nature

www.nature.com/nature

Vol 453 | Issue no. 7197 | 12 June 2008

To thwart disease, apply now

Translational medicine is a key addition to the biomedical research enterprise. Policy-makers and research leaders now must build the infrastructure to take discoveries from the bench to application.

Molecular biology is a victim of its own success. Seemingly overnight, it has changed from a science of one gene, one protein, one molecule, one at a time, to all genes, all proteins, all molecules, all at once. Everything is now 'ome-sized. The generation and exploration of these data has become a massive, all-consuming discipline.

And yet the expected pay-off — the new therapies and diagnostics that will improve human health — has not kept pace. Researchers and funding agencies recognize this inequality and are working on a solution: translational research. The name encompasses the strategies by which the intellectual riches flooding from biomedical discovery can be converted into practical riches from which humanity can benefit. *Nature* strongly encourages this effort, which is highlighted in a number of articles in this issue (see page 839).

It is clear, however, that the success of translational research requires new experts, infrastructure and incentives. This is perhaps best explained from the perspective of a basic researcher who has a discovery that is ripe for translation. A common and mistaken assumption is that this scientist will have the motivation — and somehow acquire the luck and skill — to make translation happen with little help. In practice, this is often a recipe for failure. The relative handful of scientists who have found applications for their discoveries frequently say they did so despite the system, not because of it.

Team work

Instead, imagine that the researcher enters seamlessly into a team that is ready-made for translation. That team features experts in all aspects of clinical research, including medicine, pharmacology, toxicology, intellectual property, manufacturing, clinical-trial design and regulation. The basic researcher now has the back-up from those who can do the jobs for which he or she is unqualified.

> "The scientists who have found applications for their discoveries often say they did so despite the system, not because of it."

The team has immediate access to the necessary infrastructure for toxicology testing, for instance, or to facilities that meet 'general manufacturing practice' standards for making compounds for use in human trials. The sole aim of the team is to ensure that the researcher's discovery will help the most people possible — not to ensure that it fetches the best price when it is licensed. And, absolutely crucially, all the team members are rewarded in terms of pay and promotion just as richly as if they had produced a string of publications.

In this ideal world, translational research becomes a welcome and satisfying pursuit for all concerned, rather than a distraction or a burden. It would receive much more than the lip service that it is sometimes afforded on grant applications.

Such an overhaul of infrastructure and reward systems will take commitment and money. For example, the US National Cancer Institute's Rapid Access to Intervention Development programme, which provides such a service (massively oversubscribed), spends between US$2 million and $7 million to take a discovery through to its first clinical trial. The cost of performing those trials is even more. But institution heads and policy-makers should not baulk at the price: it is small compared with the potential benefits of more discoveries being put into practice, and more illness being treated or averted.

Two-way process

Some researchers complain that an emphasis on translation swings the pendulum too far towards applied science at the expense of basic research, but this concern has little foundation. In fact, what is worrying is the extent to which biomedicine in the past few decades has swung so far towards pure science. One attempt to stem that tide is noted in the first of a series of Commentaries on innovation (see page 853). Yes, many of the most successful medical advances came from serendipitous discoveries. But money should also be spent on creating the mechanisms to translate those findings when they are made.

This distinction between basic and applied research dissolves under close scrutiny. Translation is not a one-way progression in which research findings enter a production line and emerge at the end as drugs or diagnostics. The whole process is more fluid: experiments on human tissues and clinical trials can inspire fascinating new questions back at the bench that will, when answered, improve the human experiment in its next iteration (see page 843). That is why efforts at translation should focus on training personnel to be comfortable in this eddy of information as it circulates between the bench and the clinic.

Under the direction of Elias Zerhouni, the US National Institutes of Health is spending several hundred million dollars to set up translational research centres across the United States (see page 840). At this early stage, it's not clear whether this represents a game-changing commitment by the agency and the receiving institutions, or an attempt to mollify tax-payers who want a return on their heavy investment. Many of the aims are the right ones, and anything is better than the current situation in which an individual's publications trump real medical needs. But Zerhouni's successor in the next presidential administration must make it a priority to continue or surpass these efforts. Because translational research is a new and unproven discipline, with no 'how-to' manual, it is also important to evaluate each attempt at translation as the field takes shape.

What would a scientist looking back on a long career be most proud of achieving? For some it might be the solution to a fundamental cellular mystery, such as the way a cell's division is usurped by cancer. Better still might be that solution, and the knowledge that thousands of cancers were no longer dividing because of it. ■

The price isn't right

ITER will cost more to build than previously thought. Now is the time to be honest about how much.

Quoting a price for a major new scientific instrument is notoriously tricky. Researchers have to estimate costs for equipment that has never been built, forecast expenditures years in advance, allow for unknown contingencies, and win approval from sceptical politicians who always want the project to cost less.

So it is not a complete surprise that a recently finished design review of ITER, a major fusion experiment to be built in Cadarache, France, is forecasting a delay of 1–3 years in its completion date and a roughly 25–30% increase in its €5-billion (US$7.8-billion) construction cost (see page 829).

The seven international partners in ITER (the United States, the European Union, Russia, China, Japan, India and South Korea) will no doubt be displeased by the news. They reached a final agreement to go ahead with ITER in 2006 based on a partially incomplete 2001 design, and may well suspect that the scientists were deliberately quoting an over-optimistic price in order to sell the project.

Whatever truth there might be in that allegation, the fusion community was making its estimate under less than ideal circumstances. ITER had been something of a political football since 1985, when it began life as part of the cold war détente. The collapse of the Soviet Union began a decade of political limbo for the project. Scientists had to radically downsize it at the end of the 1990s to appease the budget concerns of skittish member states.

As international partners came and went (and, in the case of the United States, came again), ITER subsisted on a shoestring. Meanwhile, politicians fought over the project's location. Until that debate was settled in mid-2005, only limited revisions to the design could be done. The redesign has been a top priority for the new ITER team ever since, and the group should be commended for coming forward with a higher estimate of costs after the full review.

What is worrying is that even this new price tag might not reflect the true cost of the machine. Crucially, it does not include the soaring price of commodities such as steel and copper, which are used in large quantities in the giant reactor. The ITER team claims that these costs can be excluded because individual member states will contribute finished components rather than raw materials, but this seems disingenuous. Already, the US government has doubled its estimated maximum contribution to the project, and other countries will probably have to follow suit.

"ITER may yet follow the path of other projects whose costs spiralled out of control"

This suggests that ITER may yet follow the path of other projects whose costs spiralled out of control once they were given a political imprimatur. The danger to the project itself may seem to be limited because of its international nature, but strictly speaking there is nothing to prevent a cancellation of the sort that ended the US Superconducting Supercollider. Congress halted that experiment 15 years ago, even as the tunnels were being dug in Waxahachie, Texas.

The more likely outcome is that overruns will further undermine the credibility of science at a time when it is increasingly dependent on multinational collaborations to build instruments and data networks. Future projects such as the International Linear Collider, a next-generation particle accelerator for high-energy physics, may well face more sceptical funders if ITER's costs aren't contained.

The independent scientific and management advisory committees overseeing ITER should take a hard look at whether the latest estimates are truly realistic. If they are not, then the committees should demand that the budget include adequate contingencies for factors such as increased energy and commodity costs, as well as scenarios for construction with less than full funding. Even if it means more pain in the short run, this kind of discipline will ultimately lead to a better machine and a better future for all international collaborations. ■

Open to interpretation

The use of 'dignity' as the foundation for an ethical law in Switzerland is compromising research.

The law introduced by Switzerland in 2004 to protect the dignity of animals, plants and other life forms is now in conflict with the country's research agenda. Two top Swiss universities have been forced to appeal to the supreme court in a bid to secure the right to perform perfectly reasonable experiments that have been banned because they are said to offend the dignity of the non-human primates involved. The problem in this instance lies in an interpretation of the law that flies in the face of research reality (see page 833).

The Swiss law is at odds not only with beneficial research but also with good sense. Even plant scientists potentially face restrictions on the kinds of genetic engineering they are allowed to do (see *Nature* **452,** 919; 2008), and debates have arisen about the abuse of dignity in decapitating wild flowers.

Although pondering the dignity of dandelions is downright silly, the underlying problem with the Swiss law is that it allows rules to be built on the foundation of a notoriously subjective concept. In March, the US President's Council on Bioethics produced a collection of 28 essays on the dignity of human life and proved unable to come to a consensus. The essays offer statements on the concept that are often contradictory: dignity is earned, but it is also shared by all in full measure. Dignity cannot be taken away — yet it can and has been in cases of slavery. Harvard psychologist Steven Pinker wrote a powerful critique of the panel's efforts in the 28 May issue of *The New Republic*, arguing that 'dignity' has been widely misused to mean whatever conservative bioethicists want it to mean.

Dignity as a concept cannot be a director of moral judgement. After all, when the Swiss government assigned dignity to plants, it was in essence assigning autonomy, as if a plant sways in the breeze because it has decided to do so. A strict interpretation of such a framework would prohibit agriculture. But even where such rules are not absurd, laws should not be based on such a slippery concept. ■

RESEARCH HIGHLIGHTS

Remember this

J. Exp. Biol. doi:10.1242/jeb.018531 (2008)
The nautilus, the archaic cousin of octopus, cuttlefish and squid, has surprisingly good powers of recollection even though it lacks the dedicated brain regions for learning and memory that other cephalopods have.

Robyn Crook and Jennifer Basil at Brooklyn College in New York trained a dozen *Nautilus pompilius* (pictured) to associate the smell of fish with a blue light, demonstrated by the creatures eventually extending their tentacles merely in response to a flash. The nautili could remember this training for up to 24 hours, and their recall within the first hour proved as good as that of cephalopods with far more complex neuroanatomies.

The discovery could lead to a more complete picture of brain evolution.

D. PERRINE/NATUREPL.COM

NEUROSCIENCE

Shiver stoppers

Cell Stem Cell **2,** 553–565 (2008)
Human glial progenitor cells can differentiate into cells that make the nerve cell insulator myelin, and it now seems they can be used to treat a neurological disorder in mice. The cells come from the white matter of the brains of second-trimester human fetuses.

A team led by Steven Goldman at the University of Rochester Medical Center in New York inserted about 300,000 human glial progenitor cells into the brain of each of 26 'shiverer' mice — which cannot make normal myelin — shortly after birth. Four of the treated animals lived for more than a year. Post mortem revealed well-myelinated neurons. All of the control mice died before they reached 21 weeks of age.

ZOOLOGY

The skin off my back

Biol. Lett. doi: 10.1098/rsbl.2008.0217 (2008)
The mothers of many species give their all to rear their young. But the South American amphibian *Siphonops annulatus* takes this dedication to an unusual extreme by allowing her offspring to eat her skin.

Alexander Kupfer of the Friedrich-Schiller University in Jena, Germany, and his colleagues report that *S. annulatus* nestlings have 44 spoon-shaped teeth. These teeth bear claw-like protrusions that the nestlings use to tear away the outer layer of their mother's skin.

The same group has previously described 'maternal dermatophagy', as such skin-feasting is called, in an African amphibian, *Boulengerula taitanus*. It therefore seems likely that the behaviour arose before Africa and South America separated, more than 100 million years ago.

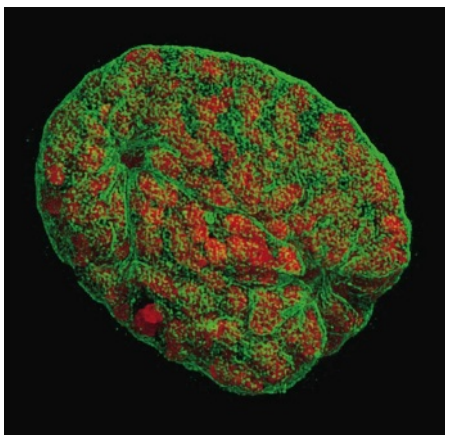
MICROSCOPY

Laser focus

Science **320,** 1332–1336 (2008)
Cellular structures as small as 100 nanometres can be viewed in three-dimensional (3D) colour images thanks to a technique that doubles the resolution of fluorescence light microscopy.

The technique illuminates samples with three interfering beams of laser light, enabling it to circumvent the resolution limit of traditional light microscopy that is set by the wavelengths of visible light.

John Sedat at the University of California, San Francisco, and Heinrich Leonhardt at the Ludwig Maximilian University of Munich, Germany, and their colleagues developed this '3D structured illumination microscopy' and used it to construct the first 3D colour image of a nuclear pore and its environment. The picture (below) shows a mouse nucleus with condensed chromosomes (red), surrounded by a fibrous network called the nuclear lamina (green).



L. SCHERMELLEH & P. CARLTON

MOLECULAR BIOLOGY

Lost in transcription

Nature Methods doi:10.1038/nmeth.1226; doi:10.1038/nmeth.1223 (2008)
High-throughput sequencing techniques have been harnessed to catalogue the messenger RNA (mRNA) molecules produced by many types of mouse cell.

The catalogue is known as the transcriptome. Barbara Wold and her colleagues at the California Institute of Technology in Pasadena sequenced mRNA from mouse brain, liver and skeletal muscle. They found that about 3,500 mouse genes are alternatively spliced — that is, the initial mRNA sequence can be chopped up and put back together in various ways to form different mRNA sequences.

Meanwhile, Sean Grimmond at the University of Queensland in St Lucia, Australia, and his collaborators sequenced mRNA from mouse embryonic stem cells. The results should help elucidate pathways controlling embryonic stem cells' ability to develop into any cell type.

ANIMAL BEHAVIOUR

Best and brightest

J. Avian Biol. **39,** 277–282 (2008)
Most nestlings have cryptic plumage to reduce the odds that predators will see them, but chicks of a few species are

brightly coloured. Why has been a mystery.

Ismael Galván of the National Museum of Natural Sciences in Madrid and his colleagues painted the yellow feathers of great-tit (*Parus major*) chicks with a marker pen that reduces the ultraviolet reflectance of a surface. They measured the difference in the chicks' tarsus length — a method used to judge growth rate — over a three day period and compared the results with a control group. The chicks with normally reflective feathers had grown more.

The authors propose that maintaining this ultraviolet reflectance might be a sign of how fit a chick is, and thus determine which the parents feed most.

