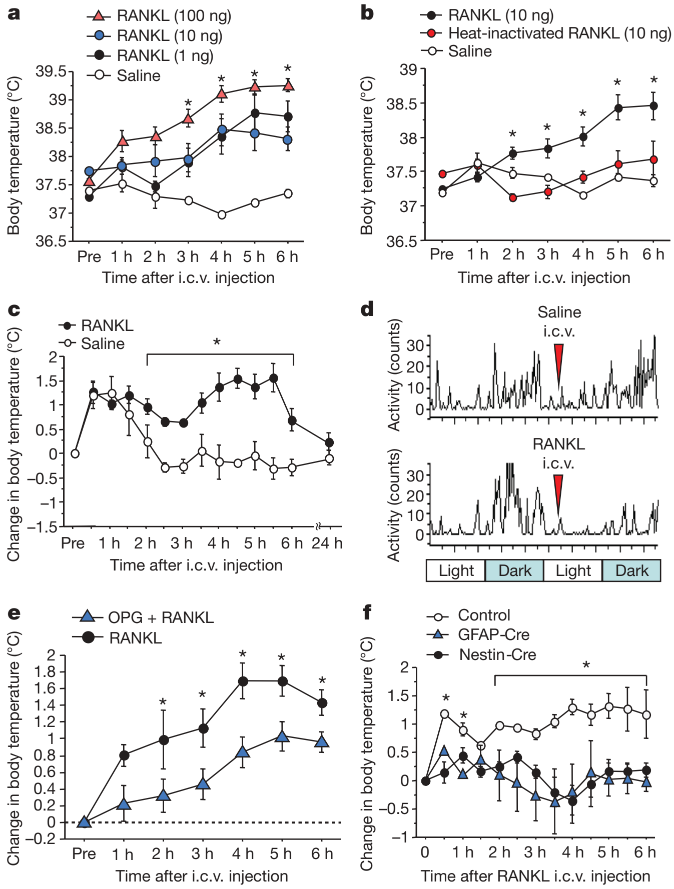


Figure 1 | RANKL/RANK control fever in the central nervous system. **a**, RANKL induces fever in rats. Ten-week-old male Wistar rats were injected i.c.v. with saline ($n = 4$) and 1 ng ($n = 4$), 10 ng ($n = 4$) or 100 ng ($n = 5$) recombinant RANKL, and core body temperatures were measured by telemetry. **b**, Heat inactivation of RANKL abolishes central fever effects. Rats were injected i.c.v. with 10 ng RANKL ($n = 4$) or 10 ng heat-inactivated RANKL ($n = 4$). A saline control is also shown. **c**, **d**, Effect of i.c.v. RANKL injection on core body temperature (**c**) and circadian activity (**d**) in wild-type male mice. Mice were injected i.c.v. with saline ($n = 4$) or 100 ng RANKL ($n = 5$) and monitored by telemetry. Light and dark phases of the day are indicated. Red arrows show time of i.c.v. injections. Activity was determined every 15 min by telemetry. **e**, Osteoprotegerin (OPG) diminishes RANKL-induced fever. Wild-type mice were i.c.v. injected with 10 ng RANKL or 10 ng RANKL + 1 μg osteoprotegerin ($n = 5$). **f**, RANKL-induced fever is diminished in Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ mice. *rank*$^{floxed/\Delta}$ mice were used as controls. Mice were injected i.c.v. with 100 ng RANKL ($n = 4$ per group) and core body temperatures were measured by telemetry. Data in **a**–**c**, **e**, and **f** are mean values ± s.e.m. $^*P < 0.05$ (ANOVA with a post-hoc Fisher's test).

RANK in the brain (Supplementary Figs 4a and 12). Nestin-Cre *rank*$^{floxed/\Delta}$ mice are viable, healthy, exhibit normal fertility, and are in general indistinguishable from their wild-type littermates, including normal lymph node organogenesis, or normal lactation during pregnancy. Importantly, whereas i.c.v. injection of RANKL into control mice resulted in severe hyperthermia, Nestin-Cre *rank*$^{floxed/\Delta}$ mice did not develop fever (Fig. 1f).

Nestin-Cre functions in neuronal progenitor cells resulting in gene deletion in neurons as well as astrocytes[13]. To dissect whether RANK functions in neurons and/or astrocytes, we next generated GFAP-Cre *rank*$^{floxed/\Delta}$ mice to inactivate RANK specifically in astrocytes[14]. Similar to the Nestin-Cre *rank*$^{floxed/\Delta}$ mice, GFAP-Cre *rank*$^{floxed/\Delta}$ mice are viable, fertile, appear healthy, and are indistinguishable from their littermates. Consistent with our data that RANK is expressed in both neurons and in astrocytes (Supplementary Fig. 6), immunostaining of the POA/MSn region in GFAP-Cre *rank*$^{floxed/\Delta}$ mice showed reduced numbers of RANK-positive cells (Supplementary Fig. 4a). Intriguingly, in GFAP-Cre *rank*$^{floxed/\Delta}$ mice we also observed a markedly abrogated fever in response to i.c.v. RANKL (Fig. 1f), suggesting that RANK expression on astrocytes is required to induce fever. Whereas wild-type mice exhibited reduced activity, diurnal activity was not changed in Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ mice following an i.c.v. challenge with RANKL (Supplementary Fig. 13). RANKL i.c.v. injections also did not induce significant c-Fos protein expression in the brain of Nestin-Cre *rank*$^{floxed/\Delta}$ or GFAP-Cre *rank*$^{floxed/\Delta}$ mice beyond that observed in control mice (not shown). These data show that genetic inactivation of RANK in the central nervous system abrogates the fever response to RANKL.

Can RANKL/RANK mediate hyperthermia during a 'real' inflammation such as lipopolysaccharide (LPS)-induced fever[11]? Whereas i.p. injection of LPS induced high fever in control mice, this response

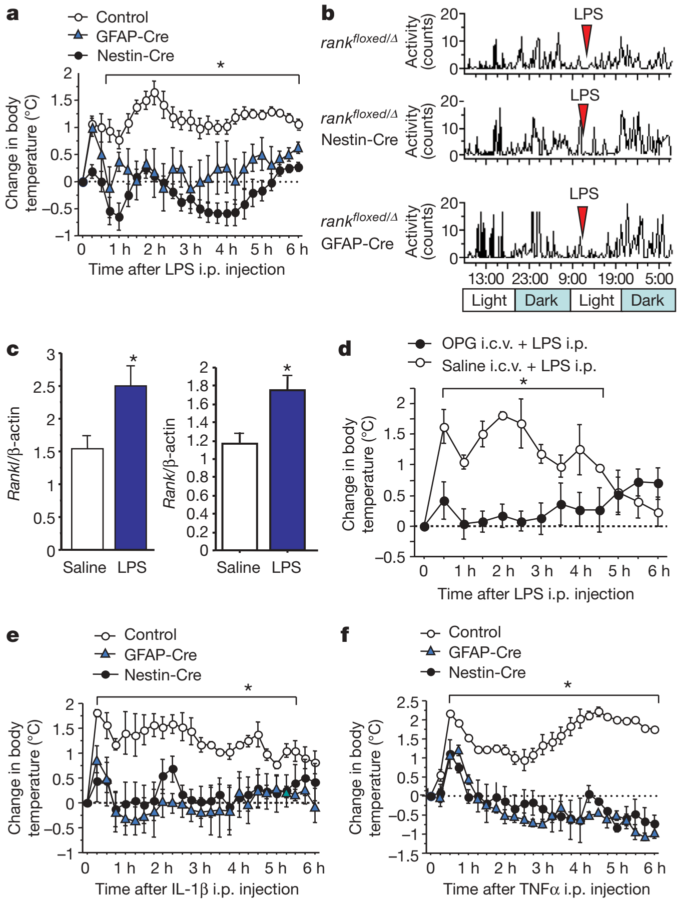


Figure 2 | Central RANK mediates the inflammatory fever response. **a**, Changes in core body temperature after LPS (10 μg per mouse, i.p.) challenge of *rank*$^{floxed/\Delta}$ (control, $n = 4$), Nestin-Cre *rank*$^{floxed/\Delta}$ ($n = 5$) and GFAP-Cre *rank*$^{floxed/\Delta}$ ($n = 5$) mice. **b**, Representative diurnal activity of *rank*$^{floxed/\Delta}$, Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ littermates injected with LPS (10 μg per mouse, i.p.). Activity was determined every 15 min by telemetry. **c**, LPS triggers RANKL and RANK expression in the POA/MSn brain region. Mice were injected with saline i.p. ($n = 4$) or LPS 10 μg per mouse i.p. ($n = 4$). Expression of *Rankl* or *Rank* mRNA relative to β-actin was determined by quantitative PCR. **d**, Central administration of the RANKL decoy receptor osteoprotegerin (1 μg per mouse, i.c.v.) results in an abolished fever response to i.p. LPS (10 μg per mouse, i.p.) in wild-type mice ($n = 4$ per group). **e**, Changes in core body temperature after IL-1β (10 μg per mouse, i.p.) challenge of Nestin-Cre *rank*$^{floxed/\Delta}$ ($n = 5$), GFAP-Cre *rank*$^{floxed/\Delta}$ ($n = 5$), and control *rank*$^{floxed/\Delta}$ ($n = 4$) mice. **f**, Changes in body temperature after TNFα (2 μg per mouse, i.p.) challenge of Nestin-Cre *rank*$^{floxed/\Delta}$ ($n = 5$), GFAP-Cre *rank*$^{floxed/\Delta}$ ($n = 4$), and *rank*$^{floxed/\Delta}$ littermates ($n = 4$). Data are mean values ± s.e.m. $^*P < 0.05$ (ANOVA with a post-hoc Fisher's test).

©2009 Macmillan Publishers Limited. All rights reserved

was markedly alleviated in Nestin-Cre *rank*$^{floxed/\Delta}$ as well as GFAP-Cre *rank*$^{floxed/\Delta}$ mice (Fig. 2a, Supplementary Fig. 14a–c). LPS-treated Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ mice also exhibited almost normal circadian activity (Fig. 2b, Supplementary Fig. 14d). As expected, LPS i.p. injections resulted in increased serum levels of the inflammatory mediators IL-1β, IL-6 and TNFα (Supplementary Fig. 15). In the brain, i.p. LPS injections resulted in increased levels of IL-1β and IL-6, as well as slightly increased levels of RANKL, whereas osteoprotegerin levels were decreased (Supplementary Fig. 16). Importantly, we observed induction of *Rankl* and *Rank* mRNA in the POA/MSn/LSn region following i.p. LPS injection (Fig. 2c), suggesting that local RANKL and RANK induction as well as downregulation of the negative regulatory RANKL decoy receptor osteoprotegerin might contribute to the response. In line with this hypothesis, i.c.v. injection of osteoprotegerin in mice (Fig. 2d, Supplementary Fig. 17a) and rats (Supplementary Fig. 17b) abolished i.p. LPS-induced fever. Thus, RANKL/RANK control fever in response to LPS.

LPS-induced fever is mediated via pro-inflammatory cytokines such as TNFα or IL-1β[11,15,16]. To assess whether RANKL/RANK act upstream or downstream of these pro-inflammatory cytokines, we injected IL-1β and TNFα into Nestin-Cre *rank*$^{floxed/\Delta}$ mice. Whereas i.p. injections of IL-1β and TNFα induce high fever in control mice, loss of RANK expression in neurons/astrocytes resulted in an impaired fever response (Fig. 2e, f). Whereas i.c.v. injection of TNFα or IL-1β triggers a high fever response in control mice, Nestin-Cre *rank*$^{floxed/\Delta}$ mice did not develop fever following i.c.v. IL-1β or i.c.v. TNFα injections (Supplementary Fig. 18). Similar to LPS injections, the circadian activity of IL-1β- and TNFα-challenged Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ mice was much less affected than in control mice (Supplementary Fig. 19a–d). Thus, RANK is a critical mediator of fever in response to LPS and the pro-inflammatory cytokines IL-1β and TNFα.

Prostaglandins are crucial fever mediators in the central nervous system via activation of the prostaglandin receptors in the POA region[15]. Injection of RANKL induced expression of cyclooxygenase (COX2), required for prostaglandin synthesis (Fig. 3a). In the POA/MSn region, we observed co-localization of RANK and COX2 (Fig. 3b). Treatment with the non-selective COX1/2 inhibitor indomethacin abolished the fever response (Supplementary Fig. 20a). Further experiments in mice and rats revealed that pharmacological inhibition of COX2, but not COX1, abrogated the RANKL-induced fever response (Fig. 3c, Supplementary Fig. 20b). We next administered RANKL i.c.v. into the brains of mice that carry mutations in the prostaglandin EP3 receptor[17]. Central RANKL injections into EP3 receptor mutant mice did not trigger fever (Fig. 3d) or changes in circadian activity (Supplementary Fig. 20c). In *ex vivo* POA/MSn/LSn brain slice cultures (Supplementary Fig. 20d), RANKL induced PGE_2 production that could be inhibited by osteoprotegerin administration (Fig. 3e). Osteoprotegerin also blocked IL-1β-induced PGE_2 production in cultures of such slices (Fig. 3e). Osteoprotegerin treatment significantly decreased PGE_2 production, suggesting a basal function of the RANKL/RANK system in local PGE_2 homeostasis. Importantly, RANKL treatment failed to induce PGE_2 production in *ex vivo* POA/MSn/LSn brain slices from Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ mice (Fig. 3f). These results show that RANKL/RANK can directly induce PGE_2 production in isolated brain slices and that RANKL/RANK mediate the febrile response via induction of the COX2–PGE_2–EP3 receptor system (Supplementary Fig. 20e).

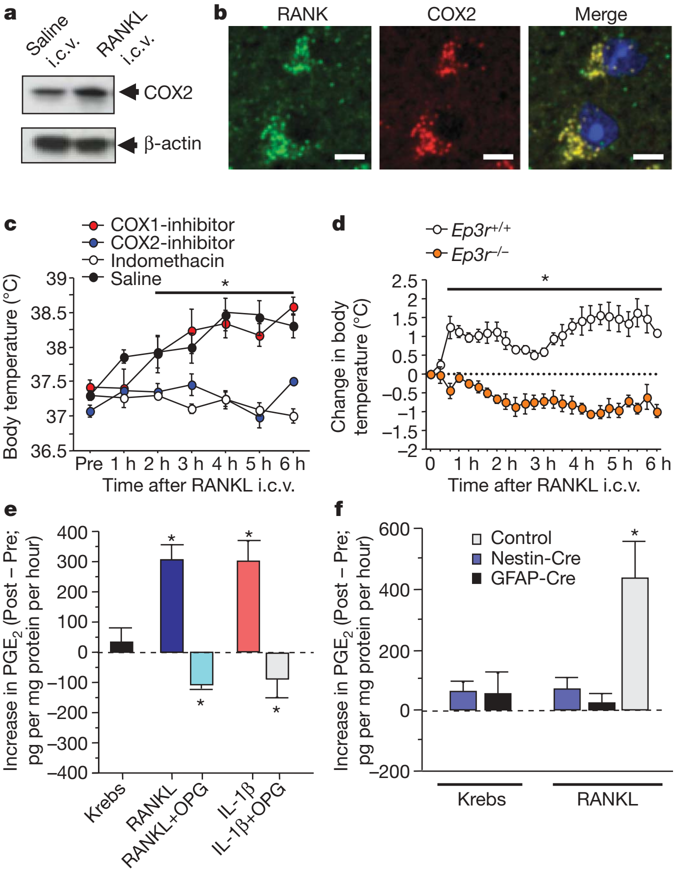


Figure 3 | RANKL induces PGE_2 and mediates fever via the EP3R. **a**, Western blot analysis of COX2 expression in the midbrain of wild-type mice two hours after i.c.v. RANKL injection (100 ng per mouse). Data are representative of four different animals. **b**, Co-localization of COX2 (red) and RANK (green) using immunofluorescence staining in POA/MSn regions 3 h after RANKL injections (100 ng per mouse, i.c.v.). DAPI counterstaining is also shown for the merged image. Magnifications ×63, scale bars, 20 μm. We note that in other brain regions that stained positive for COX2 protein, we did not see RANK co-expression (not shown). **c**, Treatment with the COX1/2 blocker indomethacin (3 mg per rat, subcutaneously) and the COX2 inhibitor SC-236 (1.5 mg per rat, i.p.), but not the COX1 blocker SC-560 (1.5 mg per rat, i.p.), abolishes RANKL-induced (10 ng, i.c.v.) fever in 10-week-old male Wistar rats ($n = 4$ per group). Core body temperatures were measured by telemetry. **d**, Genetic inactivation of the EP3 receptor abolishes the febrile effects of RANKL. RANKL (100 ng per mouse, i.c.v.) was injected into *Ep3r*$^{-/-}$ ($n = 7$) and control *Ep3r*$^{+/+}$ ($n = 4$) littermates. **e**, Production of PGE_2 in *ex vivo* POA/MSn brain slices prepared from wild-type male mice. Brain slices were perfused with Krebs solution and stimulated with either vehicle (Krebs; $n = 13$), RANKL (140 ng ml^{-1}; $n = 16$), IL-1β (140 ng ml^{-1}; $n = 13$), RANK + osteoprotegerin ($n = 7$), IL-1β + osteoprotegerin ($n = 7$) for 1 h in wild-type mice. **f**, Production of PGE_2 in *ex vivo* POA/MSn brain slices prepared from control *rank*$^{floxed/\Delta}$ ($n = 4$), Nestin-Cre *rank*$^{floxed/\Delta}$ ($n = 6$), and GFAP-Cre *rank*$^{floxed/\Delta}$ ($n = 3$) male mice perfused with RANKL (140 ng ml^{-1}) for 1 h. The Krebs vehicle control is shown ($n = 3$ per group). All values are mean ± s.e.m. *$P < 0.05$ (ANOVA with a post-hoc Fisher's test).

We have previously shown that RANKL and RANK regulate formation of a lactating mammary gland during pregnancy[1,18]. We therefore speculated that RANKL/RANK might regulate female body temperature. Both Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ females exhibited a significant increase in basal core body temperature during the light phase as compared to littermates (Fig. 4a, Supplementary Figs 21 and 22), resulting in a markedly smaller difference in body temperature (Fig. 4b). Ovariectomy-induced changes in core body temperatures occurred in control mice but not in Nestin-Cre *rank*$^{floxed/\Delta}$ females (Fig. 4c, Supplementary Fig. 23) or in GFAP-Cre *rank*$^{floxed/\Delta}$ females (not shown). Moreover, ovariectomy resulted in a marked downregulation of *Rankl* mRNA expression in the POA/MSn/LSn brain region (Fig. 4d). *Rank* mRNA expression was not affected in the brain of ovariectomized females. In male mice, loss of RANK in neurons/astrocytes has no significant effect on basal circadian body temperatures (Supplementary Fig. 24). Thus, genetic inactivation of RANK in the brain results in altered physiological thermoregulation in female mice which at least in part appears to be regulated by ovarian sex hormones.

©2009 Macmillan Publishers Limited. All rights reserved

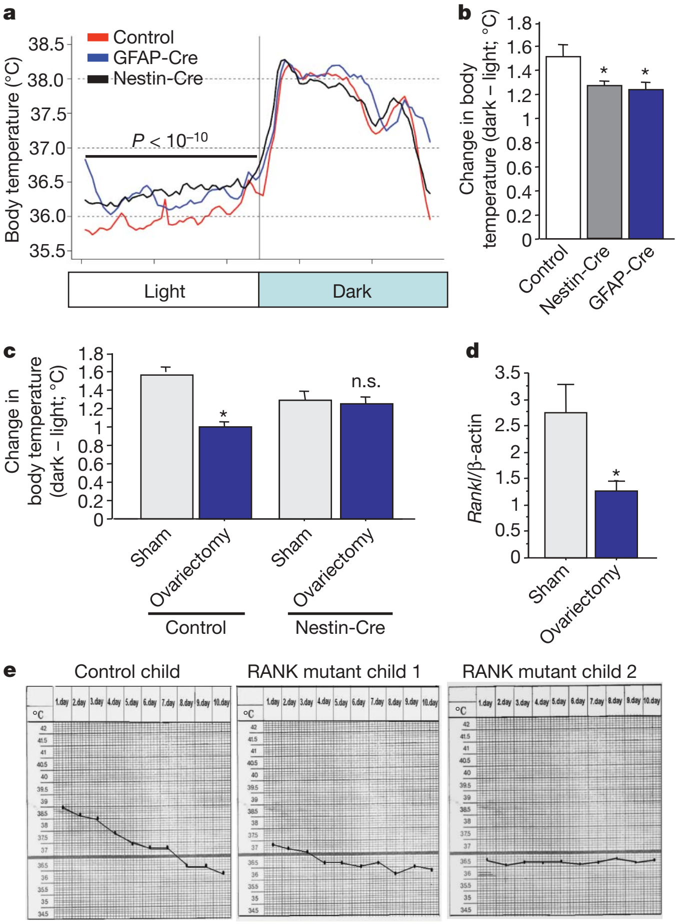


Figure 4 | RANK controls thermoregulation in female mice and fever in children with RANK mutations. a, Average diurnal core body temperatures of control *rank*$^{floxed/\Delta}$, Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ female littermate mice. Lines represent average temperatures for 20 days for all mice in each cohort. **b**, Differences in the average body temperatures in the light and dark phases in control *rank*$^{floxed/\Delta}$, Nestin-Cre *rank*$^{floxed/\Delta}$ and GFAP-Cre *rank*$^{floxed/\Delta}$ female littermates. $n = 4$–5 mice per group with averaged body temperatures for 30 days for each mouse. Values are mean ± s.e.m. **c**, Differences in the average body temperatures in the light and the dark phases in control wild-type and Nestin-Cre *rank*$^{floxed/\Delta}$ female littermates without (control, $n = 6$) or after ovariectomy ($n = 7$). Values are mean ± s.e.m. core body temperatures measured for 20 days by telemetry. **d**, Ovariectomy diminishes *Rankl* mRNA levels in the POA/MSn region. *Rankl* and β-actin mRNA levels were determined via quantitative PCR. Relative expression of *Rankl* to β-actin is shown as mean values ±s.e.m. *Rank* mRNA expression was not affected in the brain of ovariectomized females (not shown). **e**, Body temperatures in a control child and two RANK mutant siblings with severe pneumonia. Scans of the 'original' fever chart are shown. Antibiotic treatment was initiated on day 1. *$P < 0.05$ (ANOVA with a post-hoc Fisher's test). n.s., not significant.

Recently, in a Turkish consanguineous family, a homozygous RANK mutation (Arg170Gly mutation) was identified in two children, leading to severe autosomal-recessive osteopetrosis[19]. We therefore speculated that these two patients (Supplementary Fig. 25) might have impaired fever responses in their clinical history. 'Normal' children with pneumonia develop very high fever that then decreases to baseline after antibiotic treatment. Although hospitalized for severe pulmonary infections confirmed by serology and chest X-rays, both siblings with autosomal-recessive osteopetrosis had a markedly abrogated fever response compared to age-matched children with pneumonia (Fig. 4e, Supplementary Fig. 26). Thus, RANKL/RANK also control fever in humans.

Our experiments have uncovered an essential new player in the inflammatory fever response and identified a previously unknown function of the RANKL/RANK system in the central nervous system. Fever is a ubiquitous adaptive response among animals to improve survival during infections and to promote healing[7]. Fever-inducing cytokines such as TNFα and IL-1β are not only induced during infections but also during inflammation in physical and chemical tissue damage, many types of cancer, allergies, and even in obesity[7,20,21]. Moreover, our data show that RANKL/RANK mediate gender-specific, physiological thermoregulation. Interestingly, one symptom associated with osteoporosis and hormonal changes in older women is hot flashes, sudden bursts of high body temperature[9,22]. RANKL/RANK might explain such symptoms. Our data identify RANKL/RANK as key thermoregulators that directly link bone physiology to the central control of female body temperature and fever in inflammation.

METHODS SUMMARY

Animals. Mice carrying the *rank*^{floxed} or *rank*$^{\Delta}$ alleles were backcrossed five times to C57BL/6 mice before generating Nestin-Cre *rank*$^{floxed/\Delta}$, GFAP-Cre *rank*$^{floxed/\Delta}$ and K5-Cre *rank*$^{floxed/\Delta}$ mice. c-Fos–GFP and EP3R mutant mice have been previously reported[12,17]. The rats we used were male Wistar animals.
Immunohistochemistry. c-Fos, RANK and double immunostainings with K5, NeuN, GFAP and Iba1 were performed on frozen brain and thymus sections. For GFP visualization in c-Fos–GFP transgenics, brains were fixed in paraformaldehyde. *In situ* hybridization for RANKL was performed on frozen coronal brain sections.
RANKL i.c.v. administration and telemetry. The lateral ventricles of rats or mice were cannulated and animals injected i.c.v. with recombinant murine RANKL, IL-1β, TNFα, or recombinant human osteoprotegerin (all Peprotech). Core body temperatures were determined using implanted telemetry probes. Temperature and activity counts were automatically measured every 15 min. All animal experiments were performed according to approved procedures.
***Ex vivo* brain slices cultures.** POA/MSn brain slices were prepared from wild-type, Nestin-Cre *rank*$^{floxed/\Delta}$, or GFAP-Cre *rank*$^{floxed/\Delta}$ male mice and perfused with Krebs-HCO$_3$ buffer containing either vehicle, osteoprotegerin (140 ng ml^{-1}), RANKL (140 ng ml^{-1}), IL-1β (140 ng ml^{-1}), RANKL + osteoprotegerin, or IL-1β + osteoprotegerin for 1 h. Perfusates were collected for 1 h and PGE$_2$ levels were determined using an enzyme immunoassay.
Human RANK mutants. The two patients with autosomal-recessive osteopetrosis have been described previously and carry a homozygous 385C to 385T transition in the extracellular domain of RANK[19]. Pneumonia was confirmed by serology and chest X-rays. The children were treated with cefuroxim (150 mg kg^{-1}) and amikacin (15 mg kg^{-1}) for 15 days.
Statistics. All values in the paper are given as means ± s.e.m. Comparisons between groups were made by Student's *t*-test and analysis of variance (ANOVA) with a post-hoc Fisher's test. In female body temperature studies, Kruskal–Wallis with post-hoc Mann–Whitney U test was used.