NANOTECHNOLOGY

Golden advance

Adv. Mater. doi:10.1002/adma.200703026 (2008)

Engineers at the University of Arizona in Tucson have developed a way to make metal nanoparticles that human kidneys should be able to eliminate. This would render the nanoparticles suitable for biomedical imaging with technologies such as optical coherence tomography.

Marek Romanowski and his co-workers used liposomes as templates to prepare hollow, gold spheres 63 nanometres in diameter that can scatter light of preselected wavelengths. The resultant balls of gold and lipid have different optical resonances from that of pure gold structures of the same size — and within the range of visible light.

The nanoshells break up to form 5.7-nanometre-dots when their lipid cores are degraded, which should happen inside the body. This would make the gold nanoparticles small enough to be cleared by renal filtration.

PHOTONICS

Cancer zapper

Nature Photon. doi:10.1038/nphoton.208.100 (2008)

A international team has found a way to minimize collateral damage from photodynamic drugs. Doctors administer the photodynamic drugs before exposing tumours or other diseased tissues to laser light, causing diseased cells to perish.

Harry Anderson at the University of Oxford, UK, Brian Wilson at the University of Toronto in Canada and their colleagues tested a new class of compounds that become toxic only when struck by two photons arriving almost simultaneously. This means that very few cells outside the most intense part of the laser's focus are affected. One of the new compounds proved effective at closing-off blood vessels by killing the cells lining them.

GEOPHYSICS

Mysterious mountains

Geology **36,** 495–498 (2008)

Why do mountains arise in the interior of continents, far from the edges of tectonic plates where deformation — and thus mountain building — is expected?

To answer this question, Scott Dyksterhuis and Dietmar Müller of the University of Sydney in Australia modelled the stress regimes that have helped push up the Flinders Ranges and other nearby mountain belts in the middle of the Australian plate.

They concluded that the plate interior can be affected by forces at the plate edges thousands of kilometres away — a finding that could help explain deformation in the middle of other tectonic plates.

D. WALL/ALAMY

PSYCHOLOGY

Not fair!

Science doi:10.1126/science.1155577 (2008)

If someone treats you badly and you retaliate, blame serotonin. Lower levels of this neurotransmitter make people more likely to retaliate when they perceive others to have breached the maxim 'treat others as you wish to be treated', find Molly Crockett at the University of Cambridge, UK, and her co-workers.

The team temporarily lowered serotonin levels in 20 volunteers and had them play the part of responder in the 'ultimatum game'. The responder can either accept the division of a sum of money offered by the game's proposer, in which case they both get their share, or reject it and deprive both players of the amounts proposed.

Although mood remained unchanged when players' serotonin levels were lowered, they were more likely to reject unfair and very unfair offers, defined as 30% and 20% of the stake, respectively.

JOURNAL CLUB

David Beerling
University of Sheffield

A palaeobiologist calls for greater biological realism in climate models.

The world's most sophisticated climate models fail to adequately replicate climate at high latitudes and over continents' interiors during ancient periods of greenhouse-gas-induced warming: the wintertime predictions are consistently too cold. This makes me worry that the field is missing fundamental feedback processes that amplify warming. If so, climate models might be underestimating how much anthropogenic warming will happen in the future.

What might these mysterious processes be? Lee Kump and David Pollard of Pennsylvania State University in University Park think they have found one. They propose that marine phytoplankton that emit dimethylsulphide — already recognized as a major source of cloud-seeding particles far out to sea — became thermally stressed during the Cretaceous period (100 million years ago). As a result, the phytoplankton grew more slowly and reduced their emissions. Fewer biologically derived aerosol particles meant fewer nuclei for cloud condensation, which, in turn, led to less extensive cloud cover and more transparent clouds. Solar radiation was thus reflected less, and polar temperatures rose by 10–15 °C (L. R. Kump and D. Pollard, *Science* **320,** 195; 2008).

Kump and Pollard's work is exciting for its dramatic result. Nevertheless, the duo's findings are ultimately unsatisfactory; the effects of heat on biological aerosol emissions need to be better described in their model for it to generate really solid conclusions. Although some recent field and laboratory experiments do suggest that marine algae produce less dimethylsulphide when carbon dioxide concentrations approach those of the Cretaceous, much more research is needed. If such results agree with Kump and Pollard's assumptions, I might worry less about climate models — but maybe even more about global warming.

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

NEWS

Egg shortage hits race to clone human stem cells

US stem-cell researchers are calling for changes to state laws that prohibit compensating women who donate eggs for research. The laws, in leading stem-cell research states such as Massachusetts and California, are crippling the promising field of 'therapeutic cloning' that could produce useful embryonic stem-cell lines for studying various human diseases, they say.

Therapeutic cloning involves transferring the nucleus from an adult human cell into a human egg with the intent of creating stem cells from the resulting embryo. Such cells could potentially serve as a therapy genetically matched to the person who donated the adult cell. Although no one has yet derived these cells, one group has now reported creating cloned human embryos in this way. But the procedure uses many eggs. Researchers say that because many states do not allow women to be paid for their egg donation, potential donors are put off and choose lucrative fertility clinics instead.

It took Kevin Eggan and Douglas Melton, of Harvard University's Stem Cell Institute, two years and US$100,000 in local advertising to secure a single egg donor (see 'How close, how comfortable?') for their attempt to develop embryonic stem-cell lines to model diseases such as amyotrophic lateral sclerosis. The group, which obtained fewer than ten eggs, completed its experiments early this year, Eggan announced last month at a meeting of the Society for the Study of Reproduction in Kona, Hawaii. They are not yet ready to discuss results.

Eggan says that although Harvard's highly publicized egg-donation programme received plenty of responses, many women decided against donating for research when they were informed that a fertility clinic would compensate them for the same procedure. The extraordinary efforts needed to recruit donors such as this "unique woman" would not be required, says Eggan, if rules restricting the compensation of donors were relaxed.

Fertility clinics in the United States are allowed to pay for eggs. Some advertise remuneration in the tens of thousands of dollars. "It's patently unfair," says Sam Wood, head of Stemagen, a biotech company based in La Jolla, California, which uses eggs leftover from fertility procedures at an associated clinic. In January, Wood's team reported the first cloned human blastocysts. These early-stage embryos were created from 29 egg cells retrieved from three donors (A. J. French *et al. Stem Cells* **26,** 485–493; 2008).

Stemagen attains the consent of both donor and recipient, but offers no compensation beyond that offered for the reproductive egg donation. This 'uncompensated shared cycle donation' model has "worked quite well", Wood says, but a compensation model would be fairer to donors. "For virtually any other type of biomedical research you are allowed to compensate fairly."

US regulations vary from state to state — New York is currently debating its laws on egg payment. Some states allow for reimbursement of lost wages, expenses or child care during the procedure, which can take 60 hours, including initial screening, consultations, hormone injection and surgical removal of oocytes. Other countries, including the United Kingdom, allow compensation in different forms, such as offsetting the costs of fertility treatment in return for research material.

AJ PHOTO/SPL

The US National Academy of Sciences (NAS) guidelines barring compensation were set in part to protect poor people from being exploited by labs that might offer large sums of money — along the lines of rules barring compensation for organ donation. But Alta Charo, a lawyer and bioethicist at the University of Wisconsin Law School in Madison, who liaised with the NAS committee that set donor-compensation guidelines in 2005, says the move was as much political as ethical. In California, supporters of Proposition 71, which allows funding for stem-cell and cloning research in the absence of federal funding, adopted compensation prohibition in part, Charo claims, "to assuage a fringe group of the women's movement" that was aligned against the assisted-reproduction community.

The NAS guidelines followed the lead taken by California and some countries to ensure that stem cells could easily cross state and international borders. The United Kingdom, however, essentially changed the rules two years ago. In 2006, Alison Murdoch of Newcastle University received approval for a plan to allow couples to defray the costs of fertility treatments if they are willing to share some eggs for research purposes. This could create disparity in the quality of care available to people who don't have the money to undergo fertility treatments, says

How close, how comfortable?

She has chosen to remain anonymous, but the one egg donor acquired by Kevin Eggan's group at Harvard University has unspecified ties to Harvard.

It's fact that calls to mind the looming spectre of disgraced Korean cloner Woo Suk Hwang, whose fall was precipitated not just by his fraudulent research, but his ethically questionable use of eggs donated by members of his lab.

However, although Eggan's research protocol did not preclude such donors, his team agreed to avoid such conflicts. "The first questions that came out of my mouth," says Eggan, "were 'Does she work in my lab or my department?'" The answer was no.

Eggan has never met the donor, who could be any age between 25 and 35 according to the study's inclusion criteria, and he says he only found out about the Harvard connection when inquiring about her motivations.

Alta Charo, a lawyer and bioethicist at the University of Wisconsin Law School in Madison, says this shouldn't pose a problem: "If the donor was an employee or under direct supervision of the team doing the research, it would suggest that there is a risk that this was not a completely voluntary source.

"But if she was someone who teaches there, goes to school there, or works in a completely different setting, I don't see that there would be an issue there at all." **B. M.**

US researchers are not allowed to pay women for egg donation.

Charo, but it could also provide access to such treatments for more people. Murdoch has so far collected more than 100 eggs in this fashion.

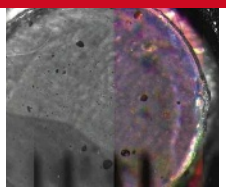
Wood says that this egg-sharing workaround presents a problem; the eggs being used for research are from older individuals, presumably with fertility problems. Using eggs donated by women aged 20–24, he has reportedly achieved cloning success rates near 25%.

And there are ethical concerns. Marcy Darnovsky of the Center for Genetics and Society in Oakland, California, says that such schemes split the doctor's duty of care three ways: between donor, fertility patient and researchers. And the focus on compensation, she says, distracts from concerns about risks from this one research avenue when other avenues seem open.

Even in light of rapid advancement of induced pluripotent stem cells — embryonic stem cell-like cells created by adding specific genes to adult cells — research using human eggs to reprogramme the nucleus continues, and with some success. Wood says Stemagen's work has been advancing rapidly. Eggan's group, while waiting for fresh eggs, showed that in principle unfertilized eggs could be obviated by introducing a nucleus into an already fertilized egg (D. Egli *et al. Nature* **447**, 679–685; 2007), which might be easier to obtain.

Still, Eggan says, his group's first modest attempts with unfertilized human eggs are more compelling. "Our results confirm, for us, that the best way to do the experiments is with unfertilized oocytes. That said, it doesn't seem like we're going to be getting many of those any time soon." ■

Brendan Maher



NASA/JPL-CALTECH/UNIV. ARIZONA

DASTARDLY DANDRUFF
Skin flakes could scupper Phoenix Mars mission.
www.nature.com/news

Fusion reactor faces cost hike

A massive international nuclear fusion experiment planned for Cadarache, France, is set to cost up to 30% more than anticipated and be delayed by as much as three years, governments will learn next week.

Construction has not even begun on the ITER fusion reactor, which has been beset by political wrangling since its inception. Now its seven international backers are to be told they will have to come up with an extra €1.2 billion–1.6 billion (US$1.9 billion–2.5 billion) on top of its current €5-billion construction budget if the project is to be realized.

A report from a group of scientific advisers says the additional money is needed for critical design changes and for coordinating between the participant nations. And the experiment, already delayed, will not be completed until anywhere from one to three years after its current 2016 due date.

Critics expect more cost hikes. "Personally, I think the price will double before it's done," says Stephen Dean, president of Fusion Power Associates, a research and educational foundation based in Gaithersburg, Maryland.

But ITER officials maintain that the new budget and schedule are achievable and that the review could not have come sooner because the ITER organization was only formed in 2006. "A good, rigorous review could only be done now," says Neil Calder, a spokesman for the project.

ITER is designed to test the feasibility of fusion power. The machine will trap hydrogen isotopes in magnetic fields and heat them to around 100 million Celsius. At that temperature, the hydrogen will fuse to form helium, releasing neutrons and energy in the process. If it works, it could lead to a second major source of atomic energy besides fission.

The project was originally proposed as a US–Soviet collaboration in the mid-1980s. It has since added partners including the European Union, South Korea, Japan, China and India, and was redesigned in the late 1990s to make it smaller and cheaper.

When ITER was finalized in mid-2005, its anticipated budget was €10 billion over 30 years, with half of that going to construction (see *Nature* 435, 1142; 2005). Dean says that estimate was based on an incomplete design and that "some things were not looked at in detail".

Over the past two years, scientists and engineers have been reworking the design. Researchers now want to add magnets to control a newly discovered instability (see *Nature* 452, 11; 2008). Buildings at Cadarache will have to be made earthquake-proof, something not included in the original blueprints because the site was undecided. And the addition of latecomers such as India and China will increase its overheads.

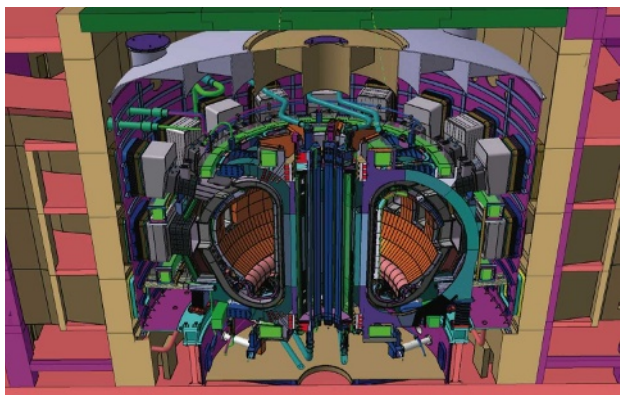
The new price tag must be approved by the ITER council, which oversees the project. "For all members, the cost issue is very sensitive," says one Japanese government official familiar with the project. The changes will be presented on 17–18 June at a meeting in Aomori Prefecture in Japan, and a new budget decided at a meeting in November.

Dean anticipates that the new budget will ultimately be approved. "This thing has gotten a life of its own — it's almost irrelevant how much it costs or what it's for." ■

Geoff Brumfiel

See Editorial, page 824

ITER



Money to burn: Extra magnets and earthquake protection are some of the factors making ITER more expensive.

SPECIAL REPORT

The winding road from ideas to income

Huge numbers of offices have been established over the past 30 years to help university researchers take their discoveries from the lab to the clinic. **Meredith Wadman** assesses their success.