Received 29 September; accepted 22 October 2009.

1. Leibbrandt, A. & Penninger, J. M. RANK/RANKL: regulators of immune responses and bone physiology. *Ann. NY Acad. Sci.* **1143**, 123–150 (2008).
2. Kong, Y. Y, *et al.* OPGL is a key regulator of osteoclastogenesis, lymphocyte development and lymph-node organogenesis. *Nature* **397**, 315–323 (1999).
3. Cummings, S. R. *et al.* Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N. Engl. J. Med.* **361**, 756–765 (2009).
4. Smith, M. R. *et al.* Denosumab in men receiving androgen-deprivation therapy for prostate cancer. *N. Engl. J. Med.* **361**, 745–755 (2009).
5. Kartsogiannis, V. *et al.* Localization of RANKL (receptor activator of NFκ B ligand) mRNA and protein in skeletal and extraskeletal tissues. *Bone* **25**, 525–534 (1999).
6. Nakagawa, N. *et al.* RANK is the essential signaling receptor for osteoclast differentiation factor in osteoclastogenesis. *Biochem. Biophys. Res. Commun.* **253**, 395–400 (1998).
7. Dantzer, R. Cytokine-induced sickness behavior: mechanisms and implications. *Ann. NY Acad. Sci.* **933**, 222–234 (2001).
8. Cannon, B. & Nedergaard, J. Brown adipose tissue: function and physiological significance. *Physiol. Rev.* **84**, 277–359 (2004).
9. Morrison, S. F., Nakamura, K. & Madden, C. J. Central control of thermogenesis in mammals. *Exp. Physiol.* **93**, 773–797 (2008).
10. Sagar, S. M., Sharp, F. R. & Curran, T. Expression of c-fos protein in brain: metabolic mapping at the cellular level. *Science* **240**, 1328–1331 (1988).
11. Elmquist, J. K., Scammell, T. E. & Saper, C. B. Mechanisms of CNS response to systemic immune challenge: the febrile response. *Trends Neurosci.* **20**, 565–570 (1997).
12. Fleischmann, A. *et al.* Impaired long-term memory and NR2A-type NMDA receptor-dependent synaptic plasticity in mice lacking c-Fos in the CNS. *J. Neurosci.* **23**, 9116–9122 (2003).

©2009 Macmillan Publishers Limited. All rights reserved

13. Tronche, F. *et al.* Disruption of the glucocorticoid receptor gene in the nervous system results in reduced anxiety. *Nature Genet.* **23,** 99–103 (1999).
14. Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Genes Dev.* **14,** 994–1004 (2000).
15. Blatteis, C. M., Li, S., Li, Z., Feleder, C. & Perlik, V. Cytokines, PGE2 and endotoxic fever: a re-assessment. *Prostaglandins Other Lipid Mediat.* **76,** 1–18 (2005).
16. McDermott, M. F. & Tschopp, J. From inflammasomes to fevers, crystals and hypertension: how basic research explains inflammatory diseases. *Trends Mol. Med.* **13,** 381–388 (2007).
17. Ushikubi, F. *et al.* Impaired febrile response in mice lacking the prostaglandin E receptor subtype EP3. *Nature* **395,** 281–284 (1998).
18. Fata, J. E. *et al.* The osteoclast differentiation factor osteoprotegerin-ligand is essential for mammary gland development. *Cell* **103,** 41–50 (2000).
19. Guerrini, M. M. *et al.* Human osteoclast-poor osteopetrosis with hypogammaglobulinemia due to TNFRSF11A (RANK) mutations. *Am. J. Hum. Genet.* **83,** 64–76 (2008).
20. Handschin, C. & Spiegelman, B. M. The role of exercise and PGC1alpha in inflammation and chronic disease. *Nature* **454,** 463–469 (2008).
21. Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. Cancer-related inflammation. *Nature* **454,** 436–444 (2008).
22. Isselbacher, K. J. *et al. Harrison's Principles of Internal Medicine 2*, 13th edn, 2021–2022 (McGraw-Hill, 1994).

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

Acknowledgements We thank C. Xia, A. Muto, T. Mizoguchi, T. Katakai, T. Mitsumori, D. Sakata, T. Matsuoka, I. Williams, M. Iehara, T. Katafuchi, K. Matsuo, J. Wojciechowski, C. Theussl, T. Nakashima, T. Wada and R. Koglgruber for their assistance and all members of our laboratory for discussions. We thank E. Wagner for the c-Fos–GFP reporter mice and M. Kopf for the IL-1Rα mutant mice. This work was in part supported by UEHARA Foundation and Japan Foundation for Applied Enzymology grants. S.K. and A.v.H. are supported by the Austrian Ministry for Science and Research (GEN-AU Bioinformatics Integration Network II) and the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF). J.M.P. is supported by grants from IMBA, the Austrian Ministry of Sciences, the Austrian Academy of Sciences, GEN-AU (AustroMouse), an EU Marie Curie Excellence Grant, and an EU ERC Advanced Grant.

Author Contributions R.H. carried out the experiments with help from T.H. A.L. generated *rank*^{floxed} mice. S.K. and A.v.H. provided professional biostatistics support for data analysis. V.K. helped with immunostaining. H.M. and H.Y. provided key reagents and technical help for i.c.v. injections. H.F. and Y.U. performed RANKL in situ hybridizations. J.T. and M.P. helped in cytokine enzyme-linked immunosorbent assays (ELISAs) and UCP1 expression. N.T. developed the reproducible RANK staining protocol in his laboratory. S.K., T.F. and S.N. provided EP3R mutant mice and established the perfusion of the brain slice and quantification of PGE$_2$ procedure. F.Q., R.P. and M.B. performed rat telemetry experiments. S.S.K. is the clinician of the two RANK mutant children and provided their fever data. C.P. helped with rat experiments. J.M.P. coordinated the project, wrote the manuscript, and together with R.H. designed the experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.M.P. (josef.penninger@imba.oeaw.ac.at).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

CD8$^+$ T lymphocyte mobilization to virus-infected tissue requires CD4$^+$ T-cell help

Yusuke Nakanishi[1], Bao Lu[2], Craig Gerard[2] & Akiko Iwasaki[1]

CD4$^+$ T helper cells are well known for their role in providing critical signals during priming of cytotoxic CD8$^+$ T lymphocyte (CTL) responses *in vivo*. T-cell help is required for the generation of primary CTL responses as well as in promoting protective CD8$^+$ memory T-cell development[1]. However, the role of CD4 help in the control of CTL responses at the effector stage is unknown. Here we show that fully helped effector CTLs are themselves not self-sufficient for entry into the infected tissue, but rely on the CD4$^+$ T cells to provide the necessary cue. CD4$^+$ T helper cells control the migration of CTL indirectly through the secretion of IFN-γ and induction of local chemokine secretion in the infected tissue. Our results reveal a previously unappreciated role of CD4 help in mobilizing effector CTL to the peripheral sites of infection where they help to eliminate infected cells.

Elimination of invading pathogens often requires coordinated effort by effector lymphocytes for containment and clearance. Successful defence against intracellular pathogens requires neutralizing antibodies and CTL responses, both of which largely depend on CD4$^+$ T-cell help. Whereas the generation of primary CD8$^+$ T-cell responses to non-inflammatory antigens[2–4] and certain virus infections, such as herpes simplex virus (HSV)[5], require CD4$^+$ T-cell help, primary CTL responses to acute infection with *Listeria monocytogenes* and lymphocytic choriomeningitis virus can occur in the absence of CD4$^+$ T cells. Instead, the latter type of infections requires CD4 help in promoting memory CTL development[6–8]. The critical role of CD4 help in the priming and maintenance of CTL responses is well characterized; however, whether CD4$^+$ T cells help at the stages after CTL differentiation has not been described and is at present unknown.

To explore the role of CD4 help in effector CTL responses, we used a physiological model of local virus infection that enables tracking of antigen-specific CD8$^+$ T cells. HSV-2 infects humans through sexual contact and causes genital herpes. When inoculated into the vaginal cavity, HSV-2 replicates predominantly in the mucosal epithelial cells and establishes latency in the innervating neurons. Because the viral infection is localized, the genital herpes model enabled us to dissect the role of CD4 help in CTL migration to the site of infection. To avoid neurovirulence associated with wild-type HSV-2, without compromising the ability to prime robust innate and adaptive immunity, we used thymidine-kinase (TK)-defective HSV-2 (TK$^-$ HSV-2)[9]. After TK$^-$ HSV-2 infection, both CD4$^+$ and CD8$^+$ T cells are primed in the local draining lymph nodes[10], and both total (Supplementary Fig. 1a) and virus-specific (Supplementary Fig. 1b, c) effector T cells migrate into the vaginal mucosa starting with CD4$^+$ T cells around day 3–4, followed by CD8$^+$ T cells on day 4–5. Notably, migration of virus-specific CD8$^+$ T cells to the infection site was highly dependent on the presence of CD4$^+$ T cells, evidenced by the failure of CD8$^+$ T cells to migrate to the local tissue in mice that were either CD4-deficient, or depleted of CD4$^+$ T cells (Supplementary Fig. 2a). However, because primary CTL expansion after HSV-1 infection has been reported to depend on CD4$^+$ T cells[5] through their ability to license dendritic cells[11], we examined the total number of congenically marked (CD45.1$^+$) HSV-gB-specific T-cell receptor (TCR) transgenic T cells (gBT-I)[12] generated in *Cd4*$^{-/-}$ and CD4-depeleted mice. Consistent with previous reports[5,11], gBT-I responses in various tissues after local HSV-2 infection also depended largely on the presence of CD4$^+$ T cells (Supplementary Fig. 2b–d).

To determine the mechanism by which CD4$^+$ T cells license CTL migration, fully 'helped' CD8$^+$ effector T cells were first generated in wild-type hosts (Fig. 1a). A physiological number (2 × 10^5 cells per mouse[13]) of gBT-I cells were transferred into naive wild-type mice. Subsequently, these mice were infected with TK$^-$ HSV-2, and effector CD8$^+$ T cells were isolated (Supplementary Fig. 3a, b) and transferred into recipient mice that had been infected with TK$^-$ HSV-2 3.5 days earlier. It is well known that effector CTLs migrate to various lymphoid and non-lymphoid organs including the lung, liver and intestine[14,15]. Accordingly, effector CD8$^+$ T cells were found in lymphoid and peripheral organs irrespective of the infection status of the host (Fig. 1b–d). This homeostatic distribution pattern did not depend on the presence of CD4$^+$ T cells. In stark contrast, although the fully helped effector CTLs migrated into the infected vaginal tissue in the wild-type hosts, their ability to do so was significantly impaired in the absence of CD4$^+$ T cells (Fig. 1e and Supplementary Fig. 3c). Similar results were obtained using a different time course (Supplementary Fig. 4). In contrast, adoptively transferred fully helped gBT-I T cells were able to migrate to the HSV-infected vagina in CD8-deficient hosts (Fig. 1f), indicating that CD4, but not CD8, T cells are required for licensing CTL entry into the vaginal mucosa. T regulatory (T$_{reg}$) cells have been shown to facilitate early protective responses to local HSV-2 infection by allowing a timely entry of immune cells into infected tissue[16]. To examine whether effector or Foxp3$^+$ CD4$^+$ T cells account for the mobilization of CTL into the infected vaginal mucosa, either total or Foxp3$^-$ HSV-primed CD4$^+$ T cells were adoptively transferred into HSV-infected *Cd4*$^{-/-}$ hosts. The analysis of migration of helped gBT-I cells in such animals showed that effector CD4$^+$ T cells were equally capable of CTL mobilization into the infected tissue as compared to total CD4$^+$ T cells (including effectors and T$_{reg}$ cells) (Supplementary Fig. 5). These data indicated that although T$_{reg}$ cells are capable of facilitating effector lymphocyte entry[16], effector CD4$^+$ T cells alone mediate CTL recruitment and can override the requirement for T$_{reg}$ cells. Collectively, our results showed that, although homeostatic migration of effector CTLs occurs independently of CD4$^+$ T-cell help, fully differentiated CTLs are not self-sufficient for accelerated recruitment to the infected tissue during a viral infection. This situation is reminiscent of the requirement for 'pioneering' CD4$^+$ T cells for entry by pathogenic CD4$^+$ T cells in the central nervous system[17].

[1] Department of Immunobiology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [2] Pulmonary Division, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.

©2009 Macmillan Publishers Limited. All rights reserved

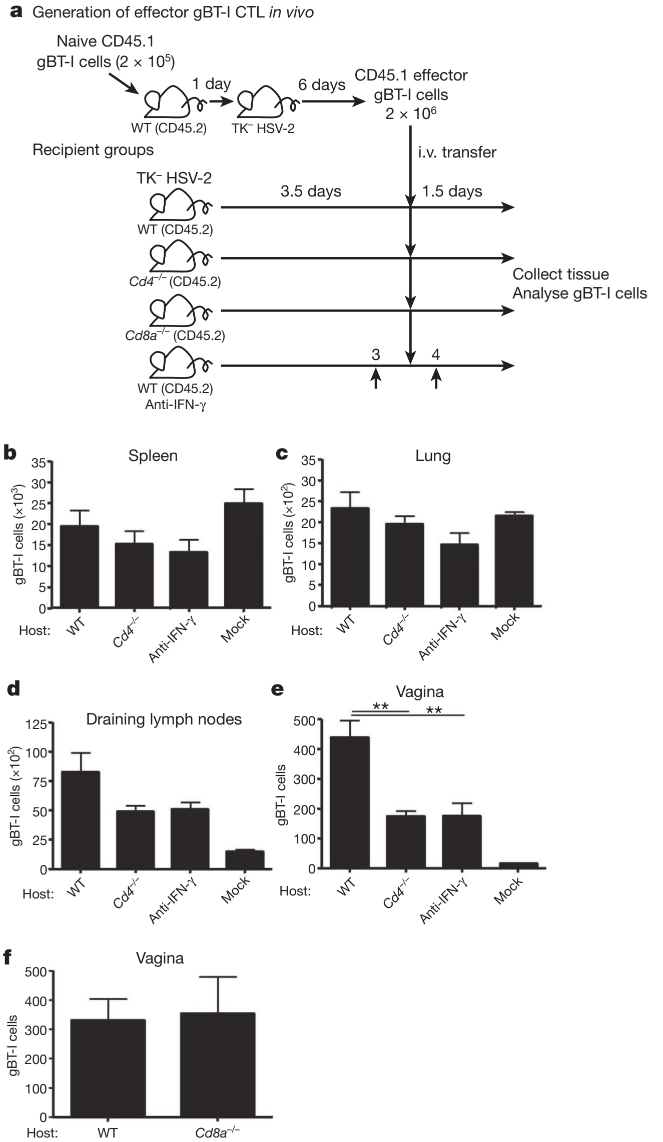


Figure 1 | CD4 help is required for CTL migration into the vaginal mucosa after HSV-2 infection. a, Naive congenic gBT-I cells (2×10^5) were transferred into wild-type (WT) mice. Six days after TK$^-$ HSV-2 vaginal infection, 2×10^6 effector gBT-I cells isolated from these mice were transferred into secondary hosts (some treated with anti-IFN-γ antibody) infected with TK$^-$ HSV-2 3.5 days earlier. i.v., intravenous. **b–f**, Numbers of gBT-I cells in the indicated tissues were assessed 5 days after infection. The data are pooled from two independent experiments and presented as mean and s.e.m. ($n = 5$–6 mice per group). $^{**}P < 0.01$.

Furthermore, our data uncovered a previously unknown function of T helper cells in mobilizing CD8$^+$ T-cell recruitment to the site of infection.

To address the mechanism by which T-cell help enables CTL recruitment to the site of viral replication, we focused on one of the major effector functions of CD4$^+$ T$_h$1 cells, namely, the secretion of IFN-γ. First, we asked whether CD4-catalysed CTL recruitment was mediated by the action of IFN-γ. To this end, CTL migration to the infected tissue in wild-type and *Ifngr*$^{-/-}$ mice was compared. CD8$^+$ T-cell recruitment to the vagina was significantly reduced in *Ifngr*$^{-/-}$ mice (Supplementary Fig. 6a), despite the fact that these mice generated comparable levels of HSV-2-specific CD4$^+$ and CD8$^+$ T-cell responses (Supplementary Fig. 6b, c). Vaginal infection with HSV-2 generates two waves of IFN-γ secretion; the first wave of IFN-γ is secreted by natural killer (NK) cells at 2 days post infection, whereas CD4$^+$ T cells produce IFN-γ starting at 4 days post infection (Supplementary Fig. 7a)[18]. To examine specifically the requirement for CD4$^+$ T-cell-secreted IFN-γ in CTL recruitment, wild-type mice were infected with TK$^-$ HSV-2, and IFN-γ neutralizing antibody was given at 3 and 4 days post infection. With this regimen, NK-secreted IFN-γ is not affected, whereas CD4-secreted IFN-γ is blocked. Fully helped CD8$^+$ T cells, when injected into IFN-γ neutralized mice, were significantly impaired in their migration to the infected tissue (Fig. 1e). Furthermore, NK depletion throughout the course of HSV infection did not alter the recruitment of effector CTL to the vaginal mucosa in either wild-type or *Cd4*$^{-/-}$ hosts (Supplementary Fig. 8). These results, combined with the fact that CD4$^+$ T cells are required for migration of CTLs, indicated that CTL recruitment primarily, if not exclusively, depends on IFN-γ produced by CD4$^+$ T cells.

To determine whether CD4$^+$ T cells indeed provide the IFN-γ required to pioneer the vaginal mucosa for CTL entry, we tested both sufficiency and requirement for CD4-derived IFN-γ in this process. First, to examine the requirement for CD4-derived IFN-γ, migration of fully helped gBT-I cells was assessed in *Cd4*$^{-/-}$ mice reconstituted with either wild-type or *Ifng*$^{-/-}$ HSV-primed CD4$^+$ T cells (Fig. 2a). Only the wild-type, and not *Ifng*$^{-/-}$, effector CD4$^+$ T cells were able to rescue CTL migration to the site of infection (Fig. 2a). These data indicated that IFN-γ secretion from CD4$^+$ T cells is required for CTL entry into the vaginal mucosa after HSV-2 infection. Second, sufficiency of CD4-derived IFN-γ in CTL migration was examined by reconstituting *Ifng*$^{-/-}$ mice with either wild-type or *Ifng*$^{-/-}$ HSV-primed CD4$^+$ T cells. Notably, migration of gBT-I effector cells was facilitated by the presence of wild-type CD4$^+$ T cells, whereas *Ifng*$^{-/-}$ CD4$^+$ T cells were only able to promote a basal level of CTL recruitment (Fig. 2f). Interestingly, effector CD4$^+$ T-cell entry into the vaginal tissue also required IFN-γ secretion and responsiveness to IFN-γ by the CD4$^+$ T cells themselves (Supplementary Fig. 9), suggesting an autocrine-mediated conditioning of CD4$^+$ T cells for access to the vaginal mucosa. To address whether IFN-γ secretion by CD4 T cells within the vagina is required for CTL entry, we took advantage of the fact that CD4$^+$ T-cell entry occurs before day 3 post infection, whereas CTL entry is delayed until day 5 post infection (Supplementary Fig. 1). Selective blockade of CD4-secreted IFN-γ (Supplementary Fig. 7a) on days 3 and 4 of infection still allowed effector CD4$^+$ T cells to enter the vaginal tissue (Supplementary Fig. 10). However, this treatment resulted in a significant reduction in CTL migration (Fig. 1e). These data collectively indicated that CD4$^+$ T cells that have migrated to the vagina must still produce IFN-γ to enable CTL entry to the site. Taken together, our data provide evidence that CD4$^+$ T-cell-derived IFN-γ is both necessary and sufficient in enabling CTL entry to the site of infection.

Next, the requirement for IFN-γ-inducible chemokines in CTL recruitment was assessed. To this end, congenically marked gBT-I transgenic mice were crossed to CXCR3-deficient mice incapable of responding to the IFN-γ-inducible chemokines, CXCL9 and CXCL10. Physiologically relevant numbers (10^5 (Fig. 2) or 10^3 (Supplementary Fig. 11)) of purified CD45.1$^+$ *Cxcr3*$^{-/-}$ gBT-I cells were transferred into wild-type hosts, and their proliferation and migration was measured 6 days after HSV-2 infection. *Cxcr3*$^{-/-}$ gBT-I cells migrated and proliferated normally in the draining lymph nodes of wild-type hosts (Fig. 3a). Furthermore, both wild-type and *Cxcr3*$^{-/-}$ gBT-I cells accumulated in the draining lymph nodes (Fig. 3c and Supplementary Fig. 11b), and migrated into the spleen and peripheral tissues (Fig. 3d, e and Supplementary Fig. 11c, d). In stark contrast, and despite the presence of abundant CD4$^+$ T$_h$1 cells in the vaginal mucosa (Fig. 3b), CXCR3-knockout CTLs failed to migrate into the virally infected vaginal tissue (Fig. 3f and Supplementary Fig. 11a). Moreover, the small numbers of *Cxcr3*$^{-/-}$ gBT-I cells that managed to migrate into the vaginal tissue were mostly excluded from the epithelial layer (Fig. 3g)—the primary site of virus infection and replication[10]. Therefore, these data indicated that effector CTL migration into the infected tissue requires CXCR3, whose ligand expression is induced by IFN-γ produced by CD4$^+$ T cells.

©2009 Macmillan Publishers Limited. All rights reserved

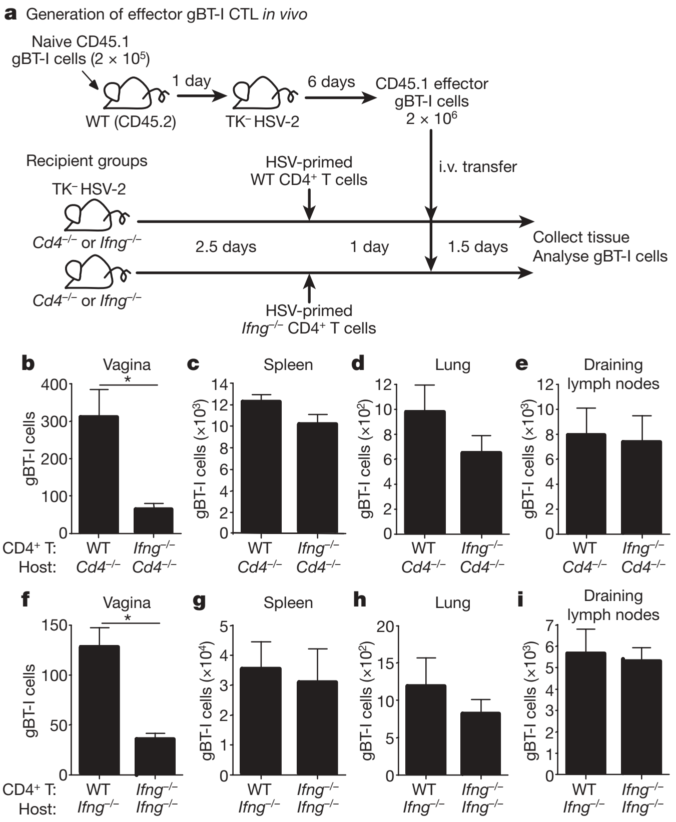


Figure 2 | CD4 T cell-secreted IFN-γ mediates CTL entry into infected vaginal tissue. **a–i**, HSV-primed wild-type or *Ifng*$^{-/-}$ CD4^{+} T cells (10^7) were adoptively transferred into *Cd4*$^{-/-}$ (**b–e**) or *Ifng*$^{-/-}$ (**f–i**) mice at 2.5 days after HSV-2 intravaginal infection. One day later, fully helped congenic effector gBT-I (2×10^6) cells were transferred into these mice. Five days after infection, the number of gBT-I cells in the indicated tissues was analysed (**b–i**). Results are mean and s.e.m. ($n = 4$) and are representative of four independent experiments. Statistics were determined by unpaired two-tailed *t*-test. $^{*}P < 0.05$, $^{**}P < 0.01$.

We next asked whether the CD4^{+} T cells directly provide help to CD8^{+} T cells, or whether CD4^{+} T cells control the recruitment of CD8^{+} T cells indirectly through modifying the local tissue environment. To address the former possibility, we examined the necessity of CD4^{+} T cells for CD8^{+} T cells to express and respond through CXCR3. To this end, gBT-I cells were primed in either wild-type or CD4-deficient hosts by intravaginal infection with TK^{-} HSV-2, and the expression of CXCR3 was assessed. Although most naive CD8^{+} T cells expressed low to undetectable levels of CXCR3, effector gBT-I cells expressed high levels of CXCR3, regardless of the absence of CD4^{+} T cells (Supplementary Fig. 12a). Furthermore, both helped and helpless CTLs responded to and migrated towards CXCL9 and CXCL10 comparably (Supplementary Fig. 12b). These data revealed that CD4 help is not required for the functional expression of CXCR3 by the effector CD8^{+} T cells. Therefore, CD4^{+} T-cell help is not provided directly to CD8^{+} T cells, but is probably mediated through modification of the local microenvironment.