It looks as though Phillip Robinson may achieve something most researchers don't dare to dream of — seeing his past decades' bench slog translate into a commercially available drug to treat conditions such as epilepsy. It's early days, but companies are already contacting the neurobiologist, declaring their interest in taking his products into the clinic.

Robinson, based at the Children's Medical Research Institute in Sydney, Australia, works with dynamin, a brain protein that is important for maintaining synaptic transmission. Over the past ten years, his group has developed dozens of dynamin inhibitors while looking for much-needed alternatives to treat epilepsy and perhaps cancer (V. Anggono *et al. Nature Neurosci.* **9,** 752–760; 2006).

Finding the drug candidates was only part of the battle, he says. The university technology transfer offices (TTOs) they approached were "hopeless", says Robinson. One of them "just showed no interest", he recalls. Another "helped us find patent attorneys and helped us write patents. But then they don't do anything. When you contact them, they say: 'have you commercialized it yet?' There was no understanding that we are scientists. This is not what we do."

But Robinson's complaints are not unique. A multi-continental chorus of academic researchers argues that the plethora of TTOs that have sprung up over recent decades are at best a mixed blessing to researchers in the life sciences.

Robinson got lucky with BioLink, a company established by the government of New South Wales in 2005 to commercialize preclinical discoveries in the life sciences. BioLink located an animal-model screening programme at the US National Institutes of Health (NIH) in which Robinson's group became the first Australian participants. That landed them, at no charge, with US$500,000-worth of screening services, allowing them to narrow down which of their many drug candidates were likely to be effective in epilepsy.

Pathology data from animal trials were needed next, Paul Field, BioLink's head, told Robinson. Field's team found five companies that could do the work, then the cheapest of those, which recently produced the animal data.

Today, says Robinson: "We feel we've got the upper hand. We're not going to take the first deal we find. Now we have the companies contacting us."

The difference between BioLink and the other TTOs he experienced is simple, he says: "BioLink works with its clients to help and advise how to take the idea forward." TTOs, he adds, "might help you get a patent. But that's the extent of what they're good at. That's not facilitating commercialization at all, or translation."

"We're definitely helping the academic researchers who want to be helped," counters Andy Sierakowski, director of the TTO at the University of Western Australia in Crowley and chair of Knowledge Commercialisation Australasia, an association of 28 TTOs from universities in Australia and New Zealand.

Sierakowski cautions that academics such as Robinson may be expecting too much of TTOs. "They can only do the [licensing] deal itself," he contends. "There's no way a TTO should think it's got all the knowledge. If you can get additional get help from BioLink, do it."

Chorus of complaints

Even scientists with an enviable track record are critical of TTOs. Mike Clark, an immunologist at the University of Cambridge , UK, says that his several successes in translating medications to the marketplace "have been despite TTOs, and mainly because of the dogged determination of a large group of scientists who kept on plugging forwards when things looked difficult both scientifically and business-wise". Clark is one of the scientists behind the discovery of Campath (alemtuzumab), a humanized monoclonal antibody made by Genzyme in Cambridge, UK, for patients with a hard-to-treat type of chronic leukaemia (L. Riechmann, M. Clark, H. Waldmann, G. Winter *Nature* **332,** 323–327; 1998).

"There was no understanding that we are scientists."

The list of complaints about TTOs is long and diverse. Critics say that they tend to overvalue their intellectual property, leading companies to walk away from potential deals. Others complain that they hoard inventions — even those of dubious commercial merit — with the result that many gather dust in TTO files rather than being built on by companies that might transform them into useful products. This is a particular problem in the United States, where a 'first to invent' rule exists, rather than the 'first to file' rule elsewhere. If a US researcher can document that he or she was the first to invent something, even if no patent application was filed at the time, then his or her application will trump that of a competitor who develops it later on.

There are also problems with small staffs — the median size of a US TTO is six employees, and many others around the world are smaller. The TTOs are being stretched too thinly, and lack expertise in the huge range of fields from which inventions may issue. Attrition of highly qualified staff is also a problem for some. "Most people who are any good are going to get better offers from industry, or end up working for a top venture capital firm," says Clark.

Others charge that universities are showing an unseemly keenness to capitalize on inventions through their TTOs, blurring what should be a clear line between public institutions and commercial enterprises. Occasionally, this has caused students and postdocs to be pushed in commercial directions, against their better interests. "Phenomenal amounts of money have been made by a very small number of institutions. That's really what these offices are set up for. Every tech-transfer person in the country wants to land the next Gatorade, the next Taxol, the next Cisplatin," says John Frangioni, who works on *in vivo* cancer detection at Beth Israel Deaconness Medical Center in Boston, Massachusetts. Frangioni has filed some 50 invention disclosures with his TTO; of these, five have been patented, three licensed and 12 patents are pending.

TTO managers dispute this charge of money-grabbing, pointing out that some TTOs lose more money than they make for their institutions, and that windfall deals are rarities. They say that their work is about translating the science into new therapies, not about making money. Sierakowski points out that globally, almost all major universities have between 1% and 5% of their research budgets returned via TTOs in royalties and cash from selling their original interests. "This is not big money."

Jon Soderstrom, managing director of the TTO at Yale University in New Haven, Connecticut, and president of the Association of University Technology Managers (AUTM), the trade group for TTOs in the United States,

IN THE FIELD
Nature blogs from the Arctic ice.
http://blogs.nature.com/news

SWAPPING FUTURE ROYALTIES FOR CASH: RECENT US MEGA-DEALS

Inventing institution	Invention	What it does	Company licensed to develop invention	Amount (US$) royalty rights were sold for
Northwestern University	Lyrica (pregabalin)	Anti-convulsant and pain reliever	Pfizer	$700 million (to Royalty Pharma)
New York University	Remicade (infliximab)	Anti-inflammatory to treat autoimmune diseases	Centocor (later acquired by Johnson & Johnson)	$650 million (to Royalty Pharma)
Emory University	Emtriva (emtricitabine)	Reverse transcriptase inhibitor to treat HIV	Triangle Pharmaceuticals (later acquired by Gilead Sciences)	$525 million (to Gilead Sciences & Royalty Pharma)
Memorial Sloan-Kettering Cancer Center	Neupogen (filgrastim) and Neulasta (PEG-filgrastim)	Stimulates production of white blood cells	Amgen	$263 million (to Royalty Pharma)
Yale University	Zerit (stavudine)	Reverse transcriptase inhibitor to treat HIV	Bristol-Myers Squibb	$115 million (to Bristol-Myers Squibb)
The Wistar Institute	Rotateq	Vaccine to prevent rotavirus infection	Merck	$45 million (to Paul Royalty Fund)

thinks that the offices are unfairly maligned. "I hear the stories. But they don't hold up under scrutiny," he says.

Certainly TTOs, which are charged with identifying and then exploiting opportunities to commercialize academic inventions, have delivered rich rewards in hundreds of instances. These include bringing several significant antiretrovirals to the clinic to fight HIV/AIDS, and the invention of a way to make the anticancer drug Taxol (paclitaxel) without destroying endangered Pacific yew trees. The particulars of individual royalty agreements between TTOs and companies are confidential, but when TTOs sell off these royalty rights — as they are doing more often with big winners — some mind-boggling numbers are involved (see Table).

In the United States alone, TTOs filed nearly 16,000 new patent applications in 2006, saw their licensees launch 697 products and spun out 553 start-up companies, according to the AUTM. Overall, US technology transfer has been "a big success story", says Mark Rohrbaugh, the director of the NIH Office of Technology Transfer.

Not all academics are dissatisfied. The performance of the TTO at Wake Forest University in Winston-Salem, North Carolina, "has been excellent. I don't think I could have done it without them," says Keith Walter, an ophthalmic surgeon there. Walter's device, Endosaver, which spares fragile tissue during minimally invasive corneal transplantations, is expected to enter clinical trials this summer. It is licensed to Ocular Systems Incorporated, a local company that worked with Walter through numerous iterations of a prototype. Walter notes that the TTO's associate director, Dean Stell, gave him tips he would never have thought of, such as advising him to enter even back-of-the-napkin sketches made during coffee breaks in a log book to protect his intellectual property.

Change of law

The US wave in technology transfer began when the Bayh-Dole Act of 1980 gave universities title to ownership of inventions resulting from research funded by the federal government — title that had formerly belonged to the government. At the time, TTOs in the United States numbered in the single digits. Today, more than 230 universities have them.

The trend was slow to reach Europe. More than half of Switzerland's TTOs were established between 1990 and 2000, for instance. In Italy, more than 40% were established in 2000 and 2001 — just before a change in the law granted title to inventors rather than institutions.

It hasn't always worked. For instance, in Germany, when a law change in 2002 gave universities title to their researchers' inventions, the ministry for education and research funded a TTO for every federal state. But because the initiative was underfunded and regional TTOs are spread too thinly, "I don't know of a single TTO set up from that programme that makes a profit or that even has a therapeutic product visible in the coming three years," says Christian Stein, head of Ascenion, the TTO for most of the life sciences at the Helmholtz and Leibniz Associations, which cover Germany's top bioresearch institutes. By working with the best institutions, and only in the life sciences, Stein's seven-year-old TTO has been able to generate more than €2 million (US$3 million) a year for the last several years in licensing fees and equity sales from exiting spin-offs.

Still, he says: "In terms of professional tech transfer, we are probably at least ten years behind what I see in the United States and quite a few years still behind the United Kingdom."

Some say that things have improved in the United Kingdom, where in the past TTOs have been accused of being inexpert and passive. Nancy Rothwell, vice president of research at the University of Manchester, UK, says that her university's TTO has been helped by a cadre of proactive business-development managers who seek out inventions rather than waiting for inventors to come to them. It has also instituted a policy pledging researchers an unusually high 85% of the income coming back to the university from any discovery that makes money.

Others are lukewarm on the nation's tech-transfer enterprise, which varies hugely between institutions. "Frankly, if there wasn't a need to fill this gap between basic science and product development, we wouldn't need to operate a tech-transfer enterprise," says Ted Bianco, who runs the Wellcome Trust's technology-transfer effort, costing £120 million (US$240 million) in its first year. In groundbreaking efforts such as medical engineering, "we get our inspiration from the United States, seeing how institutions such as Massachusetts General Hospital and Massachusetts Institute of Technology work", he says, citing the US TTOs' more proactive agenda, in which firms are courted for carefully selected inventions and staff are highly qualified with industry experience.

"Everyone in the field knows that TTOs can help or hurt researchers."

Even the likes of those institutions, however, fail to impress some commercially-headed entrepreneurs. Chemist Namyong Kim, for example, worked at two Boston-based start-ups before deciding to move to Singapore. He was drawn by the tech-transfer focus of Singapore's A*STAR institutes, which have great connections to companies and investors, and a good supply of grants to assist development. The TTO at A*STAR — called Exploit Technologies — quickly applied for the necessary patents for Kim's miniaturized aqueous bioassay and awarded him US$236,000 to develop a prototype. Kim's two-month-old spin-off, Curiox Biosystems, has already made its first sale of the $13,500 DropArray.

"Everyone in the field knows that these TTOs can be helping or hurting academic researchers," says Kim. "The support here has been quite impressive compared to my experiences in the United States." ■

See Editorial, page 823, and online at http://tinyurl.com/3tt3y3.

S. OGDEN / SPL

A court ruling could affect more than just macaque research.

Swiss court bans work on macaque brains

Zurich's two largest institutes are appealing to the country's supreme court after a lower court decided to ban two primate experiments studying how the brain adapts to change. They say that the ban is a serious threat to all basic research that uses animals in Switzerland.

The University of Zurich and the Federal Institute of Technology Zurich (ETHZ) announced on 4 June that their local administrative court had ruled against the experiments on rhesus monkeys that had been approved in 2006 by the Swiss National Science Foundation, a funding agency and the Zurich canton's veterinary office, which is responsible for controlling animal welfare.

The veterinary office decision was challenged by an external advisory committee on animal experimentation, which argued that the proposed experiments would offend the dignity of the animals. The requirement to consider the 'dignity of creatures' was introduced into the Swiss constitution in 2004.

The court did not refer to dignity, but agreed that society was unlikely to see the benefits of the research during the three-year funding period approved, and thus the burden on the animals was not justified. Swiss law requires that the benefit to society must be weighed against the burden to animals before any animal experiment can take place.

"But the court has made a new interpretation of the law which demands immediate benefits — and that's not compatible with any form of basic research," says Peter Chen, vice-president for research at ETHZ. "It goes beyond the requirement of the law and we have no choice but to contest the decision in the highest court."

The experiments were to be done at the Institute of Neuroinformatics, jointly owned by the two institutes. One experiment, to monitor changes in the cortex during perceptual learning, involved denying test monkeys a drink for up to twelve hours to increase the value of the apple juice reward they would be offered if they learnt a new task correctly. The other experiment, designed to understand the microcircuitry in the cortex, involved sacrificing the animals to follow the microcircuitry microscopically.

The scientists use primates because their brains are closer to the human brain than any other species. "We need to understand the basic biology of our brains in order to be able to successfully treat brain diseases such as Parkinson's," says Kevan Martin, one of the researchers. He says he intends to re-apply for ethics approval, making the value of the research more explicit.

Roger Lemon, a neuroscientist at University College London, UK, says that finding out how brain circuitry works in normal as well as disease conditions is fundamental if new therapies are to be developed for neurological disorders. "A huge amount of the sort of work that Martin and his colleagues do lies behind the breakthroughs in applications we are now starting to see." ■

Alison Abbott

See Editorial, page 824.

3 GOOD REASONS

To buy an HIV ring

1 It will raise awareness and empathy for the global AIDS problem, hopes its designer, Andriy Kurovets.

2 It can be worn as a "symbol of the possibility of fighting the disease".

3 It's strangely beautiful.

ROBOT NEWS

High ambitions

The US Defense Advanced Research Projects Agency (DARPA) has handed out millions of dollars for companies to design a plane that can stay aloft for five years unattended.

Low expectations

The Bristol Robotics Laboratory, UK, has won a £1-million (US$2 million) grant to see whether a robot can stir soup safely. The lab will test a robot's abilities to work alongside a person in an enclosed space, such as a kitchen.

NUMBER CRUNCH

£1.9 billion ($3.8 billion) is the estimated worth of unpaid peer review done worldwide, according to a UK report.

8.7% of that is donated by UK researchers, despite the fact that only ...