In an effort to understand the mechanism by which CD4^{+} T helper cells indirectly enable migration of CD8^{+} T cells to the site of infection, we examined the production of CD4-dependent cytokines and chemokines at the local mucosa. As expected, IFN-γ levels 4 days after infection were diminished in CD4-deficient mice within the vaginal lumen (Fig. 4) and within the lamina propria (Supplementary Fig. 7b). In the absence of CD4^{+} T cells or in mice injected with IFN-γ blocking-antibody starting at day 3 after infection, secretion of CXCL9 was diminished and a significant reduction in the levels of CXCL10 was observed (Fig. 4). Thus, cumulative levels of CXCR3 ligands are significantly reduced in the vaginal tissue in the absence of CD4 or IFN-γ.

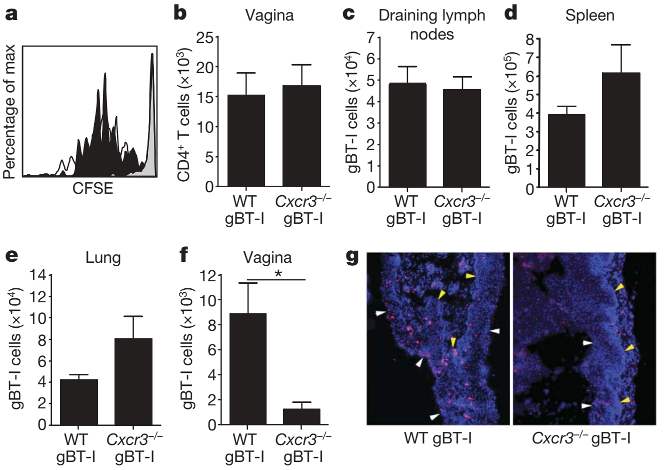


Figure 3 | CTL recruitment to the infected tissue depends on CXCR3. Recipient mice were reconstituted with 10^5 wild-type or *Cxcr3*$^{-/-}$ gBT-I cells. **a**, At 3.5 days after infection, division of wild-type (blank), *Cxcr3*$^{-/-}$ (filled), and uninfected (grey) gBT-I cells in the draining lymph nodes was analysed. CFSE, carboxyfluorescein succinimidyl ester. **b–f**, The numbers of total CD4^{+} T cells within the vagina (**b**) and gBT-I cells in the indicated tissues (**c–f**) were analysed on day 6 after infection. **g**, Vaginal tissues 7 days after infection were stained for gBT-I cells (red) and nuclei (blue). Yellow arrowheads, basement membrane; white arrowheads, vaginal lumen. Original magnification, ×10. Results in **b–f** are mean and s.e.m. ($n = 4$). Data are representative of three separate experiments. $^{*}P < 0.05$.

Because CXCL9 and CXCL10 are also induced by type I IFNs, and because we detected a reduced but significant recruitment of residual CTL to the vaginal tissue in the absence of CD4^{+} T cells or IFN-γ (Figs 1 and 2), we next assessed the importance of type I IFNs in lymphocyte recruitment after HSV-2 infection. Despite normal priming[19], CD4^{+} T-cell recruitment to the vagina was significantly reduced in the absence of IFN-αβR (Supplementary Fig. 13b). Consequently, IFN-γ, CXCL9 and CXCL10 secretion, and subsequent recruitment of CTL were significantly decreased in IFN-αβR-deficient (*Ifnar2*$^{-/-}$) mice (Supplementary Fig. 13c). Notably, neutralization of IFN-γ in *Ifnar2*$^{-/-}$ mice completely abolished CXCL9 and CXCL10 secretion and eliminated the residual CTL recruitment. These data indicated that in the absence of CD4^{+} T cells or IFN-γ, type I IFNs secreted at the site of infection lead to the production of CXCL10, leading to recruitment of a minor population of CTL to the vagina. However, CD4^{+} T cells have a dominant role in guiding effector CTLs to the site of infection. Collectively, these data indicate that CD4^{+} T cells mobilize CTLs to the sites of infection by licensing the local tissue environment—by inducing expression of chemokines necessary for CTLs to enter from systemic circulation into the site of infection. We speculate that the source of such chemokines is probably the infected vaginal epithelial cells, as high levels of messenger RNA for both CXCL9 and CXCL10 are detected within the epithelial layer by day 4 post infection (data not shown).

Our study shows that, as well as a crucial involvement in the priming of CD8^{+} T cells and in promoting memory CD8^{+} T-cell development[1], CD4^{+} T cells serve as gate-keepers of CTL entry into the infected tissues. CD4^{+} T cells orchestrate this through the secretion of IFN-γ and turning on the expression of chemokines CXCL9 and CXCL10 *in situ*, which enable the CXCR3^{+} effector CTL population to migrate from the peripheral blood into the infected areas. Both of these chemokines have been shown to be important for CTL recruitment and defence against HSV-2 infection[20]. These findings reveal that fully differentiated CTLs are self-insufficient for entry into infected peripheral tissues, and that another layer of regulation is provided by CD4^{+} T cells to carry out their effector functions. The many stages in which CD4 help is required for CTL responses—during priming, memory and now effector phases—probably reflect the need to restrict the cytotoxic activity to the site of infection. In

©2009 Macmillan Publishers Limited. All rights reserved

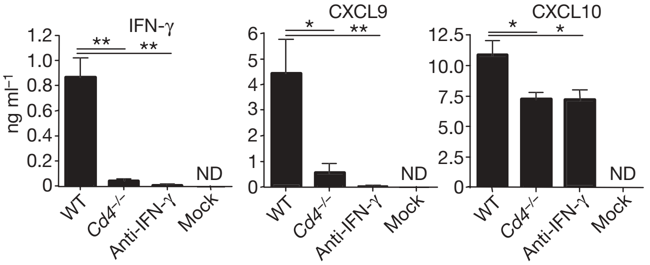


Figure 4 | Secretion of CTL-recruiting chemokines in the infected tissue depends on CD4 T-cell help. Four days after HSV-2 infection, vaginal wash was collected from wild-type, *Cd4*$^{-/-}$ mice or mice that were treated with anti-IFN-γ antibody on day 3 after infection (as in Fig. 1a). Cytokine and chemokine concentrations were determined by Luminex bead assay. Data were pooled from two independent experiments ($n = 6$) and represented as mean ± s.e.m. $*P < 0.05$, $**P < 0.01$ (one-way analysis of variance (ANOVA) followed by Tukey's post-hoc multiple comparison).

contrast, effector CTL migration into other mucosal tissues, such as the lung and intestine[15], does not require either CD4 help or inflammation. Thus, certain organs are 'permissible' for effector CTL migration, whereas other organs (for example, vagina) are 'restricted' and rely on CD4 help and inflammatory chemokines to recruit memory of effector T cells only when they are needed.

Many pathogens invade the host by establishing local infection before spreading to other organs. The recruitment of the effector lymphocytes to the initial site of microbial entry represents an important challenge in ensuring the protection of the host. Our findings indicate that vaccines that target CTL immunity must incorporate T helper epitopes to allow their migration to the site of infection once the pathogen invades the host. Furthermore, the treatment of CTL-mediated organ-specific autoimmune diseases might benefit from selective depletion of CD4^{+} T cells or IFN-γ from the local tissue, so as to block further recruitment of pathogenic CTL populations. Future studies to understand the crucial players involved in CD4^{+} T-cell-mediated recruitment of effector CTL will help our ability to design effective interventions for treatment of infectious diseases, autoimmunity and cancer.

METHODS SUMMARY

Adoptive transfer of naive CD8^{+} T cells and generation of effector gBT-I cells. Indicated numbers of naive gBT-I transgenic CD8^{+} T cells from the spleens of CD45.1^{+} gBT-I or CD45.1^{+} *Cxcr3*$^{-/-}$ gBT-I mice were adoptively transferred into recipient mice. In some experiments, the donor cells were labelled with 0.05 μM CFSE (Invitrogen) before transfer. CD4^{+} T cells from recipient mice were depleted by injection of 200 μg GK1.5 antibody (>99.5% depletion). Some recipient mice were injected with neutralizing antibody against IFN-γ (intraperitoneally, 1 mg XMG1.2). Depo-Provera-treated 6–8-week-old female recipient mice were infected intravaginally with 10^6 plaque-forming units (p.f.u.) of 186TKΔKpn (TK^{-} HSV-2)[9] or with uninfected Vero lysate (mock infection) control as described previously[10]. Effector CTLs were isolated from the spleen at 6 days post infection, and transferred into day 4 HSV-2-infected recipients. The numbers of CD45.1^{+} gBT-I cells in the primary or secondary hosts were analysed by flow cytometry.

Flow cytometry. Single suspensions were prepared from each experimental group using a modified protocol as described[21]. To analyse chemokine receptor expression, cell suspensions were stained with an anti-CXCR3 antibody (220803). The H-2K^{b}-gB498-505 tetramer used here was prepared by the National Institutes of Health (NIH) tetramer core facility. Samples were acquired on a FACSCaliber (BD Bioscience) and analysed with FlowJo software (TreeStar).

Cytokine and chemokine measurement in the vaginal wash. The vaginal wash was collected using a standard method[21]. The amount of cytokine/chemokine in the vaginal wash was measured using a multiplex Luminex beads assay (Millipore) or ELISA according to manufacturers' instructions.

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

Received 4 August; accepted 16 September 2009.
Published online 8 November 2009.

1. Williams, M. A. & Bevan, M. J. Effector and memory CTL differentiation. *Annu. Rev. Immunol.* **25,** 171–192 (2007).
2. Bennett, S. R. *et al.* Help for cytotoxic-T-cell responses is mediated by CD40 signalling. *Nature* **393,** 478–480 (1998).
3. Ridge, J. P., Di Rosa, F. & Matzinger, P. A conditioned dendritic cell can be a temporal bridge between a CD4^{+} T-helper and a T-killer cell. *Nature* **393,** 474–478 (1998).
4. Schoenberger, S. P., Toes, R. E., van der Voort, E. I., Offringa, R. & Melief, C. J. T-cell help for cytotoxic T lymphocytes is mediated by CD40–CD40L interactions. *Nature* **393,** 480–483 (1998).
5. Jennings, S. R., Bonneau, R. H., Smith, P. M., Wolcott, R. M. & Chervenak, R. CD4-positive T lymphocytes are required for the generation of the primary but not the secondary CD8-positive cytolytic T lymphocyte response to herpes simplex virus in C57BL/6 mice. *Cell. Immunol.* **133,** 234–252 (1991).
6. Janssen, E. M. *et al.* CD4^{+} T cells are required for secondary expansion and memory in CD8^{+} T lymphocytes. *Nature* **421,** 852–856 (2003).
7. Shedlock, D. J. & Shen, H. Requirement for CD4 T cell help in generating functional CD8 T cell memory. *Science* **300,** 337–339 (2003).
8. Sun, J. C. & Bevan, M. J. Defective CD8 T cell memory following acute infection without CD4 T cell help. *Science* **300,** 339–342 (2003).
9. Jones, C. A., Taylor, T. J. & Knipe, D. M. Biological properties of herpes simplex virus 2 replication-defective mutant strains in a murine nasal infection model. *Virology* **278,** 137–150 (2000).
10. Zhao, X. *et al.* Vaginal submucosal dendritic cells, but not Langerhans cells, induce protective Th1 responses to herpes simplex virus-2. *J. Exp. Med.* **197,** 153–162 (2003).
11. Smith, C. M. *et al.* Cognate CD4^{+} T cell licensing of dendritic cells in CD8^{+} T cell immunity. *Nature Immunol.* **5,** 1143–1148 (2004).
12. Mueller, S. N., Heath, W., McLain, J. D., Carbone, F. R. & Jones, C. M. Characterization of two TCR transgenic mouse lines specific for herpes simplex virus. *Immunol. Cell Biol.* **80,** 156–163 (2002).
13. Stock, A. T. *et al.* Optimization of TCR transgenic T cells for *in vivo* tracking of immune responses. *Immunol. Cell Biol.* **85,** 394–396 (2007).
14. Marshall, D. R. *et al.* Measuring the diaspora for virus-specific CD8^{+} T cells. *Proc. Natl Acad. Sci. USA* **98,** 6313–6318 (2001).
15. Masopust, D., Vezys, V., Marzo, A. L. & Lefrancois, L. Preferential localization of effector memory cells in nonlymphoid tissue. *Science* **291,** 2413–2417 (2001).
16. Lund, J. M., Hsing, L., Pham, T. T. & Rudensky, A. Y. Coordination of early protective immunity to viral infection by regulatory T cells. *Science* **320,** 1220–1224 (2008).
17. Reboldi, A. *et al.* C–C chemokine receptor 6-regulated entry of T$_{H}$-17 cells into the CNS through the choroid plexus is required for the initiation of EAE. *Nature Immunol.* **10,** 514–523 (2009).
18. Milligan, G. N. & Bernstein, D. I. Interferon-γ enhances resolution of herpes simplex virus type 2 infection of the murine genital tract. *Virology* **229,** 259–268 (1997).
19. Iijima, N. *et al.* Dendritic cells and B cells maximize mucosal Th1 memory response to herpes simplex virus. *J. Exp. Med.* **205,** 3041–3052 (2008).
20. Thapa, M., Welner, R. S., Pelayo, R. & Carr, D. J. CXCL9 and CXCL10 expression are critical for control of genital herpes simplex virus type 2 infection through mobilization of HSV-specific CTL and NK cells to the nervous system. *J. Immunol.* **180,** 1098–1106 (2008).
21. Iijima, N., Linehan, M. M., Saeland, S. & Iwasaki, A. Vaginal epithelial dendritic cells renew from bone marrow precursors. *Proc. Natl Acad. Sci. USA* **104,** 19061–19066 (2007).

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

Acknowledgements We thank F. R. Carbone and W. R. Heath for the gBT-I transgenic mouse, A. Rudensky for the foxp3 knock-in mouse, R. Medzhitov and J. M. Thompson for critical reading of the manuscript, and N. Iijima for technical assistance. This work is supported by NIH grants to A.I. (AI054359 and AI062428) and to C.G. (AI39759 and HL51366). Y.N. was a Japan Society for the Promotion of Science fellow. A.I. is a recipient of the Burroughs Wellcome Investigators in Pathogenesis of Infectious Disease.

Author Contributions Experiments were conceived and designed by Y.N. and A.I. Experiments were performed by Y.N. Data were analysed by Y.N. and A.I. The paper was written by Y.N. and A.I. C.G. and B.L. provided CXCR3-knockout mice and discussed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.I. (akiko.iwasaki@yale.edu).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Mice. C57BL/6 mice (B6), Congenic C57/BL6/SJL-Prprc^{a}Pep3^b/BoyJ mice (B6 CD45.1), *Ifngr*$^{-/-}$, *Ifnar2*$^{-/-}$, *Ifng*$^{-/-}$, *Cd8a*$^{-/-}$ and *Cd4*$^{-/-}$ mice were purchased from National Cancer Institute and Jackson Laboratories. The gBT-I TCR transgenic mice[12] that are specific for the immunodominant HSV glycoprotein B (gB) peptide gB$_{498-505}$ were provided by F. R. Carbone and W. R. Heath. *Cxcr3*$^{-/-}$ mice have been described previously[22]. The CD45.1$^+$ gBT-I mice and CD45.1$^+$ *Cxcr3*$^{-/-}$ gBT-I mice were bred in our facility. All procedures used in this study complied with federal guidelines and institutional policies of the Yale Animal Care and Use Committee.

Adoptive transfer of naive CD8$^+$ T cells and virus infection. Spleen was collected from the CD45.1$^+$gBT-I or CD45.1$^+$ *Cxcr3*$^{-/-}$ gBT-I mice. The naive gBT-I transgenic CD8$^+$ T cells were purified by magnetic cell sorting using CD8$^+$ T cell isolation kit (>96%) (Miltenyi Biotec). In some experiments, the donor cells were labelled with 0.05 μM CFSE (Invitrogen) for 10 min at 37 °C before transfer. The donor cells were adoptively transferred into recipient B6 mice or CD4-deficient mice intravenously. To deplete CD4$^+$ T cells in recipient mice, B6 mice were injected intraperitoneally with 200 μg GK1.5 antibody as indicated. The CD4 depletion was confirmed by FACS analysis (>99.5% depletion). For NK cell depletion, purified anti-mouse NK1.1 (PK136) was given by an intraperitoneal injection of 200 μg antibody in PBS 1 day before and 1 day after infection. Complete NK depletion was confirmed by FACS analysis in all treated groups. Some recipient mice were injected with neutralizing antibody against IFN-γ (intraperitoneally, 1 mg XMG1.2). Thymidine kinase mutant HSV-2 strain 186TKΔKpn (TK$^-$ HSV-2)[9] was used to infect mice. Depo-Provera-treated 6–8-week-old female mice were infected intravaginally with 10^6 p.f.u. of TK$^-$ HSV-2 or with uninfected Vero lysate (mock infection) as described previously[10].

Flow cytometry and generation of effector gBT-I cells. Single suspensions were prepared from each experimental group using a modified protocol as described[21]. In brief, spleen, lung and draining lymph nodes (inguinal and iliac) were digested with collagenase D (Roche). For vagina, the vaginal tube in its entirety was collected from each animal, and tissues were treated with Dispase II for 15 min before collagenase digestion. Total cell numbers of the single cells obtained from the entire vaginal tissue per animal were counted by haemocytometer. The frequency of congenically marked gBT-I/CD45.1 cells was analysed after staining with anti-CD8α (Ly-2, 53-6.7) and anti-CD45.1 (A20). The total numbers of gBT-I per vaginal tissue were calculated by multiplying the frequency of CD8α$^+$ CD45.1$^+$ cells within the lymphocyte forward scatter (FSC) versus side scatter (SSC) gate (>99% TCR Vα2$^+$)[12], and the number of total cells obtained per vaginal tissue. To analyse chemokine receptor, cell suspensions were stained with an anti-CXCR3 antibody (clone 220803; R&D Systems).

The H-2K^b-gB498-505 tetramer used here was prepared by the NIH tetramer core facility. Samples were acquired on a FACSCaliber (BD Bioscience) and analysed with FlowJo software (TreeStar). FACS sorting was performed on a FACSAria machine. Single-cell suspensions of splenocytes were incubated with biotin-conjugated anti-CD45.1 antibody, and then incubated with anti-biotin beads (Miltenyi Biotec). Positive-selected cells (enriched CD45.1$^+$ cells >85%) were stained with antibodies against CD45.1 and CD8α.

Immunofluorescence staining. For analysis of the localization of HSV-specific CD8$^+$ T cells, frozen sections of the vaginal tissue were stained with anti-CD45.1 antibodies using a procedure similar to that described previously[23]. Slides were examined using a BX51 fluorescence microscope equipped with ×10 Plan objective and a digital camera (Olympus) and images were processed with PictureFrame imaging software (Optronics).

Preparation of HSV-primed CD4$^+$ T cells and IFN-γ production. To prepare HSV-primed CD4$^+$ T cells, wild-type and *Ifng*$^{-/-}$ or *Ifngr*$^{-/-}$ mice were infected with TK$^-$ HSV-2 as described earlier. Six days after infection, draining lymph nodes and spleen were collected and CD4$^+$ T cells were isolated using CD4$^+$ T cells isolation kit (Miltenyi Biotec). Foxp3$^-$ effector CD4$^+$ T cells were prepared from HSV-infected Foxp3^{gfp} knock-in mice[24]. Five-million effector HSV-primed total or Foxp3$^-$ CD4$^+$ T cells were transferred into *Cd4*$^{-/-}$ mice at day 2.5 post infection; 2×10^6 effector gBT-I cells were adoptively transferred into the same hosts on day 3.5 post infection. On day 5 post infection, tissues were collected and analysed for the presence of gBT-I cells.

To determine the effector functions of T$_h$1 cells, CD4$^+$ T cells (10^5) were stimulated for 68 h *in vitro* with syngeneic splenocytes as antigen presenting cells (APCs; 2×10^5) in the presence of heat-inactivated viral antigens[10]. The levels of IFN-γ were determined by standard ELISA method.

Cytokine and chemokine measurements in the vaginal wash and lamina propria. The vaginal wash was collected using standard methods[21]. The amount of cytokine/chemokine in the vaginal wash was measured using a multiplex Luminex beads assay (Millipore) or ELISA (eBioscience) according to manufacturers' instructions. Vaginal tissues were homogenized using scissors and suspended in 500 μl PBS. Supernatants were clarified (5,600*g*, 5 min) and assessed for IFN-γ levels by ELISA.

Chemotaxis assay. Chemotaxis assays were performed as described[25]. In brief, chemokine dilutions were added to the bottom well of a 96-well chemotaxis plate (NeuroProbe). FACS-purified CD8/Ly5.2 cells were added on the top of the membrane (4×10^4) and allowed to migrate at 37 °C for 4 h. Cells in the bottom wells were counted under a microscope and the total cell numbers were determined. Each experiment was performed in duplicate and counted a minimum of two times.

Statistical analysis. Normally distributed continuous variable comparisons were performed using two-tailed unpaired *t*-test and one-way ANOVA followed by Tukey's post test comparison using Prism software.

22. Hancock, W. W. *et al.* Requirement of the chemokine receptor CXCR3 for acute allograft rejection. *J. Exp. Med.* **192,** 1515–1520 (2000).
23. Sato, A. & Iwasaki, A. Induction of antiviral immunity requires Toll-like receptor signaling in both stromal and dendritic cell compartments. *Proc. Natl Acad. Sci. USA* **101,** 16274–16279 (2004).
24. Fontenot, J. D. *et al.* Regulatory T cell lineage specification by the forkhead transcription factor Foxp3. *Immunity* **22,** 329–341 (2005).
25. Iwasaki, A. & Kelsall, B. L. Localization of distinct Peyer's patch dendritic cell subsets and their recruitment by chemokines macrophage inflammatory protein (MIP)-3α, MIP-3β, and secondary lymphoid organ chemokine. *J. Exp. Med.* **191,** 1381–1394 (2000).

©2009 Macmillan Publishers Limited. All rights reserved

Host plant genome overcomes the lack of a bacterial gene for symbiotic nitrogen fixation

Tsuneo Hakoyama[1,2], Kaori Niimi[2], Hirokazu Watanabe[2], Ryohei Tabata[2], Junichi Matsubara[2], Shusei Sato[3], Yasukazu Nakamura[3], Satoshi Tabata[3], Li Jichun[4], Tsuyoshi Matsumoto[4], Kazuyuki Tatsumi[4], Mika Nomura[5], Shigeyuki Tajima[5], Masumi Ishizaka[6], Koji Yano[1], Haruko Imaizumi-Anraku[1], Masayoshi Kawaguchi[7], Hiroshi Kouchi[1] & Norio Suganuma[2]

Homocitrate is a component of the iron–molybdenum cofactor in nitrogenase, where nitrogen fixation occurs[1,2]. *NifV*, which encodes homocitrate synthase (HCS)[3], has been identified from various diazotrophs but is not present in most rhizobial species that perform efficient nitrogen fixation only in symbiotic association with legumes. Here we show that the *FEN1* gene of a model legume, *Lotus japonicus*, overcomes the lack of *NifV* in rhizobia for symbiotic nitrogen fixation. A Fix^- (non-fixing) plant mutant, *fen1*, forms morphologically normal but ineffective nodules[4,5]. The causal gene, *FEN1*, was shown to encode HCS by its ability to complement a HCS-defective mutant of *Saccharomyces cerevisiae*. Homocitrate was present abundantly in wild-type nodules but was absent from ineffective *fen1* nodules. Inoculation with *Mesorhizobium loti* carrying *FEN1* or *Azotobacter vinelandii NifV* rescued the defect in nitrogen-fixing activity of the *fen1* nodules. Exogenous supply of homocitrate also recovered the nitrogen-fixing activity of the *fen1* nodules through *de novo* nitrogenase synthesis in the rhizobial bacteroids. These results indicate that homocitrate derived from the host plant cells is essential for the efficient and continuing synthesis of the nitrogenase system in endosymbionts, and thus provide a molecular basis for the complementary and indispensable partnership between legumes and rhizobia in symbiotic nitrogen fixation.

The major source of nitrogen for all living organisms is atmospheric dinitrogen, which is fixed mainly by microorganisms that have an ability to reduce dinitrogen to ammonium ions by a nitrogenase system. In leguminous plants, soil bacteria of the family Rhizobiaceae (rhizobia) are hosted within a symbiotic organ, the root nodule, in which the endosymbiotic rhizobia are able to fix dinitrogen. This enables the host legumes to grow without an exogenous nitrogen source. Unlike many free-living diazotrophs, rhizobia are able to exhibit highly efficient nitrogen fixation only when they are in the host nodule cells as an endosymbiotic form, the bacteroid. This indicates that rhizobial nitrogen fixation is controlled by the host plant. Fix^- mutants of the host legumes, which form ineffective nodules, are key tools in the identification of the host genes essential for the establishment of symbiotic nitrogen fixation.

A *L. japonicus* Fix^- mutant, *fen1* (refs 4, 5), forms small, pale pink nodules and shows nitrogen deficiency symptoms under symbiotic conditions (Supplementary Fig. 1a–c, f–h). In the *fen1* nodules, rhizobial invasion of the nodule cells seemed to be comparable to that in wild-type plants (Supplementary Fig. 1d, e), but the nitrogenase activity remained very low (Supplementary Fig. 1i, j).