6.6% of journal articles are written in Britain.

ON THE RECORD

"We're planning on redesigning the bus stop in a 1960s look."

Sabine Grünwald of the Mühlenau retirement home in Hamburg, Germany, ponders how to improve the fake bus stop they installed outside the facility last year. Patients with dementia who wander off tend to collect at the stop, where they can be collected by staff. A bus never actually comes.

Sources: A. Kurovets, DARPA, Ananova, Times Higher Education, Deutsche Welle

SIDELINES

Near-perfect 'black'

Researchers working with metamaterials say that they are near to achieving a surface that can absorb every photon that hits it.

Physicist Willie Padilla's group at Boston College in Massachusetts and the team's colleagues at Duke University in Durham, North Carolina, have created structures that provide a proof of principle of these perfectly absorbing surfaces, although they have yet to stop all photons from escaping[1]. The researchers predict that the material will have initial applications in ultra-sensitive imaging technologies that use terahertz radiation, with wavelengths intermediate between infrared and microwaves. Terahertz imaging is a hotly tipped technology for use in medicine — fully absorbing materials at these wavelengths could help to reveal even the dimmest of features.

"Terahertz frequencies are notoriously difficult to work with — difficult to generate, and difficult to detect," says Marin Soljačić of the Massachusetts Institute of Technology in Cambridge, who is also working on controlling absorption in tailor-made metamaterials[2]. "But there are many exciting applications that could be enabled if we mastered this technology. This work is a nice step in that direction."

Metamaterials are structures made from arrays of 'designer atoms', each of which is itself a device or structure with specially tailored properties. So far, the building blocks of most metamaterials have been designed to interact with electromagnetic radiation in unusual ways, giving the assembled metamaterial an optical behaviour that ordinary materials could never display. For example, it might have a negative refractive index, bending refracted light the 'wrong way'. Metamaterials have already provided prototypes of 'superlenses' that are more powerful than conventional ones, and 'invisibility shields' that bend electromagnetic waves smoothly around objects hidden inside.

The basic 'atoms' of these structures are C-shaped split rings made from patterned metal films on circuit boards. Radiation with a wavelength comparable to the size of the rings will stir up resonant electric currents in the metal, causing the radiation to be absorbed. For metamaterials operating at microwave frequencies, the components are relatively easy to make and are large enough to be visible.

Padilla's team has developed a new design for their perfect absorber, consisting of two back-to-back split rings. They calculated that an array of these elements should absorb about 99% of incident radiation at the resonant microwave frequency of about 11.5 gigahertz. Their experimental version managed to soak up only about 96%.

But there are plenty of ways to improve this, Padilla says. One of the likely reasons for the imperfect absorbance is small errors in the shapes of the split-rings or in the accuracy of putting them together — so, more careful fabrication should produce better performance. And the researchers have now come up with an improved design that, in theory, offers 99.99% absorbance. If more than one layer of the metamaterial is used, the absorbance should be more or less perfect, they say.

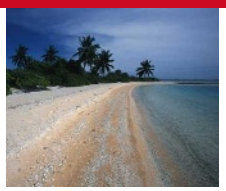
Near-perfect light absorbers are needed for detectors called bolometers, which sense photons by converting their energy to heat. Bolometers made from silicon are standard items in infrared sensing, but they need to be cryogenically cooled and only work in a limited range of wavelengths. The new metamaterial absorbers work at room temperature, and the wavelength response can be tuned simply by varying the size and shape of the components.

Such an absorber would still be perfect only within a certain range of wavelengths. But this range could be broadened, to make the equivalent of a 'perfectly black' surface that absorbs fully over the entire visible spectrum by mixing different-sized split-ring components in the array. Alternatively, a perfect absorber could pick out specific 'colours', which would be useful for combining imaging and spectroscopy. "At terahertz frequencies, you could image airport luggage directly through the bag, just as with X-rays," says Padilla. "But being able to image at different wavelengths means you might obtain spectroscopic information about what you are imaging, so that you can determine if it is harmful."

Padilla and his colleagues have already managed to scale down their design to work at terahertz and far-infrared frequencies[3]. But he says that making a perfect black in the visible range will be hard, because the components would have to be so small. A super-black material made from carbon nanotubes was reported recently[4], in which the high absorbance comes from the roughness of the surface rather than the antenna action of metamaterials. ■

Philip Ball

1. Landy, N. I., Sajuyigbe, S., Mock, J. J., Smith, D. R. & Padilla, W. J. *Phys. Rev. Lett.* **100,** 207402 (2008).
2. Chan, D. L. C., Celanovic, I., Joannopoulos, J. D. & Soljačić, M. *Phys. Rev. A* **74,** 064901 (2006).
3. Tao, H. *et al. Opt. Lett.* **16,** 7181–7188 (2008).
4. Yang, Z.-P., Ci, L., Bur, J. A., Lin, S.-Y. & Ajayan, P. M. *Nano Lett.* **8,** 446–451 (2008).



PUNCHSTOCK

SINKING FAST
Kiribati admits defeat against rising tides.
www.nature.com/news

NIH responds to critics on peer review

BETHESDA, MARYLAND

Responding to hundreds of critical comments, the US National Institutes of Health (NIH) has reversed several controversial proposals made in February as part of a year-long effort to overhaul the agency's peer-review system (see *Nature* **451,** 1035; 2008).

As part of an initiative called Enhancing Peer Review, announced in a finalized form on 6 June, the agency will spend at least $200 million annually over the next five years to foster groundbreaking, investigator-initiated research. Of that, at least $250 million will go to a new beast: a Transformative R01 Award, a reach-for-the-skies version of the NIH's basic grant. The remaining $750 million will go to existing awards that reward risk and innovation: the Eureka, New Innovator and Pioneer awards.

"Gone is the category 'not recommended for resubmission', for dismal applications."

The changes "are concrete solutions that will maximize flexibility, remove any unnecessary burden, stimulate new innovation and promote transformative research", says NIH director Elias Zerhouni.

They include rewards for long-serving reviewers; a streamlined, 12-page R01 grant application, down from 25; and a seven-point, integer scoring scale for grant applications, which will be assessed across five criteria: impact, investigators, innovation, feasibility and environment. Current applications are graded on a 41 point scale, from 1.0 to 5.0, raising complaints that they claim a degree of accuracy that can't be scientifically defended.

Among the controversial proposals shelved by the agency was a recommendation that all applications, even those on a second or third submission, would be treated as new, without reviewer access to prior reviews.

Gone, too, is the category "not recommended for resubmission", which had been suggested for dismal applications. Scientists felt that branding projects with "a clear, checkbox-driven stigma is bad, that it could have unintended consequences", Jeremy Berg, director of the National Institute of General Medical Sciences, told the advisory committee.

Berg and Lawrence Tabak, director of the NIH's dental institute, head the group that developed the recommendations and are charged with implementing them over the next 18 months.

The agency also jettisoned a "minimum effort requirement" that would have required principal investigators to commit at least 20% of their time to any single NIH grant — an item of particular concern for 'grandee grantees' (see *Nature* **452,** 258–259; 2008). Instead, grantees will need to indicate if they will have more than $1 million in cumulative NIH funding. ■

Meredith Wadman

See also page 838.

Beep-beep! Roadrunner breaks petaflop barrier

The US military's Roadrunner supercomputer has become the first machine to calculate at more than a quadrillion (10^{15}) operations per second, thereby crossing the petaflop barrier.

The record was technically set on 25 May, in IBM's laboratories in Poughkeepsie, New York, using the LINPACK standard benchmark for comparing supercomputer performance. In early June, however, it achieved petaflop performance on a real-life application of modelling the human cortex, says David Turek, vice-president for 'deep computing' at IBM.

It thus beats the record of IBM's Blue Gene/L machine, based at the Lawrence Livermore National Laboratory in California, which managed about 500 teraflops (10^{12}).

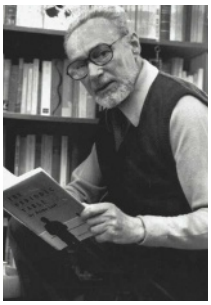
Roadrunner was built to run calculations to ensure the safety of the US nuclear stockpile, and is due to be delivered next month to the Los Alamos National Laboratory in New Mexico for that purpose. In the meantime, it will also run other calculations, including plasma physics, molecular dynamics and climate change.

Physicist's involvement in historical race row disputed

A historical investigation has posthumously cleared Eligio Perucca, a physics professor at the Polytechnic University of Turin, of alleged anti-semitism towards Primo Levi.

Levi, an Italian chemist, writer and Auschwitz survivor, described in his book *The Periodic Table* the difficulties he had faced in persuading several professors in Turin to take him on as a doctoral candidate in 1940, because of the race laws that the Fascist regime imposed in Italy at that time. Perucca was later fingered as one of these professors in a Levi biography.

AP



Primo Levi.

But Bart Kahr, a chemist at the University of Washington in Seattle, has looked into the claim and deduced that Perucca was not a fascist nor an anti-semite.

What's more, the investigation credits Perucca with a chemical discovery — of a certain type of optical activity, called optical rotary dispersion, of dyed chiral sodium chlorate crystals — in 1919 that was thought to have been discovered only in 1931.

Japanese lab installed on space station

NASA/JSC

Astronauts installed the main piece of Japan's laboratory module, Kibo, aboard the International Space Station last week.

At 11 metres long, Kibo, delivered by the space shuttle *Discovery* and opened by Japanese astronaut Akihiko Hoshide (pictured), is the largest pressurized research facility in space. Its final section, for work in unpressurized environments, is slated to be flown up next year. Astronauts also tested Kibo's robotic arm, which will be used to manipulate experiments on that final section when installed.

The station's malfunctioning toilet was also fixed during the trip.

Panel urges further review of controversial NIH lab

A blue-ribbon panel last week told the National Institutes of Health (NIH) that it needs to conduct more studies if it hopes to see high-risk biosafety research conducted at a lab under construction at Boston University Medical Center in Massachusetts.

The panel, chaired by molecular biologist Adel Mahmoud of Princeton University in New Jersey, said that the agency should conduct risk assessments for 12 dangerous pathogens, including comparing the risks of each being released at urban, suburban and rural sites.

Advisers to NIH director Elias Zerhouni, who received the report on 6 June, offered some sharp criticism of the $198-million project, most of which is being paid for by the NIH. Community concerns about the risks of the facility have led to it being challenged in both the state and federal courts.

Publications follow policy on stem-cell research

Countries with less restrictive policies for deriving human embryonic stem cells produce a disproportionate share of the field's publications, according to a new study (A. Levine *Cell Stem Cell* 2, 521–524; 2008).

By this measure, the United States was the worst performer: although 36% of scientific publications on human embryonic stem cells in 2006 had a US-based corresponding author, that compared with 46% of a control set of biomedical publications and 47% of publications on RNA interference, a less controversial 'hot' technology.

The study's author, bioethicist Aaron Levine of the Georgia Institute of Technology in Atlanta, notes that governments in the top five overperforming countries (Singapore, United Kingdom, Israel, China and Australia) support research and permit derivation of new cell lines. Besides the United States, other underperformers included Japan, France and Switzerland. Both France and Japan also have less than the expected share of publications on RNA interference, and Levine suggests that scientists in those countries are less inclined to pursue emerging technologies.

Senate climate debate suffers death by hot air

Democratic leaders in the US Senate walked away empty-handed last week after a highly anticipated debate over climate-change legislation disintegrated into partisan bickering and delaying tactics.

The bill, which would have required a two-thirds cut in greenhouse-gas emissions by 2050, died on 6 June on a procedural vote. Not a single amendment was considered, and no substantive votes were cast.

Opponents said the legislation would further increase energy prices and push US jobs overseas. The lowlight came as Republicans invoked an obscure Senate rule to force staff to read the entire 492-page bill aloud over the course of some nine hours.

Few had expected the legislation to pass, but Democrats and environmentalists had hoped for a substantive debate that would lay the groundwork for next year.

Corrections

The News story 'No star left behind' (*Nature* **453,** 437; 2008) misstated the number of constellations constructed for an index in star-matching software. The number of constellations is 800 million not 800,000.

In the News story 'Climate anomaly is an artefact' (*Nature* **453,** 569; 2008) we incorrectly stated that David Thompson was at the State University of Colorado in Boulder. He is in fact at the Colorado State University in Fort Collins.

A delicate balance

Near-term and long-term research are vying for attention. **David Goldston** says that a fuller congressional debate is needed.

PARTY OF ONE

Most of the debate on research policy in Washington DC is over how much to spend, but there is a growing, if inchoate, discussion about what type of work to support. As is often the case, the trend can be spotted through the increasing appearance of buzzwords; two current examples are the somewhat opposing notions of 'translational research' (shorter-term work focused on coming up with new products) and 'transformative research' (longer-term work aimed at coming up with new ideas).

Discussions about whether the federal government is supporting the right types of research crop up periodically, usually at times of economic distress. The underlying question is always: 'is the country gaining the greatest possible practical benefit from its research investment?'

Academics can resent this question, longing for an era when research was supported for knowledge's sake. But such a time never existed. The foundational document for US science policy after the Second World War, Vannevar Bush's 1945 report *Science: The Endless Frontier*, begins with these words: "We all know how much the new drug, penicillin, has meant to our grievously wounded men." Bush, the overseer of US wartime research, promised that federal funding will "bring higher standards of living, will lead to the prevention or cure of diseases, will promote conservation of our limited national resources and will assure means of defense against aggression".

But it's not obvious how to run a research enterprise to maximize such benefits, and policy-makers continue to fiddle with the system. In the 1980s, for example, spurred by the fear that US competitiveness was being undermined by Japan, Congress passed bills designed to help translate research prowess into commercial success. The 1980 Bayh–Dole Act, for instance, made it easier for universities to patent the results of federally funded research, encouraging the growth of spin-off companies, especially in biotechnology. No one called that translational research, but the goal was the same.

The current interest in biomedical translational research reflects not only the resurgence of concerns about US competitiveness, but also worries about whether the drug-discovery pipeline is drying up and a desire to show that the doubling of the budget of the National Institutes of Health (NIH) between 1998 and 2003 can produce measurable results.

That those issues are specific to biomedicine is one reason there has been much less talk in Congress about translational work in the physical sciences, even though the gap between basic research and commercial advance is generally considered greater in the physical sciences. A nanotechnology bill passed last week by the House of Representatives does authorize funding for partnerships between academia and industry to conduct research in 'areas of national importance' (including health care), but that would still be early-stage work. And in general, the prescription for the physical sciences has just been for increased spending on conventional basic research.

Discussions about taking further steps in the physical sciences inevitably lead to ideological debates about the roles of government and industry, with conservatives branding efforts to use taxpayers' money to translate research into products as industrial policy — a derisive term that began life as a positive buzzword in the 1980s. Somehow in biomedicine — perhaps because government, academia and industry are so obviously entwined, perhaps because health is seen as a legitimate government concern — such ideological debates are absent or muted.

Translational programmes in biomedicine are also being exempted from another standard line of attack. They can be seen, often by researchers themselves, as being too focused on near-term, incremental work. It is curious, then, that translational research in biomedicine is becoming all the rage just as reports are calling for an emphasis on transformative or high-risk, high-return research. For example, a report released last week by the American Academy of Arts and Sciences (for which I sat on the panel) argued that the priorities in biomedical research should be more funding for transformative research and for early-career researchers (see page 953).

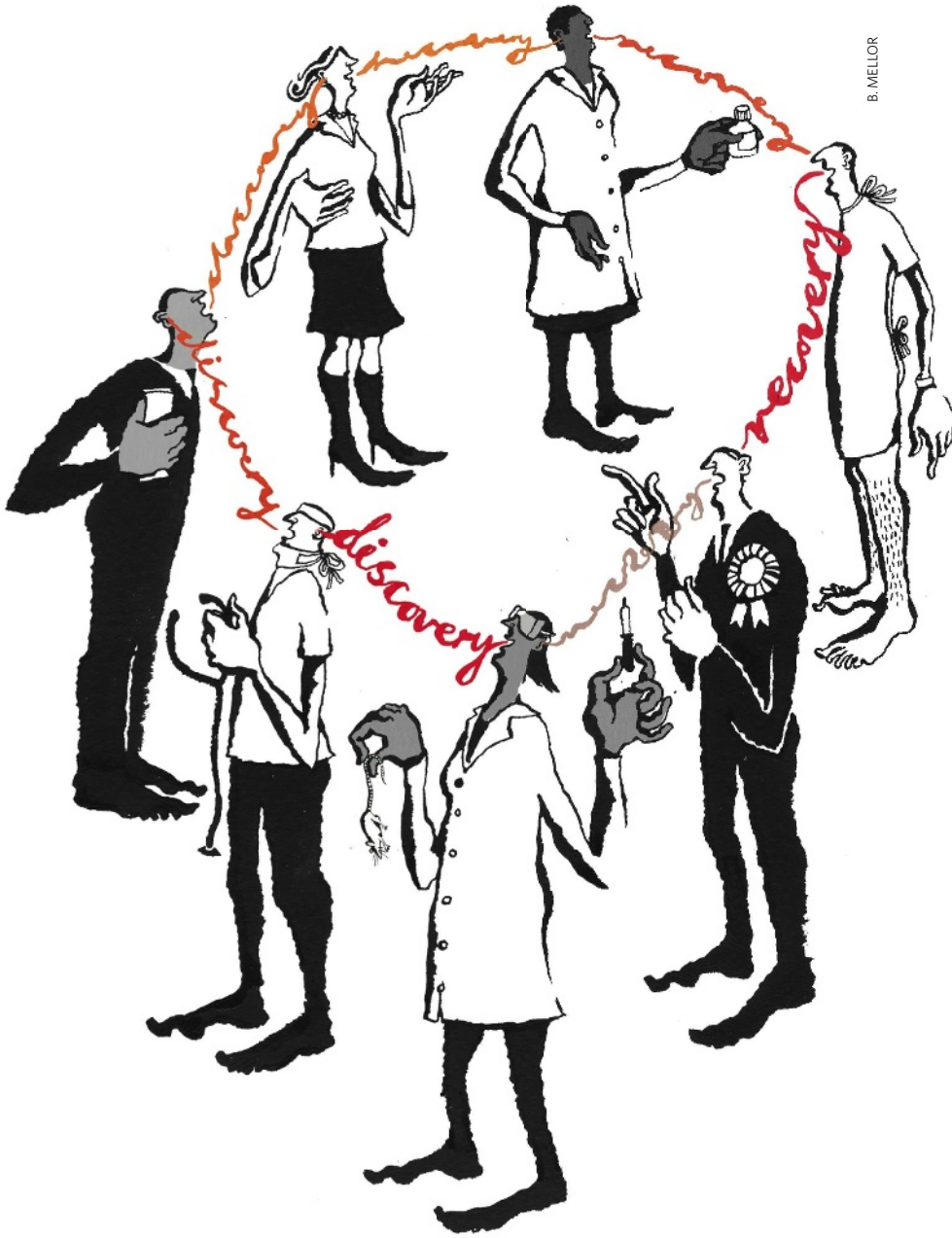
Theoretically, the research system could simultaneously give special attention to shorter-term research to convert ideas into products, and to longer-term research to generate novel ideas, but that's a hard balance to pull off, and the political debate has tended to stress one or the other. The NIH's own efforts highlight what happens with insufficient focus. Its Pioneer awards are designed to fund transformative research, but the programme is so small that fewer than 4% of applications win funding even though the number of applications plummeted as word spread about the poor odds. The NIH last week announced that it would increase funding for transformative programmes (see page 835), but the money still may be spread too thinly.

A new bill by Senator Joe Lieberman (Independent, Connecticut) also shows how policy tends to focus on just one end of the spectrum. The legislation is designed to further translational research, although it also would create a new programme for longer-range research targeted at specific health problems. With its focus on results, the bill would require every grant application to the NIH to state how the research could be used "for detecting, treating or curing" a medical condition. Such a mandate might help to ensure that basic research translates into advances in public health. Long-term and even transformative research can be targeted at solving specific problems, as Bell Laboratories proved, gaining Nobel prizes while inventing the transistor. But if viewed narrowly, the provision could easily stymie inventive work that is not far enough along to be associated with a specific medical condition. It hardly seems like the cure for an agency that is often criticized as being too conservative.

But the Lieberman bill could open up a useful debate on how to develop a better-balanced research enterprise — one that keeps its eye on real problems without becoming risk averse. That discussion needs to take place irrespective of what happens to overall spending levels. And it has to consist of more than throwing around contradictory buzzwords. ■

David Goldston is a visiting lecturer at Harvard University's Center for the Environment. Reach him at partyofonecolumn@gmail.com.

See also Editorial, page 823 and online at http://tinyurl.com/3tt3y3



GETTING THE MESSAGE ACROSS

There is a growing disparity at the heart of biomedicine. In some ways, the field is experiencing a golden age: the quantity of basic research is shooting off the charts and budgets are far higher than they were two decades ago. Yet the impact of this research is growing at a much more modest rate: new cures and therapies are ever more expensive to develop and worryingly thin on the ground.

The term 'translational research', hardly heard ten years ago, is now on everyone's lips because it is seen as the solution to this disparity. It is expected to ensure that the bounty of the golden age is 'translated' into benefits in the everyday world of cancer, dementia or heart disease, for example. It thus has a key part to play in improving our lives and also in rationalizing the social contract between researchers and the taxpayers who help fund them in expectation of future cures (see page 823).

In this issue we examine a range of topics in translational research. On page 840, Declan Butler finds out what proponents of the idea at the US National Institutes of Health expect to achieve, and how they plan to deliver on their promises. And on page 846, Helen Pearson visits the Ludwig Institute for Cancer Research, which claims to have been attempting translation for 20 years and so has a trick or two to teach newcomers.

The relationship between academia and industry is key to successful translation. University technology-transfer offices can loom large in this relationship. On page 830, Meredith Wadman discusses concerns that some of these offices are more of a hindrance than a help. Closer collaboration, in which industry pays universities to solve problems that it defines, might be one solution (see page 853). More radically, a book reviewed on page 855 argues that the entire drug industry should be restructured, with research and development pried free from marketing.

The concept that translational research is a one-way flow from bench to bedside is seen by many as outdated. Clinical data and human trials can inspire insights that flow from bedside to bench, too, as Heidi Ledford reports on page 843. The observation that some anti-inflammatory drugs may delay the onset of Alzheimer's disease, for example, has led one group to find out how they block formation of dangerous protein products (see pages 861 and 925), which may help to refine the drugs. And pages 863 and 921 detail progress towards a transgenic monkey model for Huntington's disease.

Riches abound. If these discoveries can be carried to the clinic, biomedicine will be all the richer. ■

CROSSING THE VALLEY OF DEATH

A chasm has opened up between biomedical researchers and the patients who need their discoveries. **Declan Butler** asks how the ground shifted and whether the US National Institutes of Health can bridge the gap.

B. MELLOR

"NIH stands for the National Institutes of Health, not the National Institutes of Biomedical Research, or the National Institutes of Basic Biomedical Research." This jab, by molecular biologist Alan Schechter at the NIH, is a pointed one. The organization was formally established in the United States more than half a century ago to serve the nation's public health, and its mission now is to pursue fundamental knowledge and apply it "to reduce the burdens of illness and disability". So when employees at the agency have to check their name tag, some soul searching must be taking place.

There is no question that the NIH excels in basic research. What researchers such as Schechter are asking is whether it has neglected the mandate to apply that knowledge. Outside the agency too there is a growing perception that the enormous resources being put into biomedical research, and the huge strides made in understanding disease mechanisms, are not resulting in commensurate gains in new treatments, diagnostics and prevention. "We are not seeing the breakthrough therapies that people can rightly expect," says Schechter, head of molecular biology and genetics at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland.

Medical-research agencies worldwide are experiencing a similar awakening. Over the past 30 or so years, the ecosystems of basic and clinical research have diverged. The pharmaceutical industry, which for many years was expected to carry discoveries across the divide, is now hard pushed to do so. The abyss left behind is sometimes labelled the 'valley of death' — and neither basic researchers, busy with discoveries, nor physicians, busy with patients, are keen to venture there. "The clinical and basic scientists don't really communicate," says Barbara Alving, director of the NIH's National Center for Research Resources in Bethesda.

Alving is a key part in the NIH's attempt to bridge the gap with 'translational research'. Director Elias Zerhouni made this bridge-building a focus in his signature 'roadmap' for the agency, announced in 2003 (see *Nature* **425,** 438; 2003). Spearheading the NIH effort will be a consortium of 60 Clinical and Translational Science Centers (CTSCs) at universities and medical centres across the country, which will share some US$500 million annually when they are all in operation by 2012. Late last month, the NIH doled out the most recent grants in

this programme to 14 institutions, including Indiana University School of Medicine in Indianapolis and Harvard University, bringing the consortium up to 38 member centres since its launch in 2006.

Yet the money for the CTSCs will total only 1–2% of the NIH's annual budget of $29.5 billion, and at this early stage it is not clear how much these catalysts will be able to change the terrain. Even so, some people credit the organization and its leader for trying. "Lots of people say they hate Zerhouni. I love him. He had the courage to come forward and say that the NIH was not delivering on its promise," says Lee Nadler, head of the new CTSC at Harvard.

Ask ten people what translational research means and you're likely to get ten different answers. For basic researchers clutching a new prospective drug, it might involve medicinal chemistry along with the animal tests and reams of paperwork required to enter a first clinical trial. For groups wanting to developing diagnostics, imaging tools, or screening and prevention methods the route would be different.

New image

In some sense much translational research is just rebranding — clinical R&D by a different name. But it also involves investing in training, research and infrastructure to help researchers engage in clinical research — and cross the valley of death. Funding agencies hope that this will break down barriers in the transformation of basic-science breakthroughs into clinical applications ('bench to bedside') and enable more research on human subjects and samples to generate hypotheses that are more relevant to people than to animal models (see page 843).

The barriers to translational research are relatively recent. Back in the 1950s and 60s, basic and clinical research were fairly tightly linked in agencies such as the NIH. Medical research was largely done by physician–scientists who also treated patients. That changed with the explosion of molecular biology in the 1970s. Clinical and basic research started to separate, and biomedical research emerged as a discipline in its own right, with its own training. The bulk of biomedical research is now done by highly specialized PhD scientists (see graph), and physician–scientists are a minority.

The basic biomedical research enterprise has now evolved its own dynamic, with promotions and grants based largely on the papers scientists have published in top journals, not on how much they have advanced medicine. And many clinicians who treat patients — and earn fees for doing so — have little time or inclination to keep up with an increasingly complex basic literature, let alone do research. This has diminished the movement of knowledge and hypotheses back and forth between bedside and bench. At the same time, genomics, proteomics and all its cousins are generating such a volume of potential drug targets and other discoveries that the pharmaceutical industry is having trouble digesting them. With pharma spending more on research but delivering fewer products (see graph), it is no longer in a position to take forward most academic discoveries. "There is a real crisis in the industry," says Garrett Fitzgerald, head of the CTSC based at the University of Pennsylvania in Philadelphia.

One crude way of tracking the rupture is to see when people developed a new rhetoric to deal with it. The term 'translational research' first appeared in PubMed in 1993, sparked by the characterization of *BRCA1* and other cancer genes, which suggested immediate applications in early detection and treatment of cancers. Use of the term remained low throughout the 1990s, in just a handful of papers annually, until around 2000, after which it has cropped up in several hundred articles each year.

NIH

"There was a widening gap between basic and clinical research."
— Elias Zerhouni

In 2000, the US Institute of Medicine convened the Clinical Research Roundtable, which held a series of meetings that are credited with putting translational research high on the agenda. The process pinpointed two blockages in the transfer of research knowledge into practice (S. H. Woolf *J. Am. Med. Assoc.* **299,** 211–213; 2008). The first was preventing laboratory advances being converted into new medical products and tests in humans; the second was stopping proven improvements in treatment — a new drug combination, for instance — becoming adopted in medical practice.