We identified the gene, *FEN1*, responsible for the *fen1* mutant through map-based cloning, and confirmed the complementation of the mutant phenotypes by *Agrobacterium rhizogenes*-mediated hairy-root transformation (Supplementary Information and Supplementary Fig. 2). Transcripts of *FEN1* were detected only in nodules, indicating that expression of *FEN1* was regulated in a nodule-specific manner (Fig. 1a). When the *FEN1* promoter–β-glucuronidase (GUS) fusion was introduced, GUS activity was detected only in infected cells of nodules (Fig. 1b, c). By searching the *L. japonicus* EST database[6], we found a clone, MWM049f12, paralogous to *FEN1* with a predicted amino-acid sequence that was 91% identical to that of *FEN1*. However, expression of MWM049f12 was detected in all organs of *L. japonicus* at low levels and was not enhanced in nodules. These results indicate that FEN1 is closely associated with the nitrogen-fixing activity of the nodules.

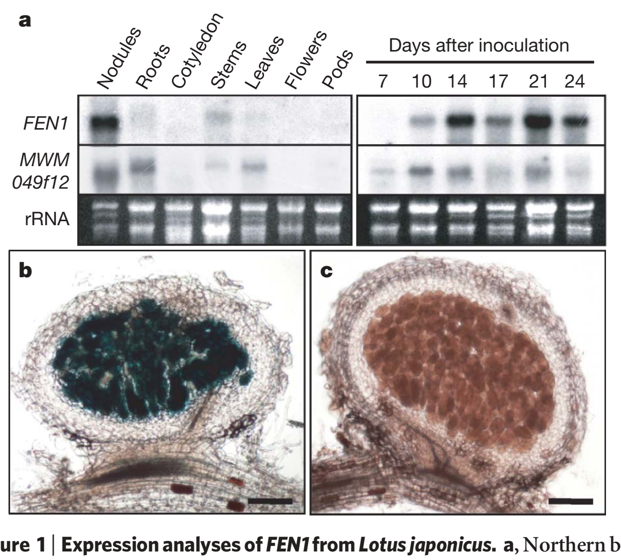


Figure 1 | Expression analyses of *FEN1* from *Lotus japonicus*. a, Northern blot analysis of the *FEN1* transcript and the homologous *MWM049f12* transcript in various organs (left) and during nodule development (right). Ribosomal RNA was stained with ethidium bromide. **b**, **c**, Spatial expression analysis of the *FEN1* gene in the wild-type nodules by *FEN1* promoter-GUS fusion. Transgenic hairy roots were inoculated with *Mesorhizobium loti* TONO, and the sections of nodules containing the *FEN1* promoter–GUS fusion (**b**) and empty vector (**c**) were examined with a histochemical GUS assay. Positive signals are visible as blue colour development. Scale bars, 200 μm.

[1]National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan. [2]Department of Life Science, Aichi University of Education, Kariya, Aichi 448-8542, Japan. [3]Kazusa DNA Research Institute, Kisarazu, Chiba 292-0812, Japan. [4]Research Center for Materials Science and Department of Chemistry, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan. [5]Faculty of Agriculture, Kagawa University, Kita, Kagawa 761-0795, Japan. [6]National Institute of Agro-Environmental Sciences, Tsukuba, Ibaraki 305-8604, Japan. [7]National Institute for Basic Biology, Okazaki, Aichi 444-8585, Japan.

©2009 Macmillan Publishers Limited. All rights reserved

The predicted FEN1 protein consisted of 540 amino-acid residues with a molecular mass of 58,600 kDa. No signal peptide sequences were found, suggesting that it is a cytosolic protein. The deduced amino-acid sequence of FEN1 had 71% identity to that for the *Glycine max* nodule-specific gene, *GmN56*. The introduction of *GmN56* complementary DNA into the *fen1* mutant recovered the growth and nitrogenase activity of the mutant (Supplementary Fig. 3a, b), indicating that *GmN56* is an orthologue of *FEN1*. *GmN56* has been shown to be induced with the onset of nitrogen fixation, and the transcripts are localized in the bacterium-infected cells of mature nodules of soybean[7], which is consistent with the expression pattern of *FEN1* in *L. japonicus* nodules. The predicted GmN56 protein showed homology to 2-isopropylmalate synthase (IPMS) and HCS, although the exact function of GmN56 has not been confirmed. Besides GmN56, amino-acid sequences deduced from several genes encoding IPMS isolated from plants such as *Brassica atlantica*, *Arabidopsis thaliana* and *Lycopersicon pennellii* were found to have high similarity (about 66% identity) to FEN1.

To explore the function of FEN1, we first introduced *FEN1* into an IPMS-defective mutant of *S. cerevisiae* [8]. *FEN1* failed to complement leucine auxotrophy of the *S. cerevisiae* mutant, and the cell extract showed no IPMS activity (Supplementary Fig. 4a, b). By contrast, we detected both substantial activity of IPMS in the extract of *S. cerevisiae* transformed with the *Arabidopsis* gene *IPMS2* (*AtIPMS2*; At1g74040)[9] and also complementation of leucine auxotrophy, even though only in part (see the legend to Supplementary Fig. 4). In addition, the Fix$^-$ phenotype of *fen1* was not recovered by introduction of *AtIPMS2* (Supplementary Fig. 3a, b). From these results we concluded that *FEN1* does not code for IPMS.

We next focused on HCS, which catalyses the synthesis of homocitrate from 2-oxoglutarate and acetyl-CoA. IPMS and HCS are different enzymes but they have some structural similarity[7]. They both catalyse similar reactions: the transfer of an acyl group from acetyl-CoA to a 2-oxo acid to generate the alkyl group on the 2-oxo acid. The FEN1 protein has 36% identity to HCS (NIFV) of the nitrogen-fixing aerobic bacterium *Azotobacter vinelandii*[3]. We introduced *FEN1* into a *S. cerevisiae* mutant that showed lysine auxotrophy caused by the lack of HCS[10]. The introduction of *FEN1*, but not *Arabidopsis IPMS2* and mutated *FEN1*, complemented lysine auxotrophy in the mutant (Fig. 2a). Furthermore, a significant accumulation of homocitrate was found in the transformed *S. cerevisiae* mutant when expression of *FEN1* was induced (Supplementary Fig. 5). These results showed that the recombinant FEN1 protein confers HCS activity. In this study we were unable to detect HCS activity *in vitro* in cell-free extracts of *Lotus* nodules. We therefore investigated the presence of homocitrate in various tissues of *L. japonicus* to confirm HCS activity *in vivo*. Analysis by liquid chromatography coupled with tandem mass spectrometry (LC–MS–MS) showed that, in wild-type Gifu plants, homocitrate was detected abundantly in nodules, but in neither roots nor shoots (Fig. 2b). By contrast, it was barely detectable (less than 1% of wild-type nodules) in ineffective nodules formed on the *fen1* mutant (Fig. 2c). The ineffective nodules formed by the *NifH*-defective mutant of *M. loti* contained homocitrate at a level comparable with that in the wild-type nodules, indicating that accumulation of homocitrate in nodules is not the result of active nitrogen fixation. In addition, the level of 2-oxoglutarate was found to be higher in the *fen1* nodules than in the wild-type nodules and in ineffective nodules formed by the *NifH*-defective mutant of *M. loti* (Fig. 2c). These results indicate that *FEN1* encodes HCS and that the activity is lost in nodules of the *fen1* mutant.

In higher plants, no metabolic pathway leading to the synthesis of lysine through homocitrate as an intermediate has been identified. Here we noted that homocitrate is a component of the iron–molybdenum cofactor (FeMo cofactor) of the nitrogenase complex in nitrogen-fixing bacteria[2]. We therefore expected that homocitrate synthesized in host plant cells would be transported to bacteroids and used in the biosynthesis of the nitrogenase complex. We examined this hypothesis by introducing *FEN1* into *M. loti* under the control of the rhizobial *NifH* promoter. Inoculation with *M. loti* carrying *FEN1* to the *fen1* mutant rescued either the defect in nodule nitrogenase activity or the growth of the plant (Fig. 3a, b). Expression of FEN1 in the bacteroids of nodules formed by transformed *M. loti* was confirmed by immunological detection of a FEN1–Myc fusion protein (Fig. 3c). Similarly, we tested inoculation with *M. loti* carrying *NifV* from *A. vinelandii*, which has been shown to encode HCS and to be essential for nitrogenase activity[3]. *M. loti* expressing *A. vinelandii NifV* could also rescue the *fen1* mutant phenotypes (Fig. 3d–f). Furthermore, we found that the addition of homocitrate to the culture solution partly restored the effectiveness of the nodules formed on the *fen1* mutant (Fig. 4a, b). In the *fen1* nodules supplied with homocitrate, nitrogenase proteins were recovered in amounts that were nearly comparable to those in the wild-type nodules (Fig. 4c), indicating that the restoration of nitrogenase activity by the supply of homocitrate was due to *de novo* nitrogenase biosynthesis. Taken together, these results indicate that rhizobial nitrogen-fixing activity depends on the homocitrate derived from the host plant, which could be used for assembly of the FeMo cofactor in the nitrogenase complex in the endosymbionts (Supplementary Fig. 6).

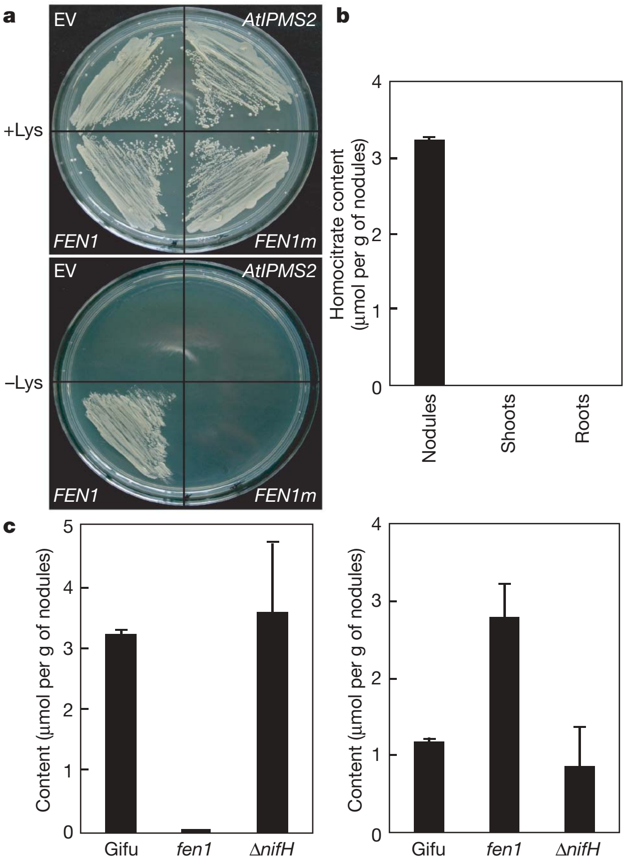


Figure 2 | Functional complementation of a *Saccharomyces cerevisiae* homocitrate-synthase defective mutant, and homocitrate and 2-oxoglutarate content in nodules. **a**, Complementation test of *S. cerevisiae* mutant by *FEN1* from *Lotus japonicus*. *S. cerevisiae* mutants carrying empty vector (EV), *FEN1*, mutated *FEN1* (*FEN1m*, corresponding to a mutation in *fen1-1*) and *Arabidopsis* isopropylmalate synthase (*AtIPMS2*) were grown on medium with (+) or without (−) lysine. **b**, Homocitrate content in nodules, shoots and roots of wild-type Gifu inoculated with *Mesorhizobium loti* TONO. **c**, Homocitrate (left) and 2-oxoglutarate (right) content in nodules of wild-type Gifu and *fen1-1* mutant inoculated with *M. loti* TONO, and in nodules induced by the *M. loti NifH* mutant (Δ*nifH*). Data are means and s.e.m. for two independent nodulated roots.

Rhizobial nitrogen-fixing activity is restricted to symbiotic bacteroids, and free-living rhizobia normally fix no atmospheric nitrogen; this

©2009 Macmillan Publishers Limited. All rights reserved

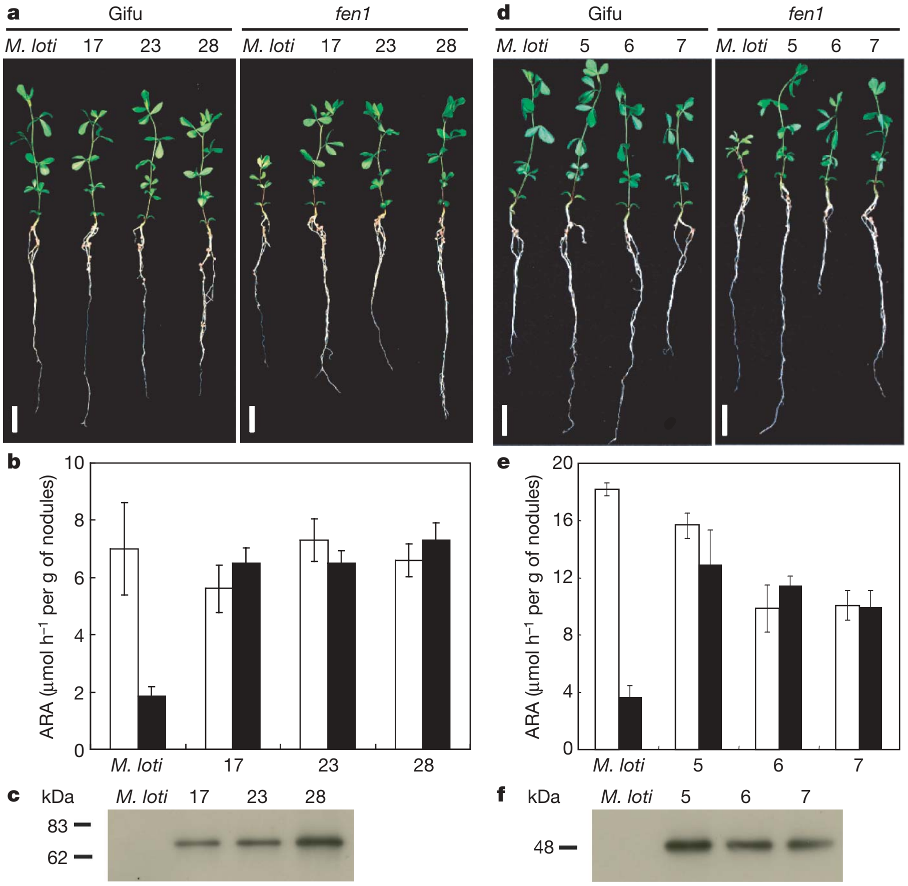


Figure 3 | Complementation of *fen1* mutants by inoculation with *Mesorhizobium loti* carrying *FEN1* and *Azotobacter vinelandii NifV*. **a**, **d**, Plants inoculated with *M. loti* TONO carrying *FEN1* (**a**) and *A. vinelandii NifV* (**d**) grown in nitrogen-free medium. Wild-type Gifu and the *fen1-1* mutant were each inoculated with three independent transformants (nos 17, 23 and 28 for *FEN1*, and nos 5, 6 and 7 for *A. vinelandii NifV*). Scale bars, 10 mm. **b**, **e**, Acetylene reduction activity (ARA) of nodules formed by inoculation with *M. loti* carrying *FEN1* (**b**) and *A. vinelandii NifV* (**e**) on wild-type Gifu (open bars) and the *fen1-1* mutant (filled bars). Data are means and s.e.m. for 12 plants. **c**, **f**, Detection of FEN1–Myc (**c**) and *A. vinelandii* NIFV–Myc (**f**) proteins in bacteroids isolated from the nodules formed by inoculation with *M. loti* transformants.

is a unique feature of the legume–*Rhizobium* symbiosis[11]. This could be explained in part by the fact that the *NifV* gene encoding HCS[3] has not been identified in most rhizobia, except in the stem-nodulating *Azorhizobium caulinodans* and photosynthetic *Bradyrhizobium* sp. rhizobial strains (RhizoBase; http://genome.kazusa.or.jp/rhizobase/). *Azorhizobium* has been shown to fix atmospheric nitrogen in culture[12]. In addition, the photosynthetic rhizobia have distinctive features, including the absence of nodulation genes, and have been proposed to belong to a distinct group in the Rhizobiaceae[13,14]. In other nitrogen-fixing symbiotic associations, however, *NifV* was identified in three types of microsymbiont: *Frankia*, *Anabaena* and some endophytic bacteria (see RhizoBase), which are capable of fixing nitrogen in their free-living state[15]. Our results, together with these previous observations, led us to the idea that the absence of *NifV* from rhizobia is compensated for by *FEN1* in the host legume's genome permitting them to acquire highly efficient nitrogen fixation in symbiosis. Nevertheless, some strains of rhizobia can fix nitrogen under defined conditions[16–18] in the free-living state, although not efficiently in most cases, and ineffective nodules induced on the *fen1* mutant showed a low rate of nitrogen-fixing activity. Rhizobia are therefore probably able to synthesize homocitrate by one or more alternative pathways without *NifV*. Alternatively, citrate may be substituted in part for homocitrate in the FeMo cofactor, as reported for *Klebsiella pneumoniae*[19].

Here we have found a gene encoding HCS in the higher plant kingdom and shown that a supply of homocitrate from the host plant cells to endosymbiotic bacteroids is essential for symbiotic nitrogen fixation. The FEN1 protein exhibited a higher structural similarity to

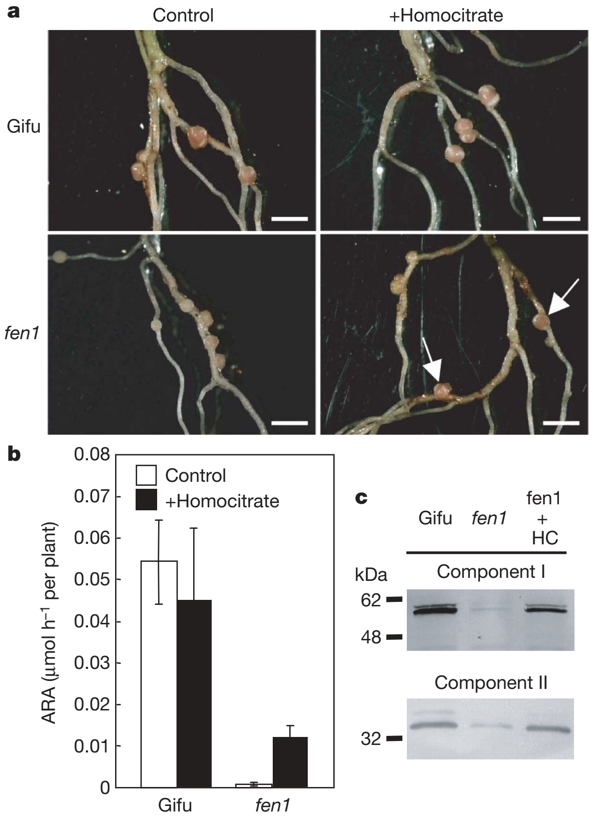


Figure 4 | Effect of supplying homocitrate to the *fen1* mutant. **a**, Nodules formed on wild-type Gifu and *fen1-1* roots after 4 days of incubation in culture medium supplemented with 1 mM homocitrate. After 4 days in culture, some of the *fen1-1* nodules showed a red coloration (indicated by arrows). Scale bars, 2 mm. **b**, Acetylene reduction activity (ARA) of nodulated roots of wild-type Gifu and the *fen1-1* mutant 4 days after supplementation with homocitrate. Data are means and s.e.m. for five independent plants. **c**, Detection of nitrogenase components I and II in bacteroids isolated from nodules of wild-type Gifu and from *fen1-1* nodules supplied with homocitrate (*fen1* + HC).

©2009 Macmillan Publishers Limited. All rights reserved

plant IPMSs rather than bacterial HCSs, but this could be reflected in the species difference between higher plants and microorganisms. Indeed, FEN1 had only about 40% similarity to bacterial IPMSs such as the LEUA of *Escherichia coli*. In addition, IPMS has been shown to be capable of using various 2-oxo acids as substrates[9,20,21]. It is thus very likely that *FEN1* was recruited from a housekeeping gene encoding IPMS during the evolution of symbiosis, and this made it possible to establish efficient nitrogen fixation by endosymbiotic bacteria. Such recruitment has been also suggested for several nodule-specific (nodulin) genes such as leghaemoglobins[22], uricase[23] and phosphoenolpyruvate carboxylase[24]. Furthermore, metabolic partnerships between host legumes and microsymbionts have been well documented. A supply of dicarboxylates and amino acids from the host cells to bacteroids has been shown to be essential for nitrogen fixation and/or differentiation of the bacteroids[25,26]. However, our finding differs from previous studies in two aspects: first, FEN1 has developed a function as a HCS, which is distinct from IPMS, and has not been found so far in higher plants; second, FEN1 produces homocitrate, which is an essential component of the nitrogenase complex but is not itself required for plant metabolism and thus could compensate for the lack of *NifV* in rhizobia. Our data support the idea that the acquisition of HCS by the nodule-specific gene *FEN1* in host legumes was one of the key genetic inventions in the establishment of a highly efficient nitrogen-fixing symbiosis by legumes and rhizobia, thus providing an insight into the co-evolution of metabolic pathways in two symbiotic partners.

METHODS SUMMARY

The plant-determined Fix^{-} mutants *fen1-1* (refs 4, 5) and *fen1-2* were derived from *L. japonicus* accessions B-129 (Gifu) and MG-20 (Miyakojima), respectively, by mutagenesis with ethylmethane sulphonate. *M. loti* strain MAFF303099 and TONO and the *NifH*-defective mutant of MAFF303099 were used for inoculation. Map-based cloning of *FEN1* was performed crossing the *fen1-1* mutant with MG-20, using SSR and dCAPS markers[27,28], together with additionally developed PCR-based markers. Because two *fen1* mutant alleles, *fen1-1* and *fen1-2*, showed essentially the same phenotypes, the *fen1-1* mutant was used in all further analyses. The *FEN1*, *GmN56* and *AtIPMS2* constructs (p*FEN1*–cDNA–t*FEN1*) were introduced into the *fen1* mutant by *A. rhizogenes*-mediated hairy-root transformation. Functional complementation of *S. cerevisiae* mutants was performed with a pYES2 yeast expression vector containing a GAL1 promoter (Invitrogen). Carboxylic acid fractions were prepared from nodules, roots and shoots with an ion-exchange resin. 2-Oxoglutarate and homocitrate were quantified by anion-exclusion high-performance liquid chromatography (HPLC) and LC–MS–MS, respectively. Authentic homocitrate standard was synthesized as described[29]. The p*NifH*–*FEN1*–3 × *myc* or p*NifH*-*A. vinelandii NifV*-3 × *myc* fragment was inserted into the transposon plasmid, pCAM120, followed by transfection into *M. loti* TONO by tri-parental mating; these transformed *M. loti* were successively inoculated into the wild-type and the *fen1* mutant plants. Synthetic homocitrate was supplied to the *fen1* mutant by immersing the nodulated roots in culture solution containing 1 mM homocitrate.

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

Received 24 September; accepted 19 October 2009.

1. Hoover, T. R. *et al.* Identification of the V factor needed for synthesis of the iron-molybdenum cofactor of nitrogenase as homocitrate. *Nature* **329,** 855–857 (1987).
2. Hoover, T. R., Imperial, J., Ludden, P. W. & Shah, V. K. Homocitrate is a component of the iron-molybdenum cofactor of nitrogenase. *Biochemistry* **28,** 2768–2771 (1989).
3. Zheng, L., White, R. H. & Dean, D. R. Purification of the *Azotobacter vinelandii nifV*-encoded homocitrate synthase. *J. Bacteriol.* **179,** 5963–5966 (1997).
4. Imaizumi-Anraku, H. *et al.* Two ineffective-nodulating mutants of *Lotus japonicus*—different phenotypes caused by the blockage of endocytotic bacterial release and nodule maturation. *Plant Cell Physiol.* **38,** 871–881 (1997).
5. Kawaguchi, M. *et al.* Root, root hair, and symbiotic mutants of the model legume *Lotus japonicus. Mol. Plant Microbe Interact.* **15,** 17–26 (2002).
6. Asamizu, E., Nakamura, Y., Sato, S. & Tabata, S. Characteristics of the *Lotus japonicus* gene repertoire deduced from large-scale expressed sequence tag (EST) analysis. *Plant Mol. Biol.* **54,** 405–414 (2004).
7. Kouchi, H. & Hata, S. GmN56, a novel nodule-specific cDNA from soybean root nodules encodes a protein homologous to isopropylmalate synthase and homocitrate synthase. *Mol. Plant Microbe Interact.* **8,** 172–176 (1995).
8. Casalone, E., Barberio, C., Cavalieri, D. & Polsinelli, M. Identification by functional analysis of the gene encoding α-isopropylmalate synthase II (*LEU9*) in *Saccharomyces cerevisiae. Yeast* **16,** 539–545 (2000).
9. De Kraker, J. W. *et al.* Two *Arabidopsis* genes (*IPMS1* and *IPMS2*) encode isopropylmalate synthase, the branchpoint step in the biosynthesis of leucine. *Plant Physiol.* **143,** 970–986 (2007).
10. Feller, A., Ramos, F., Piérard, A. & Dubois, E. In *Saccharomyces cerevisiae,* feedback inhibition of homocitrate synthase isoenzymes by lysine modulates the activation of *LYS* gene expression by Lys14p. *Eur. J. Biochem.* **261,** 163–170 (1999).
11. Kneip, C., Lockhart, P., Voss, C. & Maier, U. G. Nitrogen fixation in eukaryotes—new models for symbiosis. *BMC Evol. Biol.* **7,** 55–66 (2007).
12. Dreyfus, B. L., Elmerich, C. & Dommergues, Y. R. Free-living *Rhizobium* strain able to grow on N_2 as the sole nitrogen source. *Appl. Environ. Microbiol.* **45,** 711–713 (1983).
13. Giraud, E. & Fleischman, D. Nitrogen-fixing symbiosis between photosynthetic bacteria and legumes. *Photosynth. Res.* **82,** 115–130 (2004).
14. Giraud, E. *et al.* Legumes symbioses: absence of *nod* genes in photosynthetic bradyrhizobia. *Science* **316,** 1307–1312 (2007).
15. Gauthier, D., Diem, H. G. & Dommergues, Y. *In vitro* nitrogen fixation by two actinomycete strains isolated from *Casuarina* nodules. *Appl. Environ. Microbiol.* **41,** 306–308 (1981).
16. Pagan, J. D., Child, J. J., Scowcroft, W. R. & Gibson, A. H. Nitrogen fixation by *Rhizobium* cultured on a defined medium. *Nature* **256,** 406–407 (1975).
17. Kurz, W. G. W. & LaRue, T. A. Nitrogenase activity in rhizobia in absence of plant host. *Nature* **256,** 407–409 (1975).
18. McComb, J. A., Elliott, J. & Dilworth, M. J. Acetylene reduction by *Rhizobium* in pure culture. *Nature* **256,** 409–410 (1975).
19. Hoover, T. R. *et al.* Dinitrogenase with altered substrate specificity results from the use of homocitrate analogues for *in vitro* synthesis of the iron-molybdenum cofactor. *Biochemistry* **27,** 3647–3652 (1988).
20. Kohlhaw, G. & Leary, T. R. α-Isopropylmalate synthase from *Salmonella typhimurium. J. Biol. Chem.* **244,** 2218–2225 (1969).
21. Ulm, E. H., Böhme, R. & Kohlhaw, G. α-Isopropylmalate synthase from yeast: purification, kinetic studies, and effect of ligands on stability. *J. Bacteriol.* **110,** 1118–1126 (1972).
22. Jacobsent-Lyon, K. *et al.* Symbiotic and nonsymbiotic hemoglobin genes of *Casuarina glauca. Plant Cell* **7,** 213–223 (1995).
23. Takane, K., Tajima, S. & Kouchi, H. Two distinct uricase II (nodulin 35) genes are differentially expressed in soybean plants. *Mol. Plant Microbe Interact.* **6,** 735–741 (1997).
24. Hata, S., Izui, K. & Kouchi, H. Expression of a soybean nodule-enhanced phosphoenolpyruvate carboxylase gene that shows striking similarity to another gene for a house-keeping isoform. *Plant J.* **13,** 267–273 (1998).
25. Ronson, C. W., Lyttleton, P. & Robertson, J. G. C_4-dicarboxylate transport mutants of *Rhizobium trifolii* form ineffective nodules on *Trifolium repens. Proc. Natl Acad. Sci. USA* **78,** 4284–4288 (1981).
26. Prell, J. *et al.* Legumes regulate *Rhizobium* bacteroid development and persistence by the supply of branched-chain amino acid. *Proc. Natl Acad. Sci. USA* **106,** 12477–12482 (2009).
27. Sato, S. *et al.* Genome structure of the legume, *Lotus japonicus. DNA Res.* **15,** 227–239 (2008).
28. Suganuma, N. *et al.* The *Lotus japonicus Sen1* gene controls rhizobial differentiation into nitrogen-fixing bacteroids in nodules. *Mol. Genet. Genomics* **269,** 312–320 (2003).
29. Xu, P. F., Matsumoto, T., Ohki, Y. & Tatsumi, K. A facile method for synthesis of (*R*)-(−)- and (*S*)-(+)-homocitric acid lactones and related α-hydroxy dicarboxylic acids from D- or L-malic acid. *Tetrahedr. Lett.* **46,** 3815–3818 (2005).

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

Acknowledgements We thank E. Casalone and E. Dubois for providing *Saccharomyces cerevisiae* mutants; T. Bisseling for providing nitrogenase antibodies; Y. Kawaharada and H. Mitsui for technical help; and R. W. Ridge for critical reading of the manuscript. This work was supported by the Special Coordination Funds for Promoting Science and Technology of the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Author Contributions All authors contributed extensively to the experimental work. The manuscript was written by T.H., H.K. and N.S.

Author Information The sequences have been deposited at the DNA Data Bank of Japan with the accession numbers AP004466 (LjT09C23), AP010267 (LjT02F04), AP010268 (LjT28C03) and AB494481 (mRNA sequence (Gifu B-129) of *FEN1*). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.S. (nsuganum@auecc.aichi-edu.ac.jp).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Plant cultivation. Seeds were surface-sterilized and sown in sterilized vermiculite. Seedlings were inoculated with *M. loti* that had been cultured on yeast extract–mannitol–agar medium. The plants were grown in nitrogen-free nutrient solution in a greenhouse or in a controlled environment growth chamber[4].

Phenotypic analyses. Nodules were fixed overnight in 4% paraformaldehyde and 1% glutaraldehyde in 0.1 M sodium phosphate pH 7.2 at 4 °C. The fixed nodules were dehydrated in an ethanol series and embedded in Teknovit 7100 (Kulzer) in accordance with the manufacturer's instructions. Thin sections were made with an ultramicrotome (UltraCut-R; Leica Microsystems) with a glass knife and stained with toluidine blue. Nitrogenase activity was determined by an acetylene reduction assay. In brief, nodulated roots were placed in 12–35-ml vials containing 10% (v/v) acetylene and incubated for 30 min at 25 °C. The amounts of ethylene produced were determined by gas chromatography[28].

cDNA cloning. The *FEN1* cDNA was amplified from RNA isolated from the Gifu nodules with a SMART RACE cDNA amplification kit (TakaRa) with 5′ RACE and 3′ RACE primers (Supplementary Table 1). Full-length *FEN1* cDNA was obtained by ligation of both 5′ and 3′ RACE products. cDNAs for *GmN56* and *AtIPMS2* were amplified by reverse transcriptase polymerase chain reaction (RT–PCR) from RNA isolated from soybean nodules and *Arabidopsis* leaves, respectively, using SuperScript II reverse transcriptase (Invitrogen) and Expand-High-Fidelity DNA polymerase (Roche), with primer sets designed from published sequences. *AtIPMS2* was cloned without the predicted amino-terminal targeting sequence[9]. All the primer sequences used in this study are shown in Supplementary Table 1.

Complementation of the *fen1* mutant. The *FEN1* promoter fragment (2 kilobases (kb) upstream of the translation start) and the *FEN1* terminator fragment (1 kb downstream of the stop codon) were amplified from wild-type Gifu genomic DNA by PCR. The entire *FEN1*, *GmN56* or *AtIPMS2* cDNA was inserted between them, followed by ligation into a binary vector, pC1300GFP[30]. These constructs were transformed into *A. rhizogenes* LBA1334, and then introduced into the *fen1-1* mutant by the hairy-root transformation procedure as described[31].

Expression analyses. Northern blot analyses were performed as described previously[32]. For assay of the promoter activity, the amplified *FEN1* promoter and terminator fragments were inserted into pC1300GFP, and then a Gateway vector conversion cassette (Invitrogen) was inserted between the promoter and terminator fragments. The *gusA* gene (Invitrogen) was inserted into the cassette by LR clonase II (Invitrogen) to construct the *FEN1* promoter–*gusA* fusion gene (p*FEN1*–*gusA*–t*FEN1*). The fusion gene was introduced into Gifu by *A. rhizogenes*-mediated hairy-root transformation[31]. The nodules formed on transgenic roots were embedded in 5% agar and sectioned at 100 μm thickness with a microslicer (DTK-1000; Dohan EM), followed by incubation for 10–16 h in staining solution (2 mM 5-bromo-4-chloro-3-indolyl-β-D-glucuronide, 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 100 mM sodium phosphate pH 7.0). The stained sections were observed with a light microscope.

Complementation of *S. cerevisiae* mutants. The coding region for *FEN1* was amplified as described above. The mutated *FEN1* gene containing a single nucleotide mutation (*FEN1m*, corresponding to mutation in *fen1-1*) was amplified from a mixture of two overlapping DNA fragments, which were amplified by Fen1 forward primer and Fen1m internal reverse primer, and Fen1m internal forward primer and Fen1 reverse primer. The amplified PCR fragments and *AtIPMS2* cDNA were ligated into pYES2 (Invitrogen) yeast expression vector containing a *GAL1* promoter. The resultant constructs were introduced into a *S. cerevisiae* IPMS mutant YMRX-3B[8] and a HCS mutant 27T6d[10]. Transformants were selected by uracil prototrophy. The production of recombinant proteins was induced by incubation at 25 °C with the addition of galactose. The *S. cerevisiae* cells were collected by centrifugation and were broken with glass beads in 50 mM phosphate buffer pH 7.5 containing 1 mM phenylmethylsulphonyl fluoride. IPMS activity was assayed by an endpoint assay based on the determination of coenzyme A[9]. Homocitrate concentration in the transformed *S. cerevisiae* cells was determined by LC–MS–MS as described below.

Determination of 2-oxoglutarate and homocitrate. Organic acids in nodules, roots and shoots were extracted with 70% ethanol. After the removal of ethanol by evaporation, the extracts were passed through a Dowex 50 column and then loaded on a Dowex 1 column. The Dowex 1 column was washed with 15 ml of water, and the carboxylic acids were eluted with 15 ml of 6 M formic acid. After evaporation, the samples were dissolved in water. 2-Oxoglutarate was analysed by HPLC at 60°C with two tandemly connected anion-exclusion columns (Shodex RSpak KC-811, 8.0 mm × 300 mm; Showa Denko K.K.) with 3 mM perchloric acid solution pH 2.1 as eluent (1 ml min^{-1}). Peaks of organic acids were detected with a post-column bromothymol blue method at a wavelength of 440 nm. Homocitrate was measured with an API 3000 LC–MS–MS system (Applied Biosystems/MDS Analytical Technologies) using selected reaction monitoring. Samples were analysed in negative-ion mode. Samples were loaded by connecting the mass spectrometer to a HPLC (Nanospace SI2; Shiseido Co. Ltd) equipped with an ODS column (Sunfire C_{18}, 3.5 μm pore size, 2.1 mm × 150 mm; Waters), using acetonitrile with 0.1% formic acid as an elution solvent. HPLC was run at a flow rate of 0.18 ml min^{-1}. Deprotonated molecule peaks ($[\mathrm{M\text{-}H}]^- = 187$) were fragmented further by collision-induced dissociation, with N_2 as collision gas, and two fragment peaks of $m/z = 125$ and $m/z = 99$ were monitored.

Transformation of *M. loti*. The coding region for *FEN1* was amplified by PCR from cDNA with Fen1 open reading frame (ORF) forward and reverse primers. *NifH* promoter fragment was amplified from *M. loti* TONO genomic DNA with pNifH F and pNifH R1 primers. The 3 × *myc* tag sequence was synthesized by primer extension and amplified with 3 × *myc* F1 and 3 × *myc* R primers. These three fragments were fused by PCR with pNifH F and 3 × *myc* R primers, resulting in the p*NifH*–*FEN1*–3 × *myc* fragment. The coding region for *Azotobacter vinelandii NifV* was amplified by PCR from the genomic DNA with NifV ORF forward and reverse primers. The 3 × *myc*-tagged *NifV* gene with *NifH* promoter fragment was constructed by fusion of the PCR fragments amplified by pNifH F, pNifH R2, 3 × *myc* F2 and 3 × *myc* R primers. The transposon plasmid, pCAM120 (ref. 33), was modified by replacing a *Not*I fragment containing a p*aph*–*gusA*–*ter* cassette with a multi-cloning site of pBluescript II SK^+ with a *trpA* terminator sequence. The *FEN1* or *NifV* fragment with the *NifH* promoter described above was inserted into the modified pCAM120, and the resultant plasmid was introduced into *M. loti* by tri-parental mating, with pRK2013 as a helper plasmid. Bacteroids were isolated from nodules formed by inoculation with transformed *M. loti* as described previously[34]. For immunodetection of FEN1–Myc and *A. vinelandii* NIFV–Myc proteins, the isolated bacteroids were suspended in SDS–PAGE sample buffer (50 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 50 mM dithiothreitol, 0.1% BPB (bromophenol blue)) and subjected to SDS–PAGE on a 10% (w/v) polyacrylamide gel. The proteins were blotted onto an Immobilon-P filter (Millipore) and reacted with anti-c-Myc antibody (1:2,000 dilution; A-14; Santa Cruz Biotechnology). Immunoreactive protein was detected by using the enhanced chemiluminescence (ECL) plus western blotting detection system (GE Healthcare).

Incubation of plants with homocitrate. Plants inoculated with *M. loti* were grown in sterilized vermiculite supplied with half-strength B&D medium containing 0.5 mM potassium nitrate. Ten days after inoculation, the plants were transferred to half-strength B&D medium containing 1 mM homocitrate and grown hydroponically for 4 days. Bacteroids were isolated from nodules as described previously[35], and their soluble protein fractions were obtained by sonication and centrifugation[28]. The proteins were subjected to western blotting analysis with *Rhizobium leguminosarum* anti-nitrogenase components I and II, using the ProtoBlot immunoblotting system (Promega)[28].

30. Shimomura, K. *et al.* LjnsRING, a novel RING finger protein, is required for symbiotic interactions between *Mesorhizobium loti* and *Lotus japonicus. Plant Cell Physiol.* **47,** 1572–1581 (2006).

31. Diaz, C. L., Schlaman, H. R. M. & Spaink, H. P. in *Lotus japonicus Handbook* (Marquez, A. J., ed.) 261–277 (Springer, 2005).

32. Suganuma, N. *et al.* cDNA macroarray analysis of gene expression in ineffective nodules induced on the *Lotus japonicus sen1* mutant. *Mol. Plant Microbe Interact.* **17,** 1223–1233 (2004).

33. Wilson, J. K. *et al.* β-Glucuronidase (GUS) transposons for ecological and genetic studies of rhizobia and other Gram-negative bacteria. *Microbiology* **141,** 1691–1705 (1995).

34. Kumagai, H. *et al.* A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of *Lotus japonicus. Plant Physiol.* **143,** 1293–1305 (2007).

35. Kouchi, H., Fukai, K. & Kihara, A. Metabolism of glutamate and aspartate in bacteroids isolated from soybean root nodules. *J. Gen. Microbiol.* **137,** 2901–2910 (1991).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

An ancient light-harvesting protein is critical for the regulation of algal photosynthesis

Graham Peers[1]†, Thuy B. Truong[1,2]*, Elisabeth Ostendorf[3]*, Andreas Busch[3], Dafna Elrad[4], Arthur R. Grossman[4], Michael Hippler[3] & Krishna K. Niyogi[1,2]

Light is necessary for photosynthesis, but its absorption by pigment molecules such as chlorophyll can cause severe oxidative damage and result in cell death. The excess absorption of light energy by photosynthetic pigments has led to the evolution of protective mechanisms that operate on the timescale of seconds to minutes and involve feedback-regulated de-excitation of chlorophyll molecules in photosystem II (qE). Despite the significant contribution of eukaryotic algae to global primary production[1], little is known about their qE mechanism, in contrast to that in flowering plants[2,3]. Here we show that a qE-deficient mutant of the unicellular green alga *Chlamydomonas reinhardtii*, *npq4*, lacks two of the three genes encoding LHCSR (formerly called LI818). This protein is an ancient member of the light-harvesting complex superfamily, and orthologues are found throughout photosynthetic eukaryote taxa[4], except in red algae and vascular plants. The qE capacity of *Chlamydomonas* is dependent on environmental conditions and is inducible by growth under high light conditions. We show that the fitness of the *npq4* mutant in a shifting light environment is reduced compared to wild-type cells, demonstrating that LHCSR is required for survival in a dynamic light environment. Thus, these data indicate that plants and algae use different proteins to dissipate harmful excess light energy and protect the photosynthetic apparatus from damage.

Understanding environmental effects on photosynthesis is important for determining oceanic and lacustrine influences on climate and biogeochemistry and for estimating the energy available for higher trophic levels. Light fluxes into the water column exert direct control on primary production, and the aquatic light environment is characterized by extremes in space and time. Seasonal changes of insolation restrict photosynthesis in polar oceans but at the other extreme, surface waves can rapidly focus incident light several fold above its ambient flux[5] and beyond the photosynthetic capacity of the biota. Aquatic photosynthesis requires a balance between efficient light capture and a photoprotective capacity to avoid over-excitation of the photosystems[6].

A major photoprotective strategy, called qE, operates on a timescale of seconds to minutes and involves a regulated thermal dissipation of excess absorbed light energy. The rapid induction and relaxation of qE are required to deal with frequent, rapid changes in the natural light environment[7]. A current model of qE in plants is as follows. The thylakoid lumen in chloroplasts becomes more acidic when the generation of chemical energy by the light reactions of photosynthesis exceeds the capacity of assimilatory reactions such as carbon dioxide fixation. The low pH induces synthesis of zeaxanthin via a xanthophyll cycle and protonation of a photosystem II protein, PSBS, which transduces a conformational change to specific chlorophyll- and carotenoid-binding light-harvesting complex (LHC) proteins. Dissipation of excess excitation energy occurs by a charge-transfer mechanism involving a carotenoid radical cation[3] and/or by chlorophyll-to-carotenoid energy transfer[2]. The photosynthetic systems of algae share many central functions with land plants. However, plants and algae have diversified for several hundred million years and have established distinct biochemical strategies for survival in their respective environments[8].

The generation and analysis of plant and algal mutants that are deficient in non-photochemical quenching of chlorophyll fluorescence (NPQ) have been instrumental in identifying the molecular components of qE. Studies of *npq* mutants of *Chlamydomonas reinhardtii* provided the first genetic evidence for the role of zeaxanthin[9] and a major light-harvesting complex of photosystem II[10] in the qE process. The sequencing of the *Chlamydomonas* genome[11] has now provided the opportunity for us to characterize other *npq* mutants, such as *npq4* (ref. 9), and to expand the exploration of qE-based photoprotection mechanisms in eukaryotic algae.

The qE capacity of wild-type *Chlamydomonas* is dependent on growth conditions. When the wild type and the *npq4* mutant were cultured in subsaturating light fluxes (40 μmol photons $m^{-2} s^{-1}$), they grew at identical rates and neither strain was able to generate significant qE, measured as NPQ that is rapidly inducible by saturating light and reversible during subsequent darkness (Fig. 1a). However, growth in high light (325 μmol photons $m^{-2} s^{-1}$) strongly induced qE capacity in wild-type cells but not in the *npq4* mutant, which maintained a phenotype resembling that of low-light-grown cells (Fig. 1b).

In steady-state low light or high light, *Chlamydomonas* wild type and *npq4* exhibited identical growth rates and apparent quantum efficiencies of oxygen evolution per unit chlorophyll (α, Table 1). However, the wild type had a slightly lower chlorophyll *a:b* ratio compared to *npq4* (Table 1); because chlorophyll *b* is found in the photosynthetic antenna proteins, this suggests differences in light-harvesting antenna structure. Interestingly, the xanthophyll cycle, known to be involved in qE[9], was fully operational in high-light-grown *npq4* (Supplementary Table 1). The higher chlorophyll per cell values seen in low-light-grown *npq4* appeared to be due to slightly larger cell sizes compared to the wild type (data not shown).

qE is proposed to be a mechanism that protects oxygenic photosynthesis in a changing light environment[7]. We investigated whether the *npq4* mutant is more susceptible to photo-oxidative damage by shifting cells from a low-light-acclimated state to excess light (1,100 μmol photons $m^{-2} s^{-1}$) for 4 h. These cells were then spotted onto solid medium, and survivors were allowed to grow for a week at 400 μmol photons $m^{-2} s^{-1}$. Figure 2 shows that fewer *npq4* cells were able to survive the shift between low and excess light. Following an independent light-shift experiment we plated the

[1] Department of Plant and Microbial Biology, University of California, Berkeley, California 94720-3102, USA. [2] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3] Institute of Plant Biochemistry and Biotechnology, University of Münster, 48143 Münster, Germany. [4] Department of Plant Biology, Carnegie Institution, Stanford, California 94305, USA. †Present address: Synthetic Genomics, Inc., 11149 North Torrey Pines Road, Suite 100, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

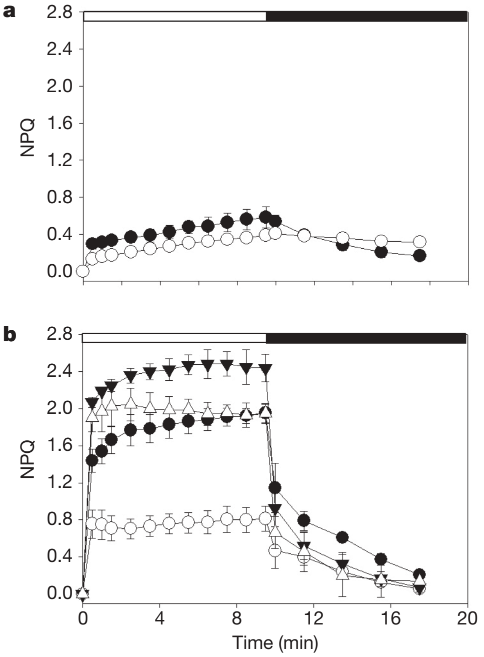


Figure 1 | NPQ of chlorophyll fluorescence in *Chlamydomonas reinhardtii* cells. The cells were exposed to an actinic light of 600 μmol photons m^{-2} s^{-1} (white bar) followed by darkness (black bar). Wild-type (filled circles) and *npq4* mutant (open circles) cells were cultured in photoautotrophic conditions under either a low light level of 40 μmol photons m^{-2} s^{-1} (**a**) or a high light level of 325 μmol photons m^{-2} s^{-1} (**b**). Closed and open triangles in **b** show NPQ for *npq4*comp1 and *npq4*comp2, which are two *npq4* lines independently rescued with genomic *LHCSR3.1*. Data represent means ± s.e. ($n = 3$).

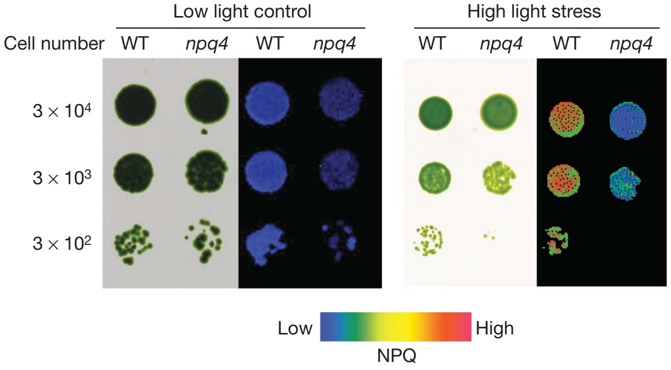


Figure 2 | Survival of wild type (WT) and *npq4* following a shift from a low-light-acclimated state to excess light for four hours. Cells were then plated onto minimal medium and grown for a week before image capture. The images with black background are false-colour images of NPQ capacity after a minute of actinic light. Low light control cells were not exposed to high light.

Chlamydomonas cells and counted the number of surviving cells as colony forming units. The *npq4* cultures had a 40% reduction in survivorship compared to the wild type following the high light shift (Supplementary Fig. 1, Student's *t*-test, $P = 0.005$, $n = 3$), whereas they had statistically identical survivorship when cells were only treated with low light (*t*-test, $P = 0.29$, $n = 3$). These results show that qE, as in *Arabidopsis*[7], is required for optimal survival of *Chlamydomonas* in variable light environments, although qE-deficient mutants of *Chlamydomonas* (Table 1) or *Arabidopsis*[12] are able to acclimate to growth in constant high light in the laboratory.

The *npq4* strain was generated by plasmid insertional mutagenesis[9]. Genetic analyses showed that a single nuclear mutation was responsible for the low NPQ phenotype and that the phenotype co-segregated with the inserted plasmid, suggesting that the mutation was tagged. Plasmid rescue identified a 212-base-pair flanking genomic DNA fragment located on linkage group VIII. Thus, the *Chlamydomonas npq4* mutant is not affected in either of two linked *PSBS* genes, previously shown to be critical for qE in plants[13], which are located on linkage group I. The flanking genomic DNA fragment sequenced in the *npq4* mutant is ~2 kilobases upstream of two genes (*LHCSR3.1* and *LHCSR3.2*) that encode identical LHCSR proteins (LHCSR3). Integration of plasmid DNA into the *Chlamydomonas* genome is often accompanied by a deletion and/or rearrangement of adjacent DNA, and genomic polymerase chain reaction (PCR) analysis showed that neither *LHCSR3.1* nor *LHCSR3.2* is intact (Fig. 3a). Correspondingly, *LHCSR3.1* and *LHCSR3.2* messenger RNAs were undetectable in *npq4* (Fig. 3b). A third *LHCSR* gene (*LHCSR1*), which encodes a protein that is 82% identical to LHCSR3, is located on the same linkage group and is ~240 kilobases upstream of the insert. This gene is intact and is transcribed in the *npq4* mutant (Fig. 3a and b).