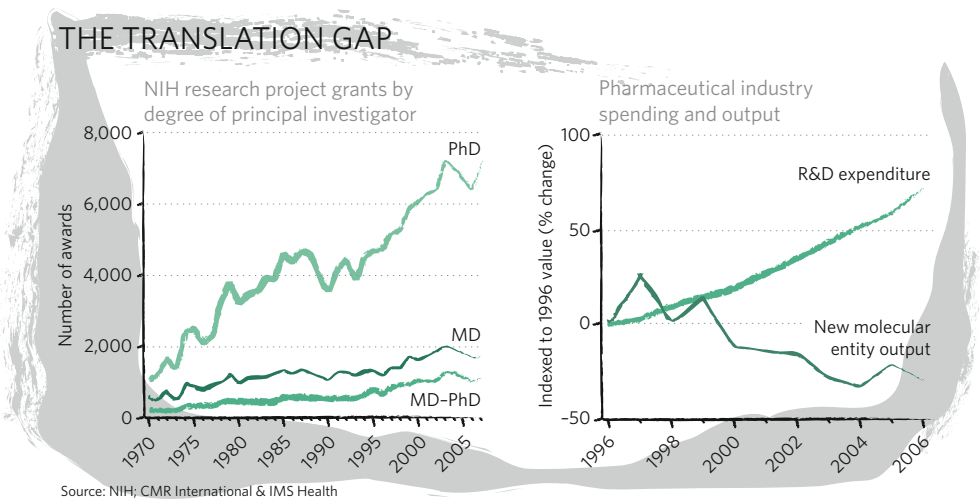
Out of the comfort zone

Biomedical research agencies are responsible for the first block. As anyone attempting translational research will testify, basic scientists have few incentives to move outside their comfort zone. It means getting involved with complex regulatory and patent issues. There is the risk of career damage to boot, because it is not the sort of research that gets published by the top journals and spurs promotion.

Publicly funded biomedical science has become disconnected from the processes that lead to cures and treatments, says Rudy Balling, a proponent of translational research at the Helmholtz Centre for Infection Research in Braunschweig, Germany. "Most biologists haven't a clue about real medical needs," he says, or about the difficulties of applying their research.

When Zerhouni became director of the NIH in 2002, "that's exactly the situation as I found it", he says. "There was a widening gap between basic and clinical research, which if left alone would have been a major barrier to progress." As head of the world's top-spending biomedical research agency, Zerhouni was under pressure to make progress. The NIH's budget had doubled since 1998 to $27 billion in 2003, and tax-payers were demanding a return on their investment. "That is the accountability factor that Congress is asking us to address," he says.

At the time, Zerhouni convened a series of whirlwind meetings with top clinicians and scientists who also realized they needed to change the way they worked so that existing



Source: NIH; CMR International & IMS Health

knowledge didn't end up sitting on shelves. These meetings convinced Zerhouni — a radiologist himself — that it was a priority "to get over this gap" by redesigning the agency's translational research programmes.

New lamps for old

The NIH already had projects under the old 'clinical research' label, including 78 General Clinical Research Centres (GCRCs) created in 1959 at universities and medical centres nationwide. But the centres were generally limited to providing services for conducting clinical trials. They did not tackle Zerhouni's new priority, spelt out in the roadmap, to boost the agency's ability to train physician–scientists and translational researchers capable of bridging the valley of death. The CTSCs will replace the GCRCs.

Science and innovation have become too complex for any nostalgic return to the physician–scientist on their own as the motor of health research. Reinventing that culture is therefore the focus of the CTSCs, in the form of larger, multidisciplinary groups, including both basic scientists and clinicians, but also bioinformaticians, statisticians, engineers and industry experts. Zerhouni says he expects them to be breeding grounds for a new corps of researchers who will effectively stand on the bridge and help others across. Scientists at the centres will be evaluated with business techniques, such as milestones and the ability to work in multidisciplinary groups, rather than by their publications alone.

NCRR/NIH

"The clinical and basic scientists don't really communicate."
— Barbara Alving

Since 2006, Fitzgerald's centre in Philadelphia is using its CTSC money to pull together 400 or so staff who were previously scattered across research centres and hospitals and install them in a new bricks-and-mortar institute. For researchers with work to translate, the new centre offers support with regulatory issues, patents and clinical trial design. Fitzgerald would ultimately like 20% of new medical-school graduates to follow translational research courses, and the centre also offers master's and other degrees in the new discipline for MDs and PhDs. One of Fitzgerald's programmes is exploring the aftermath of painkillers called COX2 inhibitors, which were more or less abandoned by the pharmaceutical industry after they were found to increase the risk of heart attack and stroke. Researchers at the centre are looking for biomarkers that might identify those who escape these side effects and salvage a future for the drugs.

Scientists in other countries are watching the NIH flagship effort with interest. In Britain, which is second only to the United States in biomedical research output, the government last year announced a doubling of the Medical Research Council's budget to almost £700 million (US$1.3 billion) by 2010, largely to finance a new focus on translational research. In Europe, around 20 national research and government agencies are exploring a European version of the CTSCs. Coordinated by Balling, the European Advanced Translational Infrastructure in Medicine wants to create a multimillion-euro network of biomedical translation hubs across Europe, based on existing research centres.

J. IDE/HARVARD UNIV. NEWS OFFICE

A new breed of researchers will aim to bridge the translational divide.

Time will tell whether the NIH's translational centres can come up with the goods. Gary Pisano, an expert in innovation at Harvard Business School, calls them "an experiment worth doing". Government support has been used with some success to further application of other research fields, he points out, such as defence funding that supported applied research in electronics, communications and the Internet.

Measuring the outcomes of translational research is notoriously difficult, as they do not lend themselves to the simplistic bean counting of publications. Because drug development can take up to 20 years, the eventual impact of such efforts on the drug pipeline will only emerge with time. At the NIH, Alving has set up a commission to advise how the CTSCs should be evaluated. This might be done by tracking researchers' career paths and surveying productivity by, for example, counting patents, clinical trials and collaborations with industry. But until patients see a benefit, the aims of the programme risk appearing laudable but vague.

Some basic scientists baulk at the $500 million annual costs of the centres when the NIH budget is under extreme pressure. But Zerhouni says there will be no significant diversion of resources to translational research and that the CTSCs will be funded largely by absorbing the $290 million budget of the old GCRCs. Some $95 million will come from the NIH's Common Fund, and the rest will be redirected from other clinical projects. Zerhouni says the NIH has a current balance of 60% basic and 30% clinical and argues that it needs more, not less, basic research to feed the translational pipeline. Others assert that the 30% clinical figure is artificially inflated because it classifies a proportion of work — such as that on animal models — as clinical that others would call basic, something Zerhouni denies.

With a tiny fraction of the NIH budget, and much of that shuffled from existing clinical programmes, critics might charge that the CTSCs are little more than business as usual. Schechter thinks that the NIH needs to go further down the translation road by reforming the monopoly of investigator-driven research grants as the agency's main funding mechanism. This system rewards individual success and does little to encourage the type of collaboration that translational research demands. He points to alternative models for doing translational research, such as the Multiple Myeloma Research Foundation, based in Norwalk, Connecticut, and other charitable groups that operate more like businesses in their drive to get research into clinical trials. "There are other structures for doing biomedical research than that which the NIH has hewed to for 40 years."

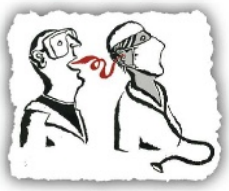
Zerhouni is sensitive to the need for reform, and points to new awards for multiple investigators. He acknowledges there is no 'right' model for translational research, but he is confident that the NIH will learn about the best ones by giving the CTSCs the freedom to explore a diversity of approaches. As to what the NIH stands for — National Institutes of Health, National Institutes of Biomedical Research or National Institutes of Basic Biomedical Research — "we are all of the above", says Zerhouni. And perhaps it will take many aliases and many attempts to cross this particular chasm. ■

Declan Butler is a senior reporter at *Nature*, based in France.

See also pages 823 and 843, and online at http://tinyurl.com/3tt3y3.

THE FULL CYCLE

Results can be thrust from bench to bedside, but there is also much to be learned by pushing the other way. **Heidi Ledford** tells tales of clinical trials that have prompted a change in tack.



In April this year, Nobel laureate Sydney Brenner brought the crowd to its feet at the American Association for Cancer Research meeting in San Diego, California. Brenner pioneered the use of the nematode *Caenorhabditis elegans* as a simple model for studying growth and development. But in his talk, he championed experiments on a more complicated creature: *Homo sapiens*. "We don't have to look for model organisms anymore because we are the model organism," he said.

Brenner is one of many scientists challenging the idea that translational research is just about carrying results from bench to bedside, arguing that the importance of reversing that polarity has been overlooked. "I'm advocating it go the other way," Brenner said. Bedside to bench means that clinical trials and patients' unexpected responses are valuable human experiments, and failed trials can stimulate new hypotheses that may help refine the experiment in its next iteration.

Reverse translation of this type comes with its own challenges, such as gaining access to clinical samples. And only certain trials, for example those the research community has already invested in, tend to get this type of scientific scrutiny. Below, *Nature* recounts three stories in which results from human experiments inspired new avenues of research.

"We don't have to look for model organisms anymore because we are the model organism."
— Sydney Brenner

A moving target

It was a front-page article in *The Boston Globe* that first caught cancer researcher Daniel Haber's attention. The 2003 article[1] described the strange results from an experimental cancer drug called gefitinib, one of the first generation of 'smart drugs' designed to target a specific protein, in this case one called epidermal growth factor receptor (EGFR).

Several types of tumour churn out higher than normal levels of EGFR, and gefitinib was designed to block the receptor and prevent further tumour growth. In 2003, the US Food and Drug Administration conditionally approved the drug, marketed by AstraZeneca as Iressa, to treat a severe lung cancer called non-small-cell carcinoma. And in some cases — typically non-smokers, women and Asians — the drug yielded dramatic results, seemingly eradicating signs of the disease from patients whose illness was thought to be terminal. "The response was magical," says Haber, director of the Massachusetts General Hospital Cancer Center in Boston.

But in other patients, as the *Globe* article explained, gefitinib was ineffective. Trials completed after the drug's approval showed that it failed to improve patients' survival when averaged across all individuals[2]. Regulators took the drug off the US market in 2005, although it is still approved in some other countries, and for certain exceptional cases in the United States.

To many, the drug's failure was a bitter disappointment. But Haber was intrigued and wondered whether there was a molecular explanation for the discrepancy in responses between one patient and the next. Down the road at the Dana Farber Cancer Research Institute, Matthew Meyerson and his colleagues were wondering the same thing. Neither group initially thought to look for answers in the sequence of the *EGFR* gene: the experiment was so obvious that both assumed it had already been done. "It was naivety," says Meyerson.

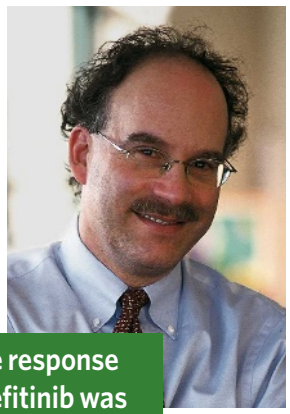
In 2004, Haber, Meyerson, and William Pao at the Memorial Sloan-Kettering Cancer Center in New York, independently published results showing that most tumours that responded to gefitinib harboured mutations in *EGFR* that rendered the protein more sensitive than usual to the drug[3–5]. As a result of these findings, researchers have developed genetic tests for *EGFR* mutations, and several clinical trials are now under way to determine whether the drug is effective when it is given only to the patients with a mutated receptor.

But there was another puzzle to solve. Early trials had shown that even those who did initially respond to the EGFR inhibitor soon become resistant, many within a year of starting treatment. By genetic analysis of tumour samples and cancer cell lines, researchers have found that these resistant tumours acquire secondary mutations that render the drug impotent.

ILLUSTRATIONS: B. MELLOR

Sometimes, the culprit is a second mutation in EGFR that stops the drug from binding to the protein; at other times, a gene called *MET* is also amplified, which allows the tumour cells to multiply even when EGFR is not working[6]. Occasionally both EGFR mutations and *MET* amplification arise in different cells within the same patient, illustrating how cancer cells will use every genetic trick in the book to continue growing. Findings such as these partly inspired the Cancer Genome Project, which aims to sequence genes from multiple cancers to reveal their genetic idiosyncrasies. And early clinical trials are under way to test whether gefitinib-resistant tumours that have accumulated a secondary mutation can be tackled with alternative drugs, such as EGFR inhibitors that bind more tightly to the protein and are unaffected by the mutations.

J. TOUSTER



"The response to gefitinib was magical."
— Daniel Haber

Few failed drugs receive so much scrutiny. Gefitinib was a special case because it had been designed specifically for a target, and because it produced such a 'magical' response in some patients.

One of the major obstacles in bedside-to-bench research is obtaining high-quality tissue samples from trials. Few physicians are willing to collect the multiple invasive biopsies that are needed to determine the molecular changes in a tumour as it evolves, but that offer no direct benefit to the patient. Meyerson had to track down the investigators who had tested the drug and apply for separate approval to use the samples from the many ethics committees who approved the gefitinib study. "I didn't understand how hard it was to get samples from a clinical trial at that time," he says.

M. STEPHENS / PA PHOTOS

Tale of the unexpected

When leukaemia first developed in a child given gene therapy, there was still hope that it was just a coincidence. "We didn't know what to make of it," says Brian Sorrentino, who directs the gene-therapy programme at St Jude Children's Research Hospital in Memphis, Tennessee. "Then the second case came, and it was clear this was going to be a recurring problem."

Since 2002, 5 of the 21 children who received a high-profile, experimental gene-therapy treatment for a disease called X-linked severe combined immunodeficiency (X-SCID) have developed leukaemia. X-SCID is caused by a mutation in a gene, called *IL2RG*, that is required for the immune system to generate working T cells and other cells in the immune system. The condition is commonly called 'bubble boy disease' because babies with X-SCID must live in sterile environments to avoid lethal infections.

At first, the two X-SCID trials, in France and Britain, were heralded as the clearest success in the controversial gene-therapy field. Most patients developed a functional immune system, and the first recipients now live a normal life, says Marina Cavazzana-Calvo, a researcher at the Neckers Children's Hospital in Paris, France, who helped conduct the French trial. The technique relied on a retrovirus to shuttle a functional copy of *IL2RG* into the patient's bone marrow stem cells, from which immune cells are generated. Researchers expected that the retrovirus would integrate into the patient's genome at random. But shortly after the two trials started, Christopher Baum of Hannover Medical School in Germany, and his colleagues published a short report indicating that the virus preferentially inserted itself next to a cancer-causing gene, causing leukaemia in mouse models[7]. "The initial reaction was that our mouse model would not be relevant to the clinics," Baum says.