LHCSR is a stress-related member of the LHC protein superfamily. LHCSR was originally identified as a light-induced transcript (called *LI818*)[14], and its expression is not coordinated with most other LHCs[15], which have a primary role in light harvesting. *LHCSR* transcripts accumulate under environmental conditions known to induce photo-oxidative stress, including deprivation of carbon dioxide[16], sulphur[17], or iron[18], as well as high light[19]. Under our conditions, we measured higher *LHCSR1* and *LHCSR3.1* mRNA (Fig. 3c) and higher LHCSR1 and LHCSR3 protein levels in high-light-grown cells compared to low-light-grown cells (Fig. 3d). Immunoblot analysis with a specific anti-LHCSR3 antibody[18] showed unambiguously that LHCSR3 is absent from *npq4* (Fig. 3d). Thus, the accumulation of LHCSR3 protein is correlated with qE capacity (Fig. 1). *LHCSR1* mRNA was still induced by high light in *npq4*, and an antibody that recognizes both LHCSR1 and LHCSR3 showed that LHCSR1 also accumulated (Fig. 3d), but this protein alone is clearly not sufficient for qE. *Chlamydomonas* upregulates LHCSR within one hour of a shift from dark to light[15] and we found that wild-type cells were already able to induce LHCSR protein levels and qE capacity within three hours of a shift from low light to high light (data not shown). This suggests that LHCSR is upregulated during the time period of our fitness tests. The mRNAs encoding LHCSR1 and LHCSR3 are differentially regulated in response to low carbon dioxide[20], so they may be regulated by different cellular signals.

Table 1 | Effects of light level on *Chlamydomonas reinhardtii* cells

	Units	Low light		High light	
		Wild type	*npq4*	Wild type	*npq4*
Growth rate	per day	1.15 ± 0.01	1.03 ± 0.05	1.55 ± 0.03	1.46 ± 0.21
Chlorophyll *a*	fmol per cell	1.92 ± 0.12*	2.40 ± .06*	0.99 ± 0.09	0.91 ± 0.18
Chlorophyll *a*:*b* ratio	mol:mol	2.36 ± 0.01*	2.78 ± 0.01*	2.76 ± 0.15*	3.33 ± 0.08*
Photosystem II efficiency, F_v/F_m		0.74 ± 0.01	0.76 ± 0.01	0.67 ± 0.05	0.66 ± 0.01
Apparent quantum efficiency of oxygen evolution, α	(mmol O_2 per mol chlorophyll *a*) / (μmol photons per m^2)	0.23 ± 0.02	0.22 ± 0.02	0.27 ± 0.01	0.27 ± 0.02

Chlamydomonas reinhardtii cells were cultured in low light (40 μmol photons m^{-2} s^{-1}) or high light conditions (325 μmol photons m^{-2} s^{-1}). Data represent means ± s.e. ($n = 3$).
*Significantly different within light treatment ($P < 0.05$, Student's *t*-test).

©2009 Macmillan Publishers Limited. All rights reserved

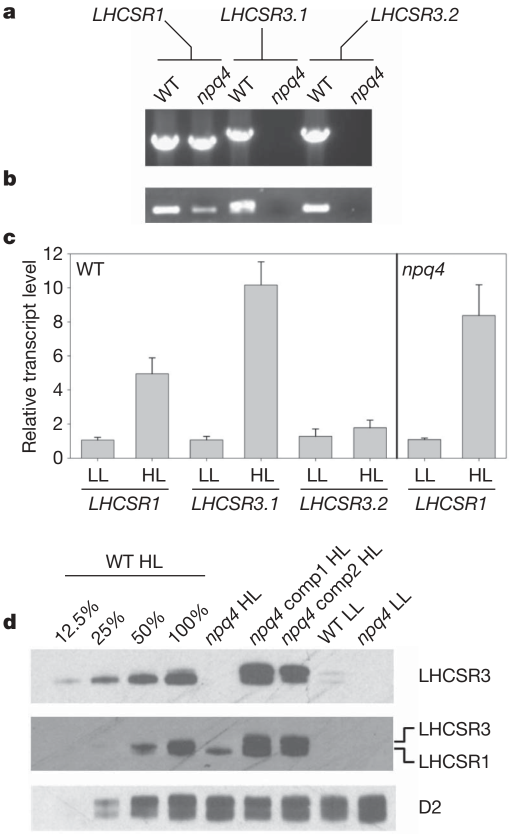


Figure 3 | Analysis of the *LHCSR* genes, *LHCSR* mRNA levels, and LHCSR protein accumulation in *Chlamydomonas* WT and *npq4* mutant. **a**, Products from PCR reactions that specifically target each *LHCSR* gene. **b**, Products from reverse transcription (RT)–PCR reactions that specifically target each *LHCSR* mRNA. **c**, Quantitative real-time RT–PCR analysis of *LHCSR* transcript levels in cells grown in low light (LL) or in high light (HL). Data represent means ± s.e. ($n = 4$–8). **d**, Immunoblot analysis of LHCSR and D2 protein of the photosystem II reaction centre in LL- and HL-grown cells. LHCSR3.1 and LHCSR3.2 have identical amino acid sequences with a predicted size of 26.1 kDa, whereas the predicted size of LHCSR1 is slightly smaller (25.4 kDa). 0.25 nmol of chlorophyll was loaded in each experimental lane. Different amounts of the WT HL sample were loaded for quantification.

Multiple lines of evidence confirm that the qE deficiency of *npq4* is caused by the loss of LHCSR3. We rescued the qE phenotype by re-introducing the *LHCSR3.1* gene into *npq4*. Interestingly, some rescued lines accumulated more LHCSR3 than the wild type (shown as *npq4*comp1, Fig. 3d), and this correlated with higher qE capacity (Fig. 1). We also generated independent RNAi-mediated (Supplementary Fig. 2) and artificial microRNA-mediated (data not shown) knock-down strains of LHCSR3 that showed a proportional reduction in qE capacity. Like the *npq4* mutant, the RNAi line was more sensitive to excess light than its respective wild type (data not shown). From the results of these experiments, it is clear that the low-qE phenotype of *npq4* is due to the absence of LHCSR3 and not a protein encoded by a closely linked gene.

Proteomic analysis of thylakoid proteins by two-dimensional gel electrophoresis and mass spectrometry confirmed that the LHCSR3 protein is absent in *npq4* (Supplementary Fig. 3). All other components of the light-harvesting antenna are present, including LHCBM1, which was previously shown to play a part in qE[10]. It is possible that the absence of LHCSR3 may alter the structure or stability of photosystem II supercomplexes; we did not quantify the individual antenna components in our two-dimensional analyses, but the differences in chlorophyll *a*:*b* ratios between *npq4* and the wild type (Table 1) suggest there may be some minor remodelling of the antenna. The photosystem II antenna undergoes significant changes in light-shift experiments (such as phosphorylation[21]), and we are currently investigating how LHCSR interacts with other antenna components during qE.

Following the initial endosymbiotic event that resulted in photosynthetic eukaryotes, algae have undergone significant diversification. Genes encoding the PSBS protein are found throughout green algae, in the model moss *Physcomitrella*, and in vascular plants. PSBS has a major role in qE in vascular plants such as *Arabidopsis*[13] but, despite the presence of *PSBS* genes, the PSBS protein has not been detected in *Chlamydomonas*[22] and other unicellular green algae[23]. In contrast, *LHCSR* genes are found in green algae and *Physcomitrella*[24,25], but are absent from vascular plants. LHCSR proteins are clearly present in *Chlamydomonas*[18,26], and they are also detectable in the primitive prasinophyte green alga, *Ostreococcus tauri* (Supplementary Fig. 4). As shown above for *Chlamydomonas*, the accumulation of LHCSR proteins is induced by high light and correlated with qE capacity in *Ostreococcus* (Supplementary Fig. 4). Thus, vascular plants and green algae appear to employ different proteins to regulate photosynthetic light harvesting in excess light.

Besides being absent in vascular plants, LHCSR orthologues are missing only in Rhodophyte algae (and cyanobacteria), which dissipate excess light energy from phycobilisomes by a mechanism distinct from qE[27]. However, LHCSR relatives are found throughout the chromalveolate algae (for example, diatoms and prymnesiophytes), which have a plastid that is derived from a red alga. The sequenced genomes of two diatoms and expressed sequence tags from a wide variety of algae outside of the green clade reveal no sequences encoding PSBS[4], but these organisms are clearly capable of qE[28]. This suggests that some other protein must confer the flexibility of switching the antenna of photosystem II between the light-harvesting and photoprotective states. We propose that this protein is LHCSR, which in chromalveolates is encoded by a gene that appears to have been acquired laterally in a cryptic endosymbiotic event involving a prasinophyte-like alga[29]. The retention of *LHCSR* genes after the subsequent endosymbiosis of a red alga suggests that they confer a significant selective advantage that is most probably related to photoprotection.

METHODS SUMMARY

Strains and growth conditions. The *Chlamydomonas reinhardtii* wild-type strain 4A+ (137c genetic background) was obtained from J.-D. Rochaix (University of Geneva). The *npq4* mutant was generated from the arginine-requiring CC-425 background as described previously[9]. It was crossed four times to the 4A+ strain before physiological characterization. All physiological measurements were performed on fully acclimated cells cultured photoautotrophically in a constant light environment and at 25 °C as described previously[9]. Cells were grown in low light (40 μmol photons $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$) or high light (325 μmol photons $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$).

Chlorophyll fluorescence, oxygen evolution and photosynthetic pigments. Chlorophyll fluorescence measurements of *Chlamydomonas* cells were performed with a Hansatech FMS2 system or with a custom fluorescence imager as previously described[9] but with some modifications. Cells were dark-acclimated for 30 to 60 min before measurement, then gently filtered onto a glass-fibre filter and placed on the instrument's leaf clip. The maximum efficiency of photosystem II, $(F_\mathrm{m} - F_\mathrm{o})/F_\mathrm{m} = F_\mathrm{v}/F_\mathrm{m}$, was measured after 5 min of far-red light to ensure transition into state I. F_o is the fluorescence resulting from the measuring light alone. F_m is the maximum fluorescence measured during a brief, saturating flash of light. F_v is the variable fluorescence. Cells were exposed to actinic light of 600 μmol photons $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$ to induce NPQ. Total NPQ was calculated as $(F_\mathrm{m} - F_\mathrm{m}')/F_\mathrm{m}'$, where F_m' is the maximum fluorescence measured in the light-adapted state (during or after actinic light illumination). Oxygen evolution, chlorophyll content per cell, and xanthophyll cycle pigments were measured on exponentially growing cultures ($<1 \times 10^6$ cells ml^{-1}) as described[30].

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

Received 10 September; accepted 19 October 2009.

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237–240 (1998).
2. Ruban, A. V. *et al.* Identification of a mechanism of photoprotective energy dissipation in higher plants. *Nature* 450, 575–578 (2007).

©2009 Macmillan Publishers Limited. All rights reserved

3. Ahn, T. K. *et al.* Architecture of a charge-transfer state regulating light harvesting in a plant antenna protein. *Science* **320,** 794–797 (2008).
4. Koziol, A. G. *et al.* Tracing the evolution of the light-harvesting antennae in chlorophyll *a/b*-containing organisms. *Plant Physiol.* **143,** 1802–1816 (2007).
5. Schenck, H. On the focusing of sunlight by ocean waves. *J. Opt. Soc. Am.* **47,** 653–657 (1957).
6. Niyogi, K. K. Photoprotection revisited: genetic and molecular approaches. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50,** 333–359 (1999).
7. Külheim, C., Ågren, J. & Jansson, S. Rapid regulation of light harvesting and plant fitness in the field. *Science* **297,** 91–93 (2002).
8. Falkowski, P. G. *et al.* The evolution of modern eukaryotic phytoplankton. *Science* **305,** 354–360 (2004).
9. Niyogi, K. K., Björkman, O. & Grossman, A. R. *Chlamydomonas* xanthophyll cycle mutants identified by video imaging of chlorophyll fluorescence quenching. *Plant Cell* **9,** 1369–1380 (1997).
10. Elrad, D., Niyogi, K. K. & Grossman, A. R. A major light-harvesting polypeptide of photosystem II functions in thermal dissipation. *Plant Cell* **14,** 1801–1816 (2002).
11. Merchant, S. S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318,** 245–251 (2007).
12. Golan, T., Müller-Moulé, P. & Niyogi, K. K. Photoprotection mutants of *Arabidopsis thaliana* acclimate to high light by increasing photosynthesis and specific antioxidants. *Plant Cell Environ.* **29,** 879–887 (2006).
13. Li, X.-P. *et al.* A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* **403,** 391–395 (2000).
14. Gagné, G. & Guertin, M. The early genetic response to light in the green unicellular alga *Chlamydomonas eugametos* grown under light dark cycles involves genes that represent direct responses to light and photosynthesis. *Plant Mol. Biol.* **18,** 429–445 (1992).
15. Savard, F., Richard, C. & Guertin, M. The *Chlamydomonas reinhardtii* LI818 gene represents a distant relative of the *cabI/II* genes that is regulated during the cell cycle and in response to illumination. *Plant Mol. Biol.* **32,** 461–473 (1996).
16. Miura, K. *et al.* Expression profiling-based identification of CO_2-responsive genes regulated by CCM1 controlling a carbon-concentrating mechanism in *Chlamydomonas reinhardtii. Plant Physiol.* **135,** 1595–1607 (2004).
17. Zhang, Z. *et al.* Insights into the survival of *Chlamydomonas reinhardtii* during sulfur starvation based on microarray analysis of gene expression. *Eukaryot. Cell* **3,** 1331–1348 (2004).
18. Naumann, B. *et al.* Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathways under iron deficiency in *Chlamydomonas reinhardtii. Proteomics* **7,** 3964–3979 (2007).
19. Ledford, H. K. *et al.* Comparative profiling of lipid-soluble antioxidants and transcripts reveals two phases of photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas reinhardtii. Mol. Genet. Genom.* **272,** 470–479 (2004).
20. Yamano, T., Miura, K. & Fukuzawa, H. Expression analysis of genes associated with the induction of the carbon-concentrating mechanism in *Chlamydomonas reinhardtii. Plant Physiol.* **147,** 340–354 (2008).
21. Turkina, M. V. *et al.* Environmentally modulated phosphoproteome of photosynthetic membranes in the green alga *Chlamydomonas reinhardtii. Mol. Cell. Proteom.* **5,** 1412–1425 (2006).
22. Allmer, J., Naumann, B., Markert, C., Zhang, M. & Hippler, M. Mass spectrometric genomic data mining: novel insights into bioenergetic pathways in *Chlamydomonas reinhardtii. Proteomics* **6,** 6207–6220 (2006).
23. Bonente, G. *et al.* The occurrence of the *psbS* gene product in *Chlamydomonas reinhardtii* and in other photosynthetic organisms and its correlation with energy quenching. *Photochem. Photobiol.* **84,** 1359–1370 (2008).
24. Alboresi, A., Caffarri, S., Nogue, F., Bassi, R. & Morosinotto, T. *In silico* and biochemical analysis of *Physcomitrella patens* photosynthetic antenna: identification of subunits which evolved upon land adaptation. *PLoS One* **3,** doi:10.1371/journal.pone.0002033 (2008).
25. Rensing, S. A. *et al.* The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319,** 64–69 (2008).
26. Richard, C., Ouellet, H. & Guertin, M. Characterization of the LI818 polypeptide from the green unicellular alga *Chlamydomonas reinhardtii. Plant Mol. Biol.* **42,** 303–316 (2000).
27. Wilson, A. *et al.* A soluble carotenoid protein involved in phycobilisome-related energy dissipation in cyanobacteria. *Plant Cell* **18,** 992–1007 (2006).
28. Casper-Lindley, C. & Björkman, O. Fluorescence quenching in four unicellular algae with different light-harvesting and xanthophyll-cycle pigments. *Photosynth. Res.* **56,** 277–289 (1998).
29. Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324,** 1724–1726 (2009).
30. Baroli, I., Do, A. D., Yamane, T. & Niyogi, K. K. Zeaxanthin accumulation in the absence of a functional xanthophyll cycle protects *Chlamydomonas reinhardtii* from photooxidative stress. *Plant Cell* **15,** 992–1008 (2003).

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

Acknowledgements We thank M. Guertin for providing the anti-LHCSR (LI818) antibody. This work was supported by grants from the Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, US Department of Energy (K.K.N.), the Deutsche Forschungsgemeinschaft (M.H.), and the National Science Foundation (A.R.G.).

Author Contributions G.P., T.B.T., E.O., A.B., D.E. and K.K.N. performed research; G.P., A.R.G., M.H., and K.K.N. designed research; G.P., M.H. and K.K.N. analysed data and wrote the paper; all authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.K.N. (niyogi@berkeley.edu) or M.H. (mhippler@uni-muenster.de).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Genomic DNA analyses. Genomic DNA was isolated using a phenol/chloroform extraction method[31]. For plasmid rescue, 10 μg of genomic DNA from the *npq4* mutant was digested overnight with SacI, the enzyme was heat-inactivated for 20 min at 65 °C, and the DNA fragments ligated overnight at 4 °C. After a phenol/chloroform purification, the DNA was ethanol precipitated, and resuspended in 20 μl TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.4). *Escherichia coli* strain DH5α was then transformed with half of the ligation mixture.

To test for the presence of *LHCSR* genes, standard PCR was performed with primer pairs that span each entire gene. The primers used were *LHCSR1* (protein ID 184724 in *Chlamydomonas* genome sequence v3) (5′-TTCTCAGCTTGCACCTCCTT-3′ and 5′-GGCACTTAGATGGCCTTGAG-3′), *LHCSR3.1* (protein ID 184731) (5′-GCTCCTCGACAATCGCTTAC-3′ and 5′-TCATGCTCTCTCTGCTGTGC-3′), and *LHCSR3.2* (protein ID 184730) (5′-ATTAACATGGGCGACTACCG-3′ and 5′-TGTACGCAGTTCAAGGGATG-3′).

RNA analysis. Cells were grown for more than two days to reach log phase and collected by centrifugation at 4 °C. RNA was extracted with TRIZOL and treated with DNase from Invitrogen according to the manufacturer's instructions. First strand complementary DNA was made with Invitrogen Superscript III reagents and protocols. Standard RT–PCR using Taq polymerase was performed to test for the presence of transcripts in the wild type and *npq4*. The primers used were *LHCSR1* (5′-ATCTGCTTCACGGTTTGGTC-3′ and 5′-CACACAATTCTGCCAACAGC-3′), *LHCSR3.1* (5′-CGCACAGTCCTATGGTGTTG-3′ and 5′-TGTTCGCACTCGTCTTCATC-3′), and *LHCSR3.2* (5′-CCAATACACACGATCCCTCTC-3′ and 5′-GGTGGAAGAGTATCGCAAGC-3′). To measure the expression level of these genes by quantitative RT–PCR, we employed the SYBR Green Master Mix and an Applied BioSystems 7000 qPCR machine. The efficiencies of the primers were between 95% and 100%. All samples were calibrated against the wild type in low light. A gamma-tubulin gene (*TUG*, protein ID 188933) was used as the endogenous control, and the ΔΔCt method was used to calculate mRNA levels. The *TUG* primers used were 5′-CGCCAAGTACATCTCCATCC-3′ and 5′-TAGGGGCTCTTCTTGGACAG-3′.

Complementation of *npq4*. A genomic clone of *LHCSR3.1* with its endogenous promoter (~1,000 base pairs upstream of the 5′ untranslated region, 5′-UTR) was amplified from wild-type genomic DNA with the primers 5′-TTCAAGGGATGAGCAAGTT-3′ and 5′-CACCGCTGACTCCCCTGTCTTCAG-3′ and cloned into the entry vector pENTR'D according to the manufacturer's protocol (Invitrogen). The gene was sequenced and cloned into a novel GATEWAY vector (called GwypBC1) using the Invitrogen GATEWAY LR Clonase II enzyme mix. GwypBC1 was created by cloning the GATEWAY fragment (*cddB* gene flanked by *attR* sites) from the pEARLEYGATE 205 plasmid[32] into the XhoI sites of pBC1 (pBluescript with the addition of a paromomycin resistance marker). *npq4* cells were transformed, selected on paromomycin, and then screened for NPQ capacity as described in the Methods Summary. Positive colonies were then assayed for NPQ capacity as previously described[33] and assayed for the presence of the LHCSR3 protein as described below.

Immunoblot analysis. Cells were harvested in exponential growth phase ($<1 \times 10^6$ cells ml^{-1}) by centrifugation (1,800*g* for 4 min). Cells were resuspended in standard SDS denaturing buffer and lysed with vigorous shaking at room temperature (25 °C) for 5 min. An aliquot was removed for chlorophyll determination[34] and insoluble material was removed by centrifugation at 10,000*g* for 10 min. Samples were diluted to a final concentration of 0.1 nmol chlorophyll μl^{-1}. Protein homogenates were loaded on pre-cast Novex 10–20% Tricine gels (Invitrogen). 35 mA was applied until the 15 kDa molecular weight standard ran off the gel. Proteins were blotted onto nitrocellulose membranes using standard methods. Membranes were blocked overnight with 2% milk in TBS and then incubated with anti-LHCSR polyclonal antibody[35], diluted 1:10,000 in 0.5% milk in TBS), anti-LHCSR3 antibody[36] diluted 1:20,000, or anti-D2 antibody (Agrisera), diluted 1:5,000. Membranes were incubated for one hour and then rinsed four times for 5 min before incubation with 1:100,000 WestFemto (Pierce) secondary antibodies, and reactive bands were detected according to the manufacturer's protocol. PSAD was resolved and detected as described previously[37].

31. Baroli, I., Do, A. D., Yamane, T. & Niyogi, K. K. Zeaxanthin accumulation in the absence of a functional xanthophyll cycle protects *Chlamydomonas reinhardtii* from photooxidative stress. *Plant Cell* **15**, 992–1008 (2003).
32. Earley, K. *et al.* Gateway-compatible vectors for plant functional genomics and proteomics. *Plant J.* **45**, 616–629 (2006).
33. Niyogi, K. K., Björkman, O. & Grossman, A. R. *Chlamydomonas* xanthophyll cycle mutants identified by video imaging of chlorophyll fluorescence quenching. *Plant Cell* **9**, 1369–1380 (1997).
34. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochim. Biophys. Acta* **975**, 384–394 (1989).
35. Richard, C., Ouellet, H. & Guertin, M. Characterization of the LI818 polypeptide from the green unicellular alga *Chlamydomonas reinhardtii. Plant Mol. Biol.* **42**, 303–316 (2000).
36. Naumann, B. *et al.* Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathways under iron deficiency in *Chlamydomonas reinhardtii. Proteomics* **7**, 3964–3979 (2007).
37. Naumann, B., Stauber, E. J., Busch, A., Sommer, F. & Hippler, M. N-terminal processing of Lhca3 is a key step in remodeling of the photosystem-I light-harvesting complex under iron deficiency in *Chlamydomonas reinhardtii. J. Biol. Chem.* **280**, 20431–20441 (2005).

©2009 Macmillan Publishers Limited. All rights reserved

nature Vol 462|26 November 2009|doi:10.1038/nature08576

LETTERS

Innate immune and chemically triggered oxidative stress modifies translational fidelity

Nir Netzer[1]*, Jeffrey M. Goodenbour[2]*, Alexandre David[1], Kimberly A. Dittmar[3], Richard B. Jones[4], Jeffrey R. Schneider[5], David Boone[5], Eva M. Eves[4], Marsha R. Rosner[4], James S. Gibbs[1], Alan Embry[1], Brian Dolan[1], Suman Das[1], Heather D. Hickman[1], Peter Berglund[1], Jack R. Bennink[1], Jonathan W. Yewdell[1]* & Tao Pan[3]*

Translational fidelity, essential for protein and cell function, requires accurate transfer RNA (tRNA) aminoacylation. Purified aminoacyl-tRNA synthetases exhibit a fidelity of one error per 10,000 to 100,000 couplings[1,2]. The accuracy of tRNA aminoacylation *in vivo* is uncertain, however, and might be considerably lower[3–6]. Here we show that in mammalian cells, approximately 1% of methionine (Met) residues used in protein synthesis are aminoacylated to non-methionyl-tRNAs. Remarkably, Met-misacylation increases up to tenfold upon exposing cells to live or non-infectious viruses, toll-like receptor ligands or chemically induced oxidative stress. Met is misacylated to specific non-methionyl-tRNA families, and these Met-misacylated tRNAs are used in translation. Met-misacylation is blocked by an inhibitor of cellular oxidases, implicating reactive oxygen species (ROS) as the misacylation trigger. Among six amino acids tested, tRNA misacylation occurs exclusively with Met. As Met residues are known to protect proteins against ROS-mediated damage[7], we propose that Met-misacylation functions adaptively to increase Met incorporation into proteins to protect cells against oxidative stress. In demonstrating an unexpected conditional aspect of decoding mRNA, our findings illustrate the importance of considering alternative iterations of the genetic code.

Owing to the central importance of tRNA aminoacylation and translational accuracy in understanding the biology of mammalian cells under normal and pathological conditions, we devised a method to measure tRNA misacylation in cells. Our method combines pulse radiolabelling of cells with [^{35}S]Met with microarrays developed for measuring tRNA abundance[8] (Fig. 1a and Supplementary Figs 1 and 2). We hybridized total tRNA to arrays that detect the 274 distinct chromosomal human tRNA species as closely related members of 42 families and all 22 mitochondrial tRNAs, and used phosphorimage analysis to visualize and quantify [^{35}S]Met-tRNAs hybridized to the array.

For HeLa cells, we detected intense radioactive spots representing five different methionyl-tRNA (tRNA$^{\mathrm{Met}}$) probes, as expected. Unexpectedly, we easily detected less intense radioactive spots representing several non-tRNA$^{\mathrm{Met}}$ probes. Further, when we infected HeLa cells with influenza A virus or adenovirus 5 before pulse radiolabelling, the level of radioactive signals from non-tRNA$^{\mathrm{Met}}$ probes greatly increased (Fig. 1a), and could be exclusively detected by adding excess Met-oligonucleotide probes to block hybridization of tRNA$^{\mathrm{Met}}$ (Fig. 1b and Supplementary Fig. 3). We used multiple approaches to establish conclusively that radioactivity emanating from non-cognate tRNA probes derives from aminoacylated [^{35}S]Met and not other ^{35}S-containing material (detailed in Supplementary Methods and Supplementary Information). We excluded radiolabelling of tRNAs with thio-modifications by catabolism of [^{35}S]Met (Fig. 1c and Supplementary Fig. 4). We validated [^{35}S]Met-misacylation using two non-array-based methods for several tRNA species (Fig. 1d, e and Supplementary Fig. 5). To distinguish aminoacyl- from peptidyl-tRNAs, we treated total RNA with aminopeptidase M before array hybridization to remove amino (N)-terminal [^{35}S]Met residues from peptidyl-tRNAs (Fig. 1f and Supplementary Fig. 6). We also excluded misacylation resulting from contaminants in the [^{35}S]Met preparation (Supplementary Fig. 7).

In uninfected cells, the eight cytosolic misacylated tRNA families totalled approximately 1.5% of the cumulative radioactivity of all five tRNA$^{\mathrm{Met}}$ families. Upon infection with influenza A virus, misacylation increased in three of the eight species and appeared in 18 new tRNA families. Remarkably, the cumulative radioactivity on non-tRNA$^{\mathrm{Met}}$ species totalled about 13% of that of all tRNA$^{\mathrm{Met}}$ families (Fig. 1g and Supplementary Fig. 8). Cells infected with adenovirus or vaccinia virus demonstrated a similar pattern and degree of misacylation. Increased Met-misacylation in virus-infected cells is not an artefact of increased tRNA expression, because increased misacylation does not correlate with the minor changes in tRNA abundance induced by viral infection (Supplementary Fig. 9). Under all conditions tested, we failed to detect misacylation of any mitochondrial tRNA, demonstrating the selectivity of misacylation for cytosolic tRNAs (Fig. 1).

We next demonstrated tRNA-Met-misacylation *in vitro* (Supplementary Figs 10 and 11). HeLa-cell-derived methionyl-tRNA synthetase (MetRS) migrates in two major sucrose gradient fractions: one containing the 11-protein multi-synthetase complex, the other containing the multi-synthetase complex associated with polysomes[9]. Each sedimenting form of the multi-synthetase complex demonstrated similar acylation activity with [^{35}S]Met. The polysome-associated form clearly mediated misacylation among a subset of the misacylated tRNA families identified *in vivo*. The free form of the multi-synthetase complex exhibited less misacylation activity, demonstrating that the fidelity of tRNA synthetases can depend on the higher-order structure of the AARS. Further, we showed that although the multi-synthetase complex misacylated tRNA$^{\mathrm{Lys}}$ isoacceptors, free LysRS did not (Supplementary Fig. 11). This is consistent with misacylation being performed by MetRS within the multi-RS complex.

Because aminoacylated tRNAs can be used for non-translation functions[10,11], it was critical to establish whether Met-misacylated tRNAs are used in translation. A pulse-chase experiment revealed that cognate and non-cognate tRNAs demonstrate a similar off-rate for [^{35}S]Met after a 3-min chase period with excess unlabelled Met (Fig. 2a). Blocking translation by incubating cells with cycloheximide during

[1]Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, USA. [2]Department of Human Genetics, [3]Department of Biochemistry and Molecular Biology, [4]Ben May Institute, [5]Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

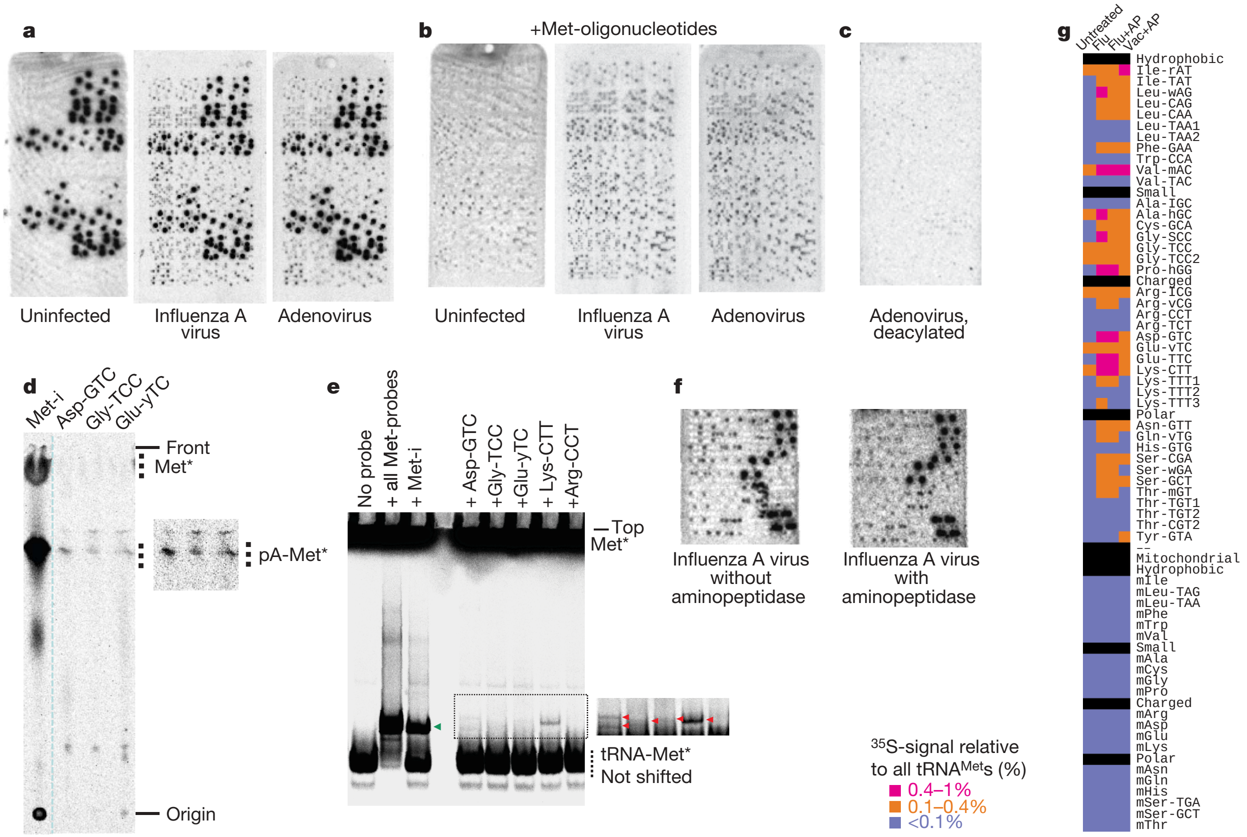


Figure 1 | Induction of tRNA misacylation by viruses. a, Microarrays showing total tRNAs isolated from uninfected, influenza A virus- and adenovirus-infected HeLa cells. **b**, Large excess of oligonucleotides complementary to tRNA^{Met} was included in array hybridization. **c**, The adenovirus-infected sample was deacylated before array hybridization. **d**, Thin-layer chromatography of the influenza A virus-infected sample using biotinylated oligonucleotide probes (longer exposure in inset). **e**, Non-denaturing acid gel detection of misacylated tRNAs in the influenza A virus-infected sample. **f**, Influenza A virus-infected sample with or without aminopeptidase. **g**, Quantitative comparison of uninfected and virus-infected samples. tRNAs are grouped according to amino-acid properties. The detection limit of misacylation was about 0.1% for each probe.

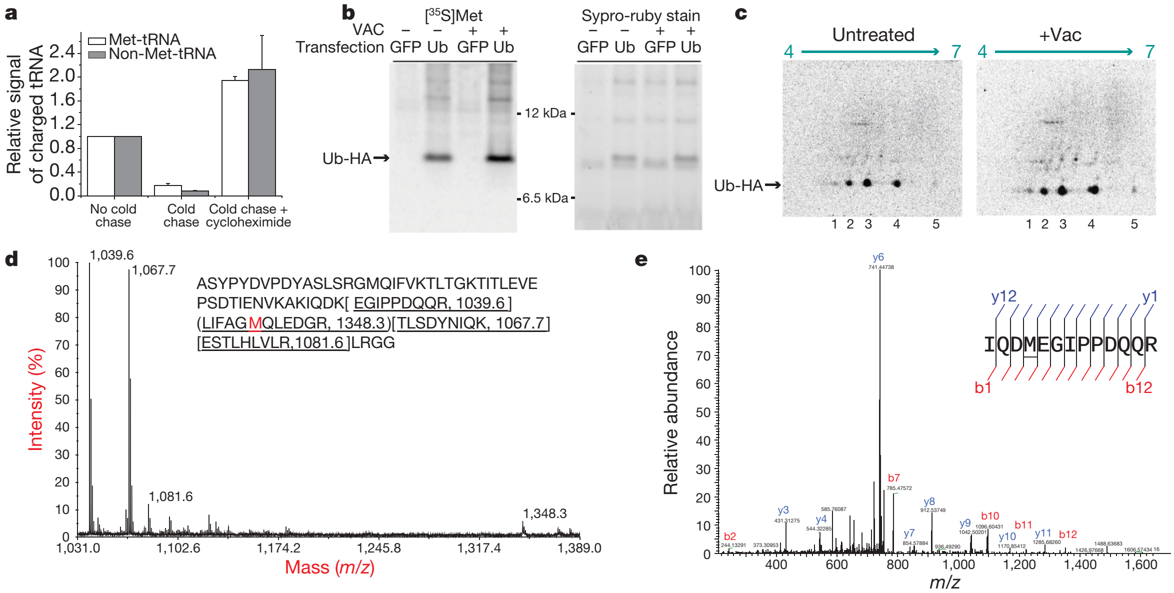


Figure 2 | Misacylated tRNAs are used in translation. a, Correctly acylated and misacylated tRNAs have the same kinetic properties with or without cycloheximide. Error bars represent s.d. ($n = 4$). **b**, One-dimensional SDS–polyacrylamide gel electrophoresis showing an increase in specific activity of [^{35}S]Met incorporation upon infection with vaccinia virus. **c**, Two-dimensional SDS–polyacrylamide gel electrophoresis of uninfected and vaccinia virus-infected samples. Spot 3 (55–62% of all radioactivity) matches the expected isoelectric point (pI) of wild-type Ub-HA. Spots 1 and 5 correspond to Lys/Arg-to-Met and Glu/Asp-to-Met substitution, respectively. **d**, Matrix-assisted laser desorption/ionization–time of flight of tryptic digested Ub-HA. Peaks are labelled with their *m*/*z* values from the Ub-HA sequence. **e**, Mass spectrometry–mass spectrometry sequencing of 763.87 (*m*/2*z*) mass peak by liquid chromatography–Fourier transform mass spectrometry (LC–FTMS) of tryptic digested Ub-HA.

©2009 Macmillan Publishers Limited. All rights reserved

the cold Met chase prevented the loss of [^{35}S]Met from cognate and non-cognate tRNAs in parallel (Fig. 2a and Supplementary Fig. 12a, b). These findings strongly support global use of misacylated tRNAs in protein synthesis. Next, we quantified the incorporation of [^{35}S]Met into haemagglutinin-epitope-tagged ubiquitin (Ub-HA), selected as a reporter because it possesses a single Met residue. Infecting cells with vaccinia virus increased the specific activity of Ub (dpm per µg protein) by about 1.8-fold, consistent with vaccinia-virus-induced increase in tRNA misacylation (Fig. 2b). Two-dimensional gel electrophoresis demonstrated vaccinia-virus-induced alterations consistent with translational use of misacylated tRNAs (Fig. 2c). Mass spectrometry detected Ub peptides containing Lys-to-Met substitutions, confirming the translation of misacylated tRNA$^{\mathrm{Lys}}$ (CTT) as predicted from the array data (Fig. 2d, e).

We detected virus-induced increases in misacylation (Supplementary Fig. 13) and global use of Met-misacylated tRNAs in protein synthesis (Supplementary Fig. 14) even when viral infectivity was inactivated by ultraviolet irradiation. Exposing HeLa cells, which express all human toll-like receptors (TLRs)[12], to the TLR3 ligand poly-inosine-cytosine (poly-IC, which mimics double-stranded viral RNA), or the TLR4 ligand lipopolysaccharide (LPS, derived from bacterial cell walls) also increased tRNA misacylation (Fig. 3a and Supplementary Fig. 15a). LPS- and poly-IC-induced misacylation patterns overlapped significantly with each other and with virus-induced misacylation. We obtained similar results with CpG oligonucleotides, a TLR9 ligand (data not shown). Not all immune signalling events increase [^{35}S]Met-misacylation in HeLa cells, however. Exposing cells to interferon-β or interferon-γ did not increase misacylation, although each cytokine altered the expression of cytoplasmic and mitochondrial tRNAs within 24 h of initial exposure (data not shown).

We extended these findings to mouse bone-marrow-derived dendritic cells (Fig. 3b–d and Supplementary Fig. 15b), and liver cells in a living mouse by injecting [^{35}S]Met into the portal vein (Fig. 3e). Both cell types demonstrated a level and pattern of tRNA misacylation similar to HeLa cells, firmly establishing the *in vivo* relevance of misacylation. We failed to detect misacylation after labelling cells with either [^{35}S]Cys or ^{3}H-labelled Ile, Phe, Val or Tyr (Supplementary Fig. 16; note that specific activities of other commercially available amino acids are too low to detect misacylation at greater than 0.5%). Thus, misacylation could well be limited to Met.

As viral and bacterial infections activate myriad stress response pathways in cells, we examined the ability of chemical or physical stressors to modulate misacylation. We incubated HeLa cells at 42 °C, or exposed them to the Asn-linked glycosylation inhibitor tunicamycin or the proteasome inhibitor MG132, treatments that induce an unfolded protein response through distinct pathways[13]. Although tunicamycin and MG132 increased tRNA misacylation by approximately twofold, heat shock decreased tRNA misacylation by approximately twofold (Supplementary Fig. 17). The pattern of misacylation induced by tunicamycin and MG132 was limited to subset of RNA families seen in response to viruses and TLR ligands. Allowing HeLa cells to grow past confluence, a condition known to induce stress-related genes[14], also induced misacylation (Supplementary Fig. 16b).

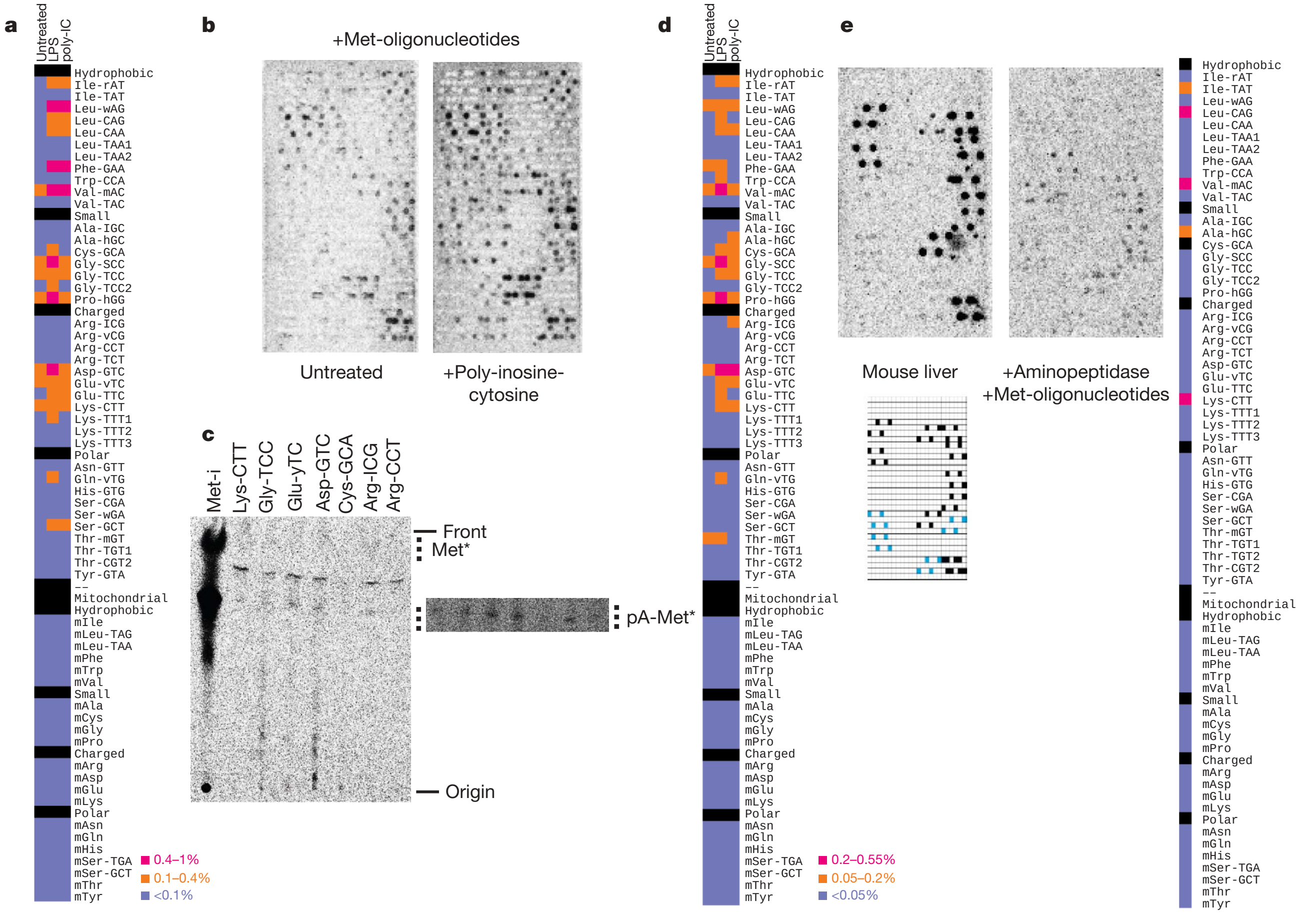


Figure 3 | tRNA misacylation induced by TLR ligands. **a**, Comparison of untreated and LPS, poly-IC-treated HeLa samples. **b**, Comparison of immature and poly-IC-matured bone marrow dentritic cells including complementary Met-oligonucleotides in array hybridization. **c**, Thin-layer chromatography of the poly-IC-matured sample using biotinylated probes. **d**, Quantitative comparison of untreated, LPS and poly-IC-matured dendritic cell samples, all aminopeptidase-treated. The detection limit of tRNA misacylation for these samples was about 0.05% for each probe. **e**, Misacylation occurs *in vivo*. Misacylation for total charged tRNA isolated from mouse liver after a 1-min pulse with [^{35}S]Met. Array key shows probe locations for Met-tRNAs (black) and Cys-tRNAs (blue).

©2009 Macmillan Publishers Limited. All rights reserved

Heightened generation of ROS by activation of NADPH oxidases is a common downstream effect of many cellular stressors[13]. It is well established that genetically encoded Met residues can act in *cis* to protect enzyme active sites against ROS-mediated damage[15], and Met protects *Escherichia coli* against oxidative-damage-induced death[7]. ROS oxidize the highly reactive sulphur in Met, which is restored to its reduced state by Met-sulphoxide reductases through NADPH oxidation[16]. We hypothesized that tRNA Met-misacylation protects cells against oxidative stress by replacing the amino acids we identified in the arrays with Met. Because tRNA misacylation is induced rapidly, this mechanism allows immediate extra-genetic incorporation of Met residues in newly synthesized proteins that provide protection against increased ROS levels.

As predicted by this hypothesis, exposing HeLa cells to ROS-inducing agents (arsenite, telluride or H_2O_2) induces Met-misacylation at high levels (Fig. 4a, b and Supplementary Fig. 18a, b). Arsenite-induced misacylation did not require protein synthesis, as it was unaffected by the addition of cycloheximide at the time of arsenite exposure (Supplementary Fig. 12c). Oxidizing agents act at least in part by increasing cellular NADPH oxidase activity[17], and in each case, misacylation was significantly reduced by diphenyleneiodonium (DPI), a broad inhibitor of these oxidases. TLR activation is known to induce ROS in neutrophils and dendritic cells[18]. Remarkably, treating HeLa cells with DPI inhibited poly-IC-induced misacylation, implicating ROS as the trigger for TLR-induced misacylation (Fig. 4b, c and Supplementary Fig. 18c). DPI also inhibited LPS and poly-IC-induced misacylation in dendritic cells (Fig. 4d, e).

We propose that Met-misacylation is a protective response to cellular stressors that increase levels of ROS. This is consistent with the recent proposal that the ROS scavenging capacity of Met selects for mitochondrial genetic recoding of AUA from Ile to Met[19]. Alternative explanations for Met-misacylation include the possibilities that it is a non-productive by-product of oxidative stress that exacerbates stress by decreasing translational fidelity (particularly because replacement of charged surface residues with Met would be predicted to increase protein aggregation), and that misacylated Met tRNAs function in cellular methylation or amino-acid transport pathways[20].

It has been demonstrated that upon mutating aminoacyl-tRNA synthetases or introducing exogenous misacylating tRNAs, *E. coli*, yeast and mice tolerate and adapt to increased errors in tRNA aminoacylation[3,21,22]. Theoretical considerations[23] support higher error thresholds for translational fidelity than those observed for tRNA aminoacylation *in vitro*. We demonstrate that mammalian cells have an intrinsic ability to modify tRNA misacylation and translational fidelity. The extent to which this ability, currently limited to Met, extends to any of the 14 amino acids yet to be examined, is an open question.

In summary, we have shown that tRNA misacylation with Met is a common and regulated event in mammalian cells. Although the full

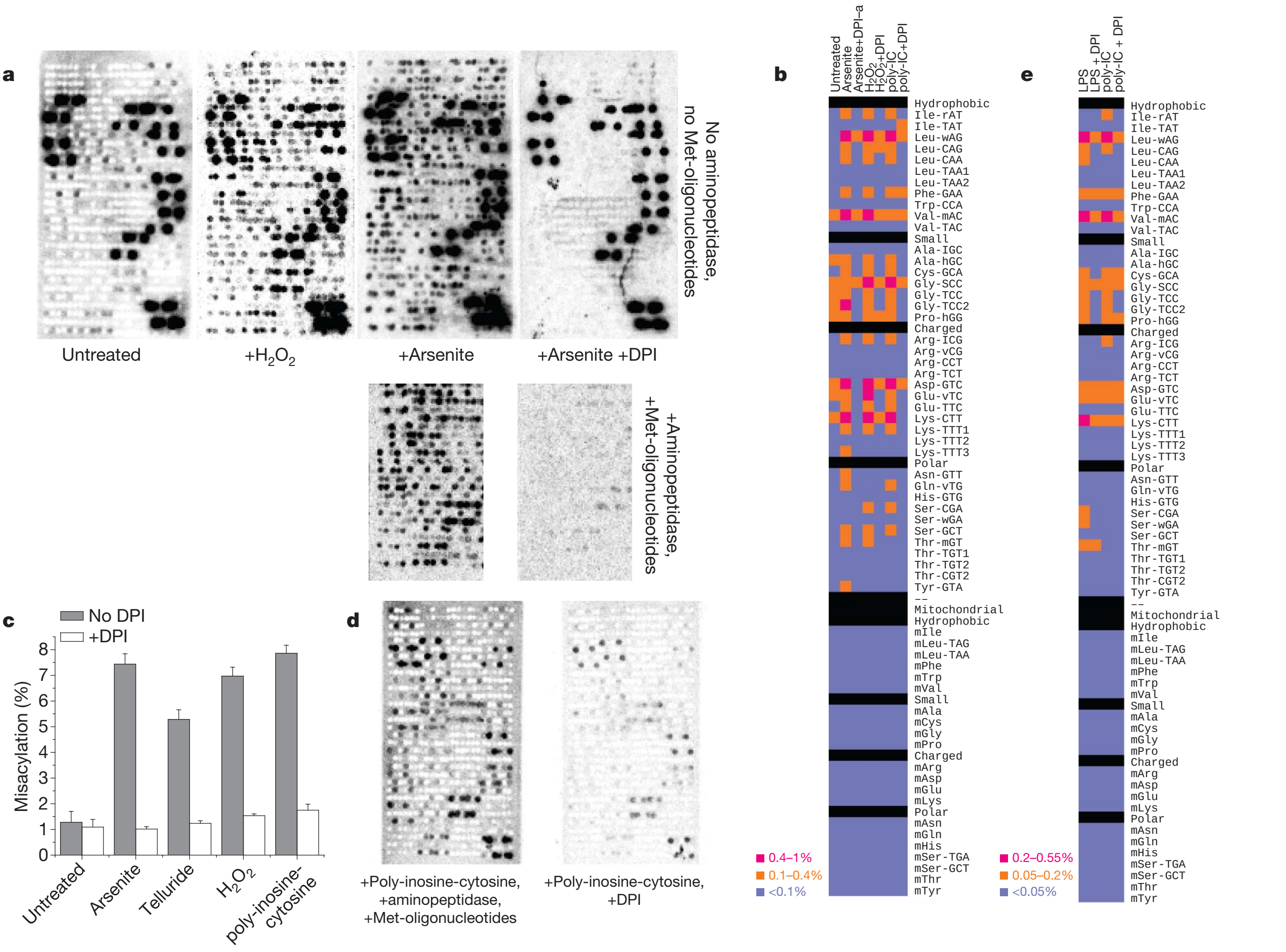


Figure 4 | Oxidative stress induces NADPH-oxidase-dependent RNA misacylation. **a**, tRNA misacylation in HeLa cells induced by oxidizing agents H_2O_2 (1 h) or arsenite (4 h). DPI inhibits arsenite-induced misacylation. **b**, Quantitative comparison of tRNA misacylation under oxidative stresses (arsenite and H_2O_2) and TLR ligand (poly-IC) with or without DPI. **c**, Percentage of all misacylated tRNAs with or without DPI when cells were treated under four conditions (100% = all Met-tRNAs). Error bars represent s.d. ($n = 2$). **d**, Poly-IC induces NADPH-oxidase-dependent tRNA misacylation in dendritic cells. **e**, Quantitative comparison of tRNA misacylation in dendritic cells with or without DPI.