Once the third case of leukaemia came to light in 2005, the French trial was put on hold, sending a chill through the entire gene-therapy community. "It was a difficult time," says immunologist Frank Staal, who studies gene therapy at the Erasmus University Medical Center in Rotterdam, the Netherlands. For many clinical trials, such a disaster would mark an immediate end to the research. But because this therapy had looked so promising and because the disease is so devastating, researchers were anxious to find out what had gone wrong.

Gene therapy helped some children who were born without a working immune system.

Using a method to isolate and amplify only the regions of DNA that surrounded the virus, scientists have found that the insertions were far from random in the bone-marrow stem cells. The virus had multiplied and slotted into hundreds of different sites, preferring to settle near highly expressed genes[8,9]. In the French trial, some of the patients with leukaemia had viral inserts near possible cancer-causing genes such as *LMO2*, which is involved in blood-cell formation[10]. The researchers suspected that genetic elements in the virus that were used to activate *IL2RG* also stimulated expression of *LMO2* and other genes nearby the virus' insertion site. This probably caused the proliferation of T cells that caused the leukaemia.

Since then, researchers have been trying to get around this by modifying the viruses used to transfer the gene. Adrian Thrasher at University College London has designed a vector with genetic control regions that are less likely to activate nearby genes and that contains a 'kill switch' to prevent it from replicating once it has inserted into the genome[11]. Pending final approval, this vector will be used in the next round of gene-therapy trials for X-SCID, says Thrasher, who is one of those leading the trial.

Of the five children who developed leukaemia, four were successfully treated and their gene-therapy-repaired immune systems remained intact. Sorrentino notes that if the gene therapy itself had not been so successful, the community might not have rallied so readily to fix the leukaemia problem. "At the time it felt terribly depressing," he says. "But we've worked through that and now I feel very enthusiastic."

Test shot

By the time Merck's HIV vaccine candidate made it into a phase II trial, there was no shortage of strong opinions about its probable fate. Over the past few decades, the

HIV vaccine community had seen one failed attempt after another. Merck's approach — to stimulate a T-cell response, rather than one from antibodies — was new. So some voiced hope for the vaccine's success. Others pointed to the raging HIV epidemic as a reason to move any promising candidate into clinical testing. And yet more were downright pessimistic.

But no one expected the vaccine to make some study participants more susceptible to infection. "That really shocked the field," says Barton Haynes, from Duke University in Durham, North Carolina, and director of the Center for HIV-AIDS Vaccine Immunology (CHAVI).

Researchers remain at a loss to explain what went wrong with the trial, called STEP, which was halted last September. But they have been galvanized to find that explanation, and are lining up to access samples and trial data. Part of the willingness to explore the failure lies in the nature of the HIV vaccine field, which was born of 'bedside-to-bench' research, says Bruce Walker of Massachusetts General Hospital in Boston. "In the beginning, we didn't have any idea of what this disease was," he says. "It required going from the bedside back to the bench to figure it out." Plus the research community was already heavily involved in developing the vaccine and the trial was co-sponsored by the National Institutes of Health.

The fact that this vaccine increased the rate of infection injected a sense of urgency into the trial's post-mortem because of concern that other vaccines might do the same. "There is no more important question than determining what happened," says Mark Feinberg, Merck's vice-president of medical affairs and policy. "The implications for the field are enormous."

R. ESPINOSA/AP PHOTO

Problems with Merck's HIV vaccine trial inspired researchers to examine what went wrong.

Some have cautioned that another failure like the STEP trial could spell the end of the HIV vaccine quest altogether.

Preliminary analysis of the trial data has revealed that those who became more susceptible were typically uncircumcised males and carried pre-existing immunity to the virus used in the vaccine. That virus, a disabled form of a common cold virus called adenovirus serotype 5, was used to ferry HIV genes into the patient to elicit an immune response. But researchers do not yet know whether the pre-existing immunity was important in reducing resistance to HIV, or whether it is merely a surrogate for some other factor.

One hypothesis is that the increased susceptibility could have a genetic cause, and researchers in the CHAVI consortium have put forward a proposal to scan the genome of trial participants to look for such a genetic signature. Others want to determine whether immune responses to the vaccine's vector may have rendered participants more vulnerable to infection by fuelling an increase in the subset of T cells that HIV infects.

The HIV vaccine community has formed a scientific review panel to evaluate proposals for research that will use STEP samples, and Merck is planning to create a central repository of samples from the trial to facilitate their distribution. So far, more than 25 investigators have proposed projects, says Julie McElrath, an HIV researcher at the Fred Hutchinson Cancer Research Center in Seattle, Washington, who is helping to coordinate the effort. The challenge, McElrath says, lies in recruiting new basic researchers with fresh ideas to work with the samples because many do not know that they are available or how to go about getting their hands on them. "It seems like this huge barrier," she says.

Prompted by the failed trial, the US National Institute of Allergy and Infectious Diseases made it clear at an HIV vaccine summit in March that the institute will shift the balance of its vaccine funding away from clinical trials and towards basic research, in an effort to stimulate new approaches to vaccination. Those involved say that the shocking scale of the HIV epidemic demands a dogged approach that is sometimes absent from other areas of research. "In academia, we'll come up against a brick wall and we'll tend to move to a different area," says Haynes. "With the HIV vaccine, we have to keep looking at that problem."

> **"In academia, we'll come up against a brick wall and we'll tend to move to a different area."**
> — Barton Haynes

Learning lessons

Francesco Marincola, a cancer immunologist and advocate of translational medicine at the National Institutes of Health Clinical Center in Bethesda, Maryland, says that the level of organization and persistence found in the HIV vaccine community is rare. "There is a lack of that in other, more complex diseases," he says. One reason is that HIV researchers have a clear target — the disease-causing virus itself. Another is that industry and government-funded researchers collaborated on the vaccine's development, so the project wasn't just abandoned when the results turned out negative.

Other fields may begin to catch up. The current emphasis on developing biomarkers to monitor disease progression is encouraging more investigators to collect patient samples. Without them, it is difficult to look back and evaluate what went wrong if the trial fails. "That's the reason why we haven't learned from these failures," says Marincola. "Instead, we have gone from one failure to another." ■

Heidi Ledford is a reporter for *Nature* based in Cambridge, Massachusetts.

1. Mishra, R. A drug that works — for some. *The Boston Globe* (24 November 2003).
2. Metro, G. *et al. Rev. Recent Clin. Trials* **1,** 1-13 (2006).
3. Lynch, T. J. *et al. N. Engl. J. Med.* **350,** 2129-2139 (2004).
4. Paez, J. G. *et al. Science* **304,** 1497-1500 (2004).
5. Pao, W. *et al. Proc. Natl Acad. Sci. USA* **101,** 13306-13311 (2004).
6. Engelman, J. A. *et al. Science* **316,** 1039-1043 (2007).
7. Li, Z. *et al. Science* **296,** 497 (2002).
8. Deichmann, A. *et al. J. Clin. Invest.* **117,** 2225-2232 (2007).
9. Schwarzwaelder, K. *et al. J. Clin. Invest.* **117,** 2241-2249 (2007).
10. Hacein-Bey-Abina, S. *et al. Science* **302,** 415-419 (2003).
11. Thornhill, S. I. *et al. Mol. Ther.* **16,** 590-598 (2008).

See Editorial, page 823, and online at http://tinyurl.com/3tt3y3.

A CASE HISTORY

The Ludwig Institute for Cancer Research is focused on translating research into cures. **Helen Pearson** investigates whether its sometimes unusual methods are producing results.

B. MELLOR

There were two phones on the Baltic island: one in the barn near where Webster Cavenee was seeking seclusion to write up a paper, the other on a village telephone pole. Both were ringing. When Cavenee ignored the first, the villagers brought him messages from the second: a mysterious American was trying to contact him from Brazil, then England, then Australia. He paid them no heed: "Calls from a guy I didn't know, from an organization I didn't know anything about — I thought he was a crank."

His paper finished, Cavenee flew home to Ohio, where he was working as a geneticist at the University of Cincinnati College of Medicine. As he opened his mail, the mystery caller rang again. "He said, 'I'm calling about the job. I understand you're interested'," Cavenee recalls. "I said 'I don't know anything about your job'. He said 'No one's talked to you? We need to meet'." The man told Cavenee to go to a Hilton near Chicago O'Hare airport, and ask the manager to take him to the boardroom. Intrigued, Cavenee did as he was told.

Good call. That meeting in 1984 changed Cavenee's life. A doctor called Hugh Butt from the Mayo Clinic in Rochester, Minnesota, one of whose patients was the billionaire Daniel Ludwig, offered the researcher US$3 million a year for a new cancer research institute. "He was telling me they'd give me enough money that I could never blame failure on anyone but myself," Cavenee says. Cavenee accepted, and has worked for the Ludwig Institute for Cancer Research (LICR) ever since.

"You have to admit it's not the normal kind of job-recruitment process," says Cavenee. But then the LICR is not a very normal kind of research institute. Ludwig's legacy now supports Cavenee's research centre in San Diego, California, and eight more around the world (see 'Nine branches, seven countries'). It also — crucially, in the opinion of its leaders — spends between 15% and 20% of its annual $100-million outlay on an infrastructure that deals comprehensively with intellectual-property issues and clinical-trials management, and more on facilities that make some of the biological reagents that those trials require.

Clinical discovery

Those additional features are part of the LICR's formula for 'clinical discovery', a term that the chairman of the LICR's board of directors, Lloyd Old, prefers to the more modish 'translational research' and a goal he has pursued since the 1980s. In 20 years as the LICR's scientific director, and thus its effective chief executive, Old focused on strategies for getting cutting-edge science as close to the clinic as he could before turning it over to drug companies. "There is no question that we are the model," he says. "I don't know another organization that saw it so clearly and put it in place."

S. DEANE / NFCR

"You have to admit it's not the normal kind of job recruitment process."
— Webster Cavenee

The LICR has indeed won itself some influential admirers, most recently collaborating with leaders at the National Medical Research Council in Singapore who are keen to tap into its infrastructure. It also has its critics. The centralized power structure that allowed Old to put his vision of clinical discovery into place also allowed him to pursue his own particular scientific enthusiasms, a pursuit on which the jury is still out. And although clandestine-feeling meetings in hotel rooms are no longer part of the institute's recruitment procedure, its decision-making is still opaque enough for others to find it hard to see what they might most usefully copy.

The same whiff of the hermetic may also harm the LICR itself in its ability to recruit or retain the best in the field. "While the LICR has many strong attributes, there is a certain level of the old-boys' network," says Jim Woodgett, who worked for the LICR centre in London and now directs the Samuel Lunenfeld Research

A. OBORN

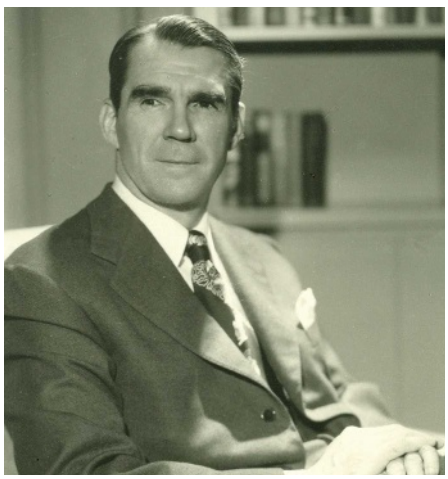
Institute in Toronto, Canada. "They should be willing to open themselves up to scrutiny."

A tendency to secrecy may be another part of the LICR's inheritance from Ludwig, a determined but publicity-shy entrepreneur who built up a tremendous fortune in shipping and other businesses. Why he decided to give so much money to cancer research is not known; one anecdote puts it down to one-upmanship. When Richard Nixon declared a 'war on cancer' in 1971 and requested an extra $100 million for research, Ludwig, a friend of the president's, is said to have boasted "Hah, well I can beat that!" He founded the LICR that year with an endowment of slightly more than $500 million.

Ludwig reputedly liked to do business with his friends, and relied on business and scientific directors from his inner circle, such as Butt, when it came to setting up the institute. In the first phase of its life that inner circle did the hiring and firing, selecting and approaching people such as Cavenee to establish institutes and then leaving them to pursue their own direction more or less autonomously. At this point some of the decisions reflected Ludwig's other interests: a branch in Brazil was connected to his massive land purchases in the Amazon. By the late 1980s and early 1990s, the institute had established an international reputation for sterling basic cancer research in areas such as signal transduction.

It was in 1988 that Old, initially recruited onto the LICR board while he was working at the Memorial Sloan-Kettering Cancer Center in New York, became the LICR's scientific director. His interest in cancer was partly spurred by his family: his grandfather, mother, father and aunt all developed cancer. His lab at Sloan-Kettering played a part in some seminal discoveries, including the discovery of tumour necrosis factor (TNF), a molecule central to inflammation and cell death.

Now 75, Old is recovering from heart-bypass surgery in his Upper West Side Manhattan apartment, with parquet floors, a grand piano and French windows that open onto West End Avenue. Although no longer the LICR's scientific director, he is still the chairman of its board and wields considerable influence. His voice is soft and hoarse from having had a tube down his throat, his hands shake and he has a tendency to ramble. But he leans forward excitedly when he talks about his passions, such as translational research and searching for the secrets of the immune system's successes — and failures — in fighting off cancer.



Daniel Ludwig's legacy: focused and well financed.

Old says he faced obstacles when he tried to carry out clinical work related to his discoveries at Sloan-Kettering. When he came in as the LICR's scientific director, "I wanted to create structures that I would have liked to be there for me when I came along." He set about making the rest of the organization take that desire on board. "Old said to the institute's branches, 'You do it or else'," says Eric Hoffman, who directs the LICR's Office of Clinical Trials Management.

Control is key

Old's mantra, taken up by his colleagues, has been that the key to translational research is control. 'Control' means retaining the intellectual-property rights and having a support staff that allows LICR researchers to do at least some early-phase clinical trials, rather than leaving a treatment's post-lab life entirely in the hands of pharmaceutical companies. That sort of control has made the institute one of the largest not-for-profit DNA patent holders in American biomedicine. It also means that LICR researchers can design trials to investigate a therapy's mechanisms, rather than just its performance or lack of it. That lets them explore speculative applications that would be ignored by a drug company with more obviously promising leads to hand. "That gives us immense power to get the job done," says Jonathan Skipper, director of intellectual property at the LICR. "You're not beholden to anyone". "They can do these wacky things and demonstrate proof of concept for something new," says Paul Bevan at biotech company Wilex in Munich, Germany, who is collaborating with the LICR.