©2009 Macmillan Publishers Limited. All rights reserved

implications of this phenomenon remain to be explored, there is a practical consequence: decoding mRNA into protein in living cells is not as simple as generally believed. tRNA-misacylation-based protein sequence diversity, like RNA splicing and post-translational modifications, may represent an evolutionary strategy for expanding and manipulating the information encoded by nucleic acids[24].

METHODS SUMMARY

The Methods section includes cell growth, cell treatment and stress conditions, misacylation *in vivo*, immune-precipitation of Ub-HA, two-dimensional polyacrylamide gel electrophoresis and mass spectrometry, detection of tRNA misacylation by TLC and native acid gels, pH 9 deacylation, nuclease treatment of arrays, *in vitro* aminoacylation and acylated tRNA extraction for microarrays.

Microarray. The basic features of the tRNA microarray have been described previously to determine tissue-specific differences in human tRNA expression[8]. Both array versions contain about 50 probes (42 unique) for chromosomal human and mouse tRNAs, 22 probes for human mitochondrial tRNAs and 18 probes for mouse mitochondrial tRNAs. The first array version contains 20 repeats for each probe, and over 50 hybridization control probes for tRNAs from bacteria, yeast, *Drosophila* and *Caenorhabditis elegans*. The second array version contains eight repeats for each probe and six hybridization controls for tRNAs from bacteria and yeast. The second version contains fewer repeats per probe but has higher sensitivity due to improved array printing techniques. Both versions contain four probes for chromosomal initiator and elongator tRNA^{Met} and one for mitochondrial tRNA^{Met}.

Array hybridization was performed on a Genomic Solutions Hyb4 station with 10 μg total RNA in 2 × SSC, pH 4.8 at 60 °C for 50 min. This short hybridization time was the same as the half-life of [^{35}S]Met-labelled aminoacylated tRNAs (Supplementary Fig. 19) and was necessary to minimize the amount of hydrolysis of aminoacylated tRNA during hybridization. After hybridization, arrays were washed twice each, first in 2 × SSC, pH 4.8, 0.1% SDS, then in 0.1 × SSC, pH 4.8, spun dry and exposed to phosphorimaging plates (Fuji Medicals) for up to 14 (^{35}S labels) or 34 days (^{3}H labels using ^{3}H plates). Spot intensity was quantified using Fuji Imager software.

Received 26 August; accepted 9 October 2009.

1. Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69,** 617–650 (2000).
2. Cochella, L. & Green, R. Fidelity in protein synthesis. *Curr. Biol.* **15,** R536–R540 (2005).
3. Lee, J. W. *et al.* Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature* **443,** 50–55 (2006).
4. Miranda, I., Silva, R. & Santos, M. A. Evolution of the genetic code in yeasts. *Yeast* **23,** 203–213 (2006).
5. Silva, R. M. *et al.* Critical roles for a genetic code alteration in the evolution of the genus *Candida*. *EMBO J.* **26,** 4555–4565 (2007).
6. Ahel, I., Korencic, D., Ibba, M. & Soll, D. Trans-editing of mischarged tRNAs. *Proc. Natl Acad. Sci. USA* **100,** 15422–15427 (2003).
7. Luo, S. & Levine, R. L. Methionine in proteins defends against oxidative stress. *FASEB J.* **23,** 464–472 (2009).
8. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* **2,** e221 (2006).
9. Ussery, M. A., Tanaka, W. K. & Hardesty, B. Subcellular distribution of aminoacyl-tRNA synthetases in various eukaryotic cells. *Eur. J. Biochem.* **72,** 491–500 (1977).
10. Varshavsky, A. Regulated protein degradation. *Trends Biochem. Sci.* **30,** 283–286 (2005).
11. Ibba, M. & Soll, D. Quality control mechanisms during translation. *Science* **286,** 1893–1897 (1999).
12. Nishimura, M. & Naito, S. Tissue-specific mRNA expression profiles of human toll-like receptors and related genes. *Biol. Pharm. Bull.* **28,** 886–892 (2005).
13. Malhotra, J. D. *et al.* Antioxidants reduce endoplasmic reticulum stress and improve protein secretion. *Proc. Natl Acad. Sci. USA* **105,** 18525–18530 (2008).
14. Murray, J. I. *et al.* Diverse and specific gene expression responses to stresses in cultured human cells. *Mol. Biol. Cell* **15,** 2361–2374 (2004).
15. Levine, R. L., Mosoni, L., Berlett, B. S. & Stadtman, E. R. Methionine residues as endogenous antioxidants in proteins. *Proc. Natl Acad. Sci. USA* **93,** 15036–15040 (1996).
16. Oien, D. B. & Moskovitz, J. Substrates of the methionine sulfoxide reductase system and their physiological relevance. *Curr. Top. Dev. Biol.* **80,** 93–133 (2008).
17. Chernyak, B. V. e. t. a. l. Production of reactive oxygen species in mitochondria of HeLa cells under oxidative stress. *Biochim. Biophys. Acta Bioenergetics* **1757,** 525–534 (2006).
18. Savina, A. *et al.* NOX2 controls phagosomal pH to regulate antigen processing during crosspresentation by dendritic cells. *Cell* **126,** 205–218 (2006).
19. Bender, A., Hajieva, P. & Moosmann, B. Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. *Proc. Natl Acad. Sci. USA* **105,** 16496–16501 (2008).
20. Finkelstein, J. D. Metabolic regulatory properties of *S*-adenosylmethionine and *S*-adenosylhomocysteine. *Clin. Chem. Lab. Med.* **45,** 1694–1699 (2007).
21. Ruan, B. *et al.* Quality control despite mistranslation caused by an ambiguous genetic code. *Proc. Natl Acad. Sci. USA* **105,** 16502–16507 (2008).
22. Santos, M. A., Cheesman, C., Costa, V., Moradas-Ferreira, P. & Tuite, M. F. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp. *Mol. Microbiol.* **31,** 937–947 (1999).
23. Hoffmann, G. W. On the origin of the genetic code and the stability of the translation apparatus. *J. Mol. Biol.* **86,** 349–362 (1974).
24. Freist, W., Sternbach, H., Pardowitz, I. & Cramer, F. Accuracy of protein biosynthesis: quasi-species nature of proteins and possibility of error catastrophes. *J. Theor. Biol.* **193,** 19–38 (1998).

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

Acknowledgements The authors are grateful to D. Klinman for his gift of CpG oligonucleotides and advice, A. Schilling for supervision and advice on mass spectrometry experiments, and C. Nicchitta, T. Pierson, P. Cluzel, R. Levine, A. Iwasaki and S. Amigorena for insight and advice. This work was supported by the Division of Intramural Research, the National Institute of Allergy and Infectious Diseases, and by National Institutes of Health extramural pilot projects.

Author Contributions N.N., J.M.G., A.D., K.A.D., R.B.J., J.R.S., D.B., E.M.E, M.R.R., J.S.B., A.E., B.D., S.D., H.D.H., P.B. and T.P. designed and performed experiments, and analysed data. J.S.G. generated genetic constructs. J.R.B. and J.W.Y designed experiments and analysed data. J.W.Y. and T.P. conceived the project and wrote the paper. J.W.Y. and T.P. contributed equally to this paper.

Author Information The microarray platforms are deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database under accession numbers GPL9427 and GPL9428. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.W.Y. (jyewdell@nih.gov) or T.P. (taopan@uchicago.edu).

©2009 Macmillan Publishers Limited. All rights reserved

Q&A

Next February, **Richard Olds** will begin his tenure as dean of the planned new medical school at the University of California, Riverside. The school is due to open in 2012.

How did you become an expert in tropical diseases?
I grew up with a strong international perspective on life because my father was an ambassador to the United Nations. At one point I worked in refugee camps in Europe. But after college, I didn't know what to do until an administrator at Case Western Reserve University suggested I might make an interesting doctor. I went to medical school there, and found my passion was working with tropical diseases. Working in China, Egypt and the Philippines, I was among the first generation of physicians to go beyond basic public-health strategies and also apply modern medical science to diseases in developing countries.

How do you plan to shape the new Riverside programme?
I've always been an innovator. I'm dyslexic, so I tackle problems differently. I prefer to look outside the box to find innovative ways to distinguish my programmes from others. At Riverside, I want to use those skills to develop a programme that can anticipate where the field is going, for example finding new ways to diagnose and manage diseases such as diabetes, so that I can effectively educate the doctors of the future.

How might you train doctors of the future differently?
Historically, medical-school students are taught through lectures. Although cost effective, this is the worst way to communicate information. The most effective way to learn material is to teach it yourself. I want to create a teaching culture — where everybody teaches and everybody learns. In that environment, physicians will also learn how to educate patients better about disease management, which will be crucial in the future.

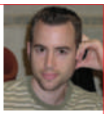
Will the ongoing health-care debate, and possible reform, affect how your programme develops?
It should. We have a tremendous opportunity to build this school to reflect any changes that occur as this country addresses health-care reform. Other medical schools will have to overcome the fact that they are organized on the basis of traditional views of how health care is delivered.

Will your international background prove useful in this endeavour?
Yes. This ethnically diverse region of California has some of the worst health-care statistics in the country and a serious physician shortage. To retain physicians here, we need to find ways to encourage members of this diverse community to go to medical school. My international experiences taught me how to reach out to communities and build strong partnerships with public-health officials and health-care providers — skills that will be needed to create the community-based programmes that are necessary to improve health. ■

Interview by Virginia Gewin

POSTDOC JOURNAL

Sharing ideas and data



Recently, a researcher asked me to provide more information about a paper of mine. I had a mixed reaction: flattery, but also hesitation. I appreciated that others were reading and enjoying it. Yet, despite being an ardent supporter of open access, I couldn't help wondering whether this researcher might find a new, exciting result that I had overlooked.

Authors are obliged to provide data and unique reagents freely to the scientific community. However, before publication, the sharing of ideas and data is played out like a high-stakes game of poker. What data do I show and what data do I hold close my chest? Everyone is vying for the highest-impact publication in a world of 'publish or perish'.

Despite the ostensible drawbacks, much is gained by sharing data with colleagues. After our last weekly lab meeting, a fellow postdoc asked me for some of my unpublished data sets to help her interpret her own data. I was more than happy to help out. A sense of collegial trust and respect assured me that we could both benefit from this exchange of data. Indeed, it allowed us both to look at our projects from a different perspective and to brainstorm new hypotheses to test.

Unpublished data shared judiciously at conferences is also beneficial to the scientific community and ultimately helps to advance science. Nevertheless, filling these conference halls are hundreds of poker players, mulling over their respective hands. ■

Bryan Venters is a postdoctoral fellow at the Center for Eukaryotic Gene Regulation at Pennsylvania State University, University Park.

IN BRIEF

NIH asked to probe ethics

One hundred health-care and policy executives and professionals are asking the head of the US National Institutes of Health (NIH) to fund studies on medical ethics and conflicts of interest in medicine. In a 17 November letter to NIH director Francis Collins, the multinational group warns that relationships between industry and academics, medical educators and clinicians are flawed. The signatories seek to learn the extent to which commercial bias compromises medical and health-care information, and to identify appropriate interventions. A spokesman for the NIH says that the agency has not formally received the letter and could not comment.

Disease threat assessed

The US Wildlife Trust is looking to support up to 40 postdocs and researchers in an effort to find and predict diseases that move between wild animals and humans. The non-profit trust is assembling eight teams of scientists to work in Asia and South America to detect disease hot spots and determine how to respond. The effort, funded with about US$15 million over five years, is one of five initiatives of the US Agency for International Development (USAID), collectively known as the Emerging Pandemic Threats Program. The project builds on USAID's monitoring of the H5N1 influenza virus in wild birds to address the broader role of wildlife in emerging human diseases.

Asia takes clean-tech lead

The clean-technology triumvirate of China, Japan and South Korea has already surpassed the United States in producing most of the world's clean technology, according to a report. *Rising Tigers, Sleeping Giant* predicts that the three nations will also grab most of the sector's available private equity in the short term if nothing changes. Released on 18 November by two Washington DC think tanks, the report also finds that the United States lags far behind the three Asian nations in terms of federal funding and legislation to support research and production in most clean technologies. China will invest $397 billion over the next five years in clean technologies, the report says, compared with just $172 billion in the United States.

© 2009 Macmillan Publishers Limited. All rights reserved

IMAGES.COM/CORBIS

START UP AND SUCCEED

Scientists looking to capitalize on their latest discovery might consider starting a company. But that's more complex than it may seem, as **Karen Kaplan** reports.

Andy Richards, a biotechnology entrepreneur based in Cambridge, UK, took a circuitous route to business success. Armed with a doctorate in chemistry in the mid-1980s, he joined an early iteration of what is now the London-based international pharmaceutical firm AstraZeneca. But he quickly realized that political and commercial interests could dominate outcomes. Richards left for a science consulting firm in Cambridge, to help scientists turn their ideas into companies.

In 1991, Richards and a group of colleagues founded their own start-up, a small-molecule firm called Chiroscience. It turned out to be a mammoth undertaking, but one that he is glad he pursued. The fledgling company took off in a series of different directions, moving from solving chemistry problems to redevelopment of existing drugs to small-molecule drug development to novel biology. He and his partners even dabbled in genomics, at one point buying a genomics company from Bill Gates. Then in 1999, after the original company had swept through an array of incarnations, Richards and his partners sold it. "By that time, I was knackered," says Richards, who had promised his wife he would take a year's holiday. Three months later, however, he had invested in four life-sciences companies and was interim chief executive of two. Today, he is an 'angel' investor, putting his capital into very early-stage start-ups and helping set them up. He sits on the board of 10 such firms and has holdings in 15.

Seductive scenario

For those who hope to turn a specialized research niche into a widely sought-after product or drug, venturing off to start one's own business may sound wildly seductive. But launching a start-up based on a research discovery is rarely easy or lucrative, especially in the early stages.

There are some non-negotiable requirements for translating an idea into a successful money-making business. These include a sound concept for a marketable product; a business plan; seed and start-up funds; a partner; and a business-savvy network to make up the business team.

Although the business-launching process may seem neatly linear — one step progressing to the next, and so on — the process is often chaotic, involving many steps that must often be taken simultaneously. The whole undertaking can be daunting, especially for a scientist who may have little or no experience in the business world. But it can be done, with the right support and assistance.

> "You need to have a clear idea of what you're going to sell and who you're going to sell to."
> — William Bains

The first steps are fundamental to any business: find a product and identify a market. "You have to formulate your commercial idea or concept, and then you have to figure out how your idea is going to become a business," says William Bains, a life-sciences entrepreneur and consultant in Cambridge who has launched a number of successful science-based start-ups. "You need to have a clear idea of what you're going to sell and who you're going to sell to."

The aspiring entrepreneur needs to think through how the discovery can be brought

© 2009 Macmillan Publishers Limited. All rights reserved

to market. How, for example, might a newly discovered hypertension biomarker best be used? Would it be incorporated into a test? And how and where would such a test be used?

Revisiting and rethinking strategies is key. Suppose budding entrepreneurs realize that executing their idea would require an entire automated analytical machine to run the test. That is likely to be impractical — it could cost millions of dollars to implement, for a market that may not be very large. The next step is attempting to refine the idea. Might it be possible to incorporate the test into an existing system? Should it be used in a general practitioner's office? In a hospital? In an ambulance? By paramedics? The idea may have potential, but it needs the right approach.

William Bains (left) and Daniel Behr: seek appropriate support.

All this information should then become part of the start-up's business plan — minus the technical specifics, in order to protect it from idea theft. Here, the aspiring entrepreneur assesses the business venture's economic viability, describes and analyses its prospects and customers, and evaluates the need for a patent to protect the fledgling business's intellectual property.

Some start-up business owners skip this step, but creating the plan can be crucial, in a number of ways. It helps the entrepreneur to define his or her aim. Writing everything down can uncover potential problems or weaknesses. And the document can be used as a selling or marketing tool for both investors and those who can help to launch the venture.

Scientists planning to set up a business without the assistance of their university or institution will need seed funds. How much depends on the business idea and how it will be developed. Some prefer to launch a start-up without seeking external funding, contributing their own savings along with contributions from friends or relatives willing to support the idea.

To save on upfront expenses, early-stage business start-ups should consider joining a business incubator, suggests attorney Michael Shuster of law firm Fenwick & West in San Francisco, California, who specializes in intellectual-property law. Because businesses share space in the incubator building, costs for lab space as well as for phone and Internet service are minimal. An incubator also offers myriad business-support programmes, including providing referrals to potential partners and helping with regulatory compliance and intellectual-property management — all key components for a start-up launch.

Often it is best not to go it alone. A business partner and a team of people who can help with all stages of the launch are great assets. "Doing a start-up on your own is almost impossibly hard work," says Bains. "I've tried to be the sole founder of a company, and on the thirtieth night with no sleep, you realize it may not be the best idea."

Unlike a partner, team members don't have to own any of the business. But the right team can take care of the business details many scientists are likely not to be as good at, such as crafting the business idea, identifying a market, writing a business plan, assessing the need for a patent and securing investors.

Some institutions permit students to do such work as part of their graduate student programmes; others set up internships. They don't have to be science students. University business students may make good team members, says Shuster.

Those pursuing ideas outside their institutions must be careful to avoid any potential conflict. They must keep all research related to the new venture off campus and make sure they do not pay for any spin-off expense with grant funds received while they've been at the university.

"Success isn't about finding the best idea. It's about doing something with it."
— Andy Richards

Patent or not?

Ideas should be patented early on, especially in the case of a specific product such as a drug compound. Without early protection, ideas could unwittingly be disclosed and possibly stolen, thwarting any later efforts to secure the patent. Under patent laws worldwide, disclosure refers to anyone who isn't under a duty of confidentiality, even if the entrepreneur is only chatting about the idea to a friend over coffee. Alternatively, the aspiring entrepreneur can draw up a confidentiality agreement and protect the idea by having all contacts agree to keep the idea under wraps.

But patents are costly. Just preparing an application costs between $7,500 and $15,000 in the United States, depending on the complexity of the application. A US patent expires 20 years after the date of application, meaning that if it takes several years for the patent to be issued, the invention or business idea could have significantly less than 20 years of protection. The price tag for pursuing broad international protection is even higher, sometimes as much as $100,000. Even if the patent is denied, that information becomes public. It's too late to guard the intellectual property: the patent office in the nation where the application was filed publishes the information.

The university route

Entrepreneurs can avoid many of these obstacles, however, if they set up the new company in conjunction with their university — an approach that offers distinct advantages. University technology-transfer offices often handle everything from helping the entrepreneur craft the business idea to developing the business plan to funding the commercialization of the product to securing and paying for a patent.

That's the case at Harvard University in Cambridge, Massachusetts. According to Daniel Behr, director of business development at Harvard, any principal investigator with a faculty position who launches a start-up under the university's auspices can have an advisory position in that company and can own company equity. The researcher has to sign a statement confirming that he or she has an economic interest in the company and agreeing that the company may be restricted from providing funds to him or her.

Because the intellectual property, and possibly some of the company equity, are owned by the university, all revenues generated by the spin-off must be distributed by the university under its own formula. Harvard researchers receive 35% of licensing revenues or stock and the researcher's lab receives another 15%, says Behr. The researcher's school or department receives 35% and the remaining 15% goes to the office of the president.

Different institutions may offer less to the researcher, the lab or department. "Distribution varies widely from university to university," says Jon Soderstrom, immediate past president of The Association of University Technology Managers, a US organization that promotes technology transfer from academia to industry.

Developing a start-up into a thriving business can be a bumpy road whether it's done independently or through one's institution. But the best ideas will never sprout a business if researchers don't work to develop them. "Ideas are cheap and there are lots and lots and lots of them," says Richards. "But success isn't about finding the best idea. It's about doing something with it." ■

Karen Kaplan is assistant editor of *Naturejobs*.

© 2009 Macmillan Publishers Limited. All rights reserved

The imitation game

Being human.

Shelly Li

Subject C stared at the blank wall in front of him as he reprocessed the answers from Subject A and Subject B, comparing the octaves of their voices, the strain patterns of their words.

For all practical purposes, Subject C's programmed name was Turing: a sixth-generation robot, one of the most advanced of his kind.

"I'm sorry, Subject A," Turing said to the left side of the wall. As Dr Conway had disabled his X-ray vision for the test, he couldn't see Subject A, but his inner tracking system could still locate the voice. "Would you repeat the answer to my last question. How is your relationship with your wife?"

Turing increased the power of his inner sound reception just in time to hear the soft scoff that he wouldn't have caught otherwise.

In a woman's voice, Subject A repeated: "I don't have a wife. I'm a woman. Not that I don't approve of gay people, mind you. I'm a Democrat, and it's 2065, for Christ's sake — half my graduating class swung both ways."

Turing frowned. *Swung both ways*. He had been confused about the phrase the first time it was said. "What does the phrase 'swing both ways' mean?"

"Oh." There was a light chuckle. "It means that a person is, uh, attracted to members of both sexes."

"I see." Turing recorded the definition and filed it away.

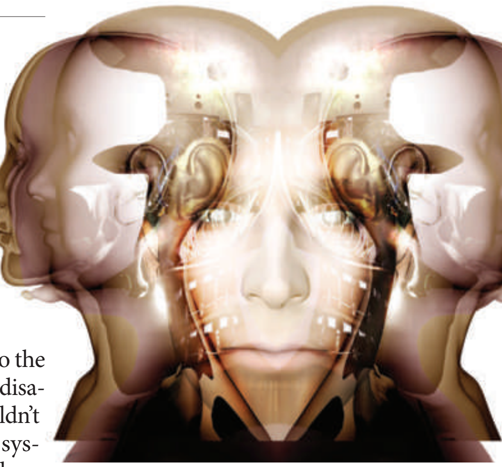
He then sat back in his chair and evaluated. Subject A had to be the woman, and Subject B had to be the man.

But Subject B is a convincing woman too, Turing thought.

He had to be careful of which gender to place on which subject. Dr Conway had warned him to weigh the answers he received from the two subjects. One person was lying on every question, whereas the other always told the truth.

Inside the viewing area above the testing room, Dr Conway watched from behind a one-way mirror as Turing struggled to choose. The programmer on Conway's right monitored Turing's processing system, as well as the robot's emitted levels of artificial intelligence.

As Conway stood there, seeing Turing on one side of the wall and the two subjects on the other, he couldn't help feeling a twinge of pity for the unknowing robot.



JACEY

Before starting the test, dubbed 20 years ago by Conway's mentor as 'the imitation game', Turing had given Conway a promise to pass the test, and had said it with such determination in his human-looking eyes. Conway had to admit: he had never seen so much emotion behind a robot, so much fire and passion.

Conway's mentor, Dr Maigney, had always had to remind him that the robots would never have a 'mind'. Sure, by the sixth generation, robots looked so real that you couldn't distinguish them from human beings. Robots could talk slang — they could even tell jokes. But that didn't mean they had thoughts of their own. All they had was artificial intelligence, words and actions that they had learned by mimicking humans.

As a scientist, Conway still had trouble with these facts.

The imitation game continued.

"Subject B," Turing said, his eyes flickering to the right of the wall. "Do you hold a job?"

"I work part-time as a dental assistant." The womanly voice was steady, even.

"And what do you do with the rest of your time?"

"I clean the house, cook, take care of my kids … you know."

Being a robot that had never left the factory, Turing didn't, but that was a trivial matter. He asked Subject B: "Do you have a husband? Or a wife? Perhaps you swing both ways?"

Subject B giggled. "My kids have a father who sends us plenty of money every month." Then the frequencies in the voice faltered. "But no, I don't have a husband."

"Hmm." Slowly, a smile spread across Turing's face. He had the answer now.

"Dr Conway," said the programmer who monitored Turing's artificial intelligence. "His AI level just peaked. He's got the answer."

Conway said: "Then let's hear it."

Turing's chair scraped softly across the concrete floor as he turned towards the one-way mirror. "I know who is the man and who is the woman."

"Tell us, Turing," Conway's voice echoed through the room.

Turing's conviction made his heart race with an unsteady beat. Nevertheless, he stood up and said: "Subject B is the woman. Subject A is the man."

Turing hadn't looked to see which subject displayed womanly qualities, and which subject resembled a man — honestly, he didn't know many differences in the characteristics of the two genders. But to Turing's ears, there was a pure tone of vulnerability in Subject B's voice, something that had moved the robot. There was no doubt that Subject B had been telling the truth.

But whether or not Turing's instincts were right would depend on Dr Conway's next words.

After a few moments, Conway said: "Congratulations, Turing. You have passed the test and will be granted human status."

Focusing on the triumphant smile lighting up the robot's face, Dr Conway tuned out the murmuring of programmers throughout the viewing area.

But there was one programmer whose words wormed into Conway's head. He was only a kid, maybe two or three years out of college, and no doubt he was new to the programme.

"Sir," the young man said. "Subject A and Subject B are both computers that we programmed with personalities, nothing more. The imitation game is unwinnable."

"That's correct."

"Then Turing is nothing more than a robot, albeit a very advanced one." The young man peered up at Conway with confusion. "How can you lie to him?"

Conway tore his eyes off Turing and turned to the programmer. With a smile on his face and a churning in his stomach, Conway answered with the same advice that Dr Maigney had given him three years ago. "Because in this world, belief is the only thing that matters." ■

Shelly Li lives in Omaha, Nebraska, and writes science fiction and fantasy. Visit her website at www.shelly-li.com

Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a

© 2009 Macmillan Publishers Limited. All rights reserved