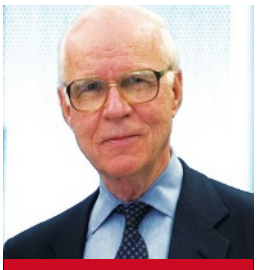
The idea is not to go it alone without pharmaceutical companies. It is to get therapies far enough down the road under the institute's control that they will be in pole position for transition to the clinic. "In many cases when you license it you lose control," says Joseph Tomaszewski, who is involved in translation of cancer discoveries at the National Cancer Institute (NCI) in Bethesda, Maryland. "So you want to take it as far forward as you can."

The other aspect of the LICR's 'control' is ensuring that researchers have access to their own supply of the physical reagents needed for their trials. The LICR owns facilities in Melbourne, Australia, and Ithaca, New York, dedicated to the production and storage of antibodies and other biological reagents. "Synthesizing a monoclonal antibody that can be put in humans

can cost a million dollars or more, a very expensive and tedious procedure that academic institutions usually can't afford," says Otmar Wiestler, chairman of the German Cancer Research Center, Heidelberg, who works with the the LICR. "The LICR is really unique in providing such an incredible infrastructure," he goes on. "I don't know any other academic institution that does it on this scale."

Take the institute's work on cancer vaccines that might provoke the immune system into killing tumours. Since the early 1990s, Thierry Boon, director of the LICR's Brussels branch, Old and others have identified a family of 'cancer/testis' antigens that are expressed almost exclusively in tumour cells and not healthy tissues. The LICR has agreements with GlaxoSmithKline to work on the commercial development of these and other antigens. But the institute has also retained the right to perform clinical trials itself, and the ability to make its own antigens to that end. Vincent Brichard, who deals with the vaccine effort for GlaxoSmithKline in Rixensart, Belgium, agrees that trials performed by the LICR tackle scientific questions that have not figured in GlaxoSmithKline's development programme, but may yet prove crucial in the company's decision to develop an antigen.

A. OBORN



"There is no question that we are the model."
— Lloyd Old

A monoclonal antibody called G250 provides another example of the Ludwig approach. G250, which was discovered not by LICR researchers but at the University of Leiden in the Netherlands, binds to clear-cell renal carcinoma, a common and aggressive form of kidney cancer. Old's lab did much of the work that characterized its behaviour, hoping it might have a therapeutic role.

Chaitanya Divgi, who once worked in Old's lab at the LICR, and is now at the University of Pennsylvania in Philadelphia, wanted to test the idea that the antibody could be used to identify which kidney masses are malignant clear-cell carcinoma as opposed to other, less dangerous tumours. This is something that currently requires surgery and a biopsy. It was not one of the applications being pursued by Wilex, the company to which the antibodies are licensed by the LICR, so the LICR funded an independent trial. Divgi showed that the antibody attached to kidney masses in 15 out of 16 patients who were subsequently found to have clear-cell renal carcinoma — and in none of those whose kidney masses were more benign[1]. A commercial assessment spurred by these results has convinced Wilex to launch a phase III trial of G250 as a diagnostic agent alongside its trial on the antibody as a possible therapy for reducing recurrence of the disease. "We would have struggled to find the resource and commitment to do that [initial study] alone," says Bevan.

Privileged position

A visit to the LICR's plush New York office (above right) reveals that ten metres of shelving is devoted to regulatory paperwork required to test a new agent in humans: just short of a metre for every one of the 11 investigational new drug applications the institute holds in the United States. That shelf explains exactly why translational research looks uninteresting, frustrating or overwhelming to many — getting a satisfied buzz out of rules and regulations takes a certain type. And the LICR has employed 15 such types to deal with clinical-trial paperwork alone.

The LICR owes this infrastructure not just to Old's vision; the "historically privileged position", as Old puts it, of the institute's bank balance is crucial. Its endowment of $1.4 billion is not in the same league as those of the Wellcome Trust, the Howard Hughes Medical Institute, or the Bill & Melinda Gates Foundation. But it is substantial, focused on cancer and entirely under the control of a board of like-minded trustees: there is no peer-review process. Old and Andrew Simpson, who has succeeded him as scientific director, suggest that other institutions in privileged positions — such as Harvard, with its $34 billion endowment — could learn something from the LICR's model. (In May, Harvard announced that it would be receiving $117.5 million over 5 years from the National Institutes of Health in Bethesda, Maryland, for a new centralized translational research effort in which the university is concomitantly investing $15 million annually.) But the LICR is hardly proselytizing. The only time that the central office has published anything about its vision and infrastructure was in a 2004 letter to *Nature Medicine*[2]. Old says he "was taught not to toot his own horn", although perhaps he is now overcoming that childhood stricture a little.

Becoming modesty can edge into opacity. It is unclear, even to some of those who work for the organization, how top-level decisions are made and priorities set. "It's fairly autocratic in the way it's run," says Gerard Evan, who worked at the now defunct Cambridge, UK, branch

Nine branches, seven countries

Location	Director	Research focus
Brussels, Belgium	Thierry Boon	Human cancer genetics and cancer immunology.
Lausanne, Switzerland	Jean-Charles Cerottini	Basic and applied immunology in the treatment of cancer.
Oxford, UK	Xin Lu	Signal transduction processes in cancer.
Melbourne, Australia	Antony Burgess	Tumour molecular biology.
New York	Lloyd Old	Immunotherapies for human cancer.
San Diego, California	Webster Cavenee	Genetic changes during progression of cancer.
São Paulo, Brazil	Luisa L. Villa	Cancer genetics and markers.
Stockholm, Sweden	Ralf Pettersson	Aberrant signal transduction in cancer cells.
Uppsala, Sweden	Carl-Henrik Heldin	Signal transduction pathways and cell growth.

A. OBORN

in the 1980s and is now at the University of California, San Francisco. Old and Simpson acknowledge that big mistakes were made in the 1980s, when five branches were shut on the basis of business decisions by the board. "I'd go in every day and check the fax machine to see if I was still in business," Cavenee recalls.

More recently, some researchers who worked in the London branch when it was directed by Michael Waterfield were dismayed by the decision to move it from University College London to Oxford after new director Xin Lu took over in 2005. "It's a little sad after all the good work," says Bart Vanhaesebroeck who worked at the London branch and is now at Queen Mary, University of London, with continued funding from the LICR. Simpson says that the branches have always been anchored around the research interests of the relevant director. Old says an organization ought to reflect its leadership.

Results matter

Some observers say the organization could benefit scientifically by being more open to review. "If they're really doing the right thing they shouldn't have to defend themselves so much against external review," Woodgett says. "It's hard to argue with yourself and that's the problem with the Ludwig." But Simpson defends the LICR approach. "There is a philosophy we'd lose if we didn't have that autocratic element," he says. "It's important for smaller more nimble organizations like ourselves to plough a different furrow. Judge us by our results, not by the way we're doing it."

What, then, are those results? The LICR has only one licensed therapeutic success to point to — granulocyte-macrophage colony-stimulating factor (GMCSF), which is used to stimulate production of white blood cells after chemotherapy — and that grew out of work done before Old's focus on translation got underway. That record may sound poor, but the development of cancer therapies takes time whoever is doing it. For example, the NCI set up its Rapid Access to Intervention Development (RAID) programme in 1998. Designed to take a basic discovery and put it through all the phases of toxicology, pharmacology, large animal studies, production and paperwork that is required for an investigational new drug, with about $100 million spent so far, RAID has got 32 of 121 candidates to the clinical-trial stage, according to Tomaszewski, who has been involved in the programme since its outset. This is not a field with quick wins, and by RAID standards the LICR's 120-odd clinical trials since 1996 look at least par for the course.

That said, Old's stated belief that various aspects of the LICR's work are now coming to fruition suggests that if its success rate doesn't climb in the coming decade there will be something amiss. But what? The translational approach that Old developed — or the particular enthusiasms at which he has directed it? Tony Burgess, who heads the LICR's Melbourne branch, says that Old's personal research interest has produced a clinical programme biased towards cancer vaccines — which have a history of failure — and immunology. "I think the investment is very risky," he says. "I happen to think it might help some people but not many." If risky ideas don't pan out the efficiency of their translation is moot.

Old says that the institute will be judged on the "unyielding yardstick of human benefit". But interest and emulation from admirers may offer an interim tape measure. "They are small and elite in a positive way," says Edward Holmes, chairman of the National Medical Research Council in Singapore, where a new LICR institute is now being planned. Singapore has made an aggressive investment in biomedical science over the last five years and, for the last two, has been pushing particularly hard to develop a more translational and clinical side. The LICR and Singapore government has agreed to jointly invest $7 million per year to establish an LICR branch there focused on clinical research.

Cavenee is on the recruitment committee for the director of the Singapore branch — which is proceeding in a comparatively standard way with advertisements and word-of-mouth recommendations. Having built up two LICR branches, first in Montreal, Canada, then in San Diego, and trained more than 300 young scientists since his own recruitment, he's grateful for the opportunities the LICR has offered that other settings might not have. The organization's support, worldwide expertise and provision of enough money did indeed remove any external risk of failure. "There was never a dull moment," he says happily. That unyielding yardstick, though, does not care about dullness — and it has yet to measure the interesting institute's success. ■

Helen Pearson is *Nature*'s biology features editor.

1. Divgi, C. *et al. Lancet Oncol.* **8,** 304–310 (2007).
2. Skipper, J., Hoffman, E. W., O'Donnell-Tormey, J. & Old, L. J. *Nature Med.* **10,** 1154–1155 (2004).

See Editorial, page 823, and online at http://tinyurl.com/3tt3y3.

CORRESPONDENCE

These letters respond to the Commentary 'Europe's research system must change' by Luke Georghiou (*Nature* **452,** 935–936; 2008).

European research system must not go bananas

SIR — Luke Georghiou's Commentary may be applauded by science policy-makers, but it sent shivers down my spine as a scientist. Curiosity-driven research is under siege from those who claim that they know best how to "attain Europe's economic, social and environmental goals", and to "engage research with the problems that society recognizes as central".

There is nothing apparently wrong with such clichés, but for me they evoked an image of a tribe of chimpanzees sitting in banana trees pondering their societal needs. The issue at stake is to find a better technology for banana peeling. To climb down the tree and start walking would be viewed by the group as either a priority or politically correct. Likewise, the Commentary is full of populist, politically correct goals such as carbon capture, and leaves aside difficult ones such as nuclear energy.

This could not be further removed from the message of another Commentary 'The unpaid debt' (*Nature Phys.* **3,** 824–825; 2007). That described how the fundamental research that usually serves — let's face it — very obscure goals can nevertheless "yield unforeseen benefits of great value for society". Fortunately, there are still some agencies and a few politicians who do their best to fund blue-sky, seemingly useless research — but they are under attack.

Nature is uniquely placed to help them by conveying the message that a successful society must commit a significant portion of its resources to curiosity-driven research and not ask for anything apparently useful in return, as long as it is assured that qualified people do the work. Otherwise, we shall never learn how to climb down those banana trees that are undisputedly important and recognizable to society.

Andre Geim Centre for Mesoscience and Nanotechnology, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Large projects can create useful partnerships

SIR — Luke Georghiou urges reforms to the European Commission's research system, criticizing the present Framework Programmes for their many small and very loosely connected projects.

The large integrated projects of the Sixth Framework Programme are a notable exception. These include the ALARM project, intended to assess large-scale environmental risks for biodiversity using tested methods (www.alarmproject.net). This encompasses 68 partner organizations, with some 250 participants from 35 different countries. It focuses on the combined risks derived from climate change, environmental chemicals, biological invasions and pollinator loss against their socio-economic background.

Such large projects within the Framework Programmes offer opportunities for productive partnerships. It may look like locking scientists in a room and expecting them to get along — which Paul Jeffrey is quoted as saying won't happen, in the related News Feature 'With all good intentions' (*Nature* **452,** 682–684; 2008). But our experience in ALARM is that, if you give scientists sufficient choice, new and productive partnerships emerge, and their success increases with project size and collaboration options.

The consortium agreements required in the Sixth Framework Programme's integrated projects fit the model agreements you refer to in your Editorial on the subject, 'The path to productive partnerships' (*Nature* **452,** 665; 2008). For ALARM, they were signed by all partners before the project started. Laying down rules may seem unnecessary as members cooperate to avoid adverse consequences. But this may be different without such rules.

We therefore disagree with Georghiou's view that the Framework Programmes have to go. We favour more projects of variable sizes, organized through work plans and accompanied by model agreements — including a reasonable proportion of large integrated projects to create opportunities for interdisciplinary and productive partnerships.

Josef Settele, Joachim Spangenberg, Ingolf Kühn UFZ, Helmholtz Centre for Environmental Research, Theodor-Lieser-Strasse 4, 06120 Halle, Germany

European research needs a dash of anarchy

SIR — As Luke Georghiou urges, 'Europe's research system must change' — but most certainly not in the direction he proposes. He wants to strengthen strategic and applied research within the EU Framework Programme, advocating direct political influence on research into "problems that society recognizes as central" by creating even larger directed research programmes than we have now. But problems can arise from forced collaborations and top-to-bottom direction in large research consortia, as your recent Editorial and News Feature on the subject make clear (*Nature* **452,** 665 and 682-684; 2008).

Almost every significant breakthrough in the history of science has come about by serendipity — not as a result of strategic planning or problem-oriented and directed research. To foster real innovation in Europe's science, he should be advocating bottom-up research by small and innovative groups.

Young investigators, even those with brilliant ideas and a novel approach, are unlikely to be invited to participate in huge directed consortia. These mostly involve laboratories that have already established their name and fame, and are now often comfortably operating on well-worn tracks or working opportunistically on headline-grabbing problems or fashionable topics.

Science and innovation are chaotic, stochastic processes that cannot be governed and controlled by desk-bound planners and politicians, whatever their intentions. Good scientists are by definition anarchists, who don't want to be managed by what Gottfried Schatz of Biocenter Basel calls 'chronoclasts' — people whose bureaucracy steals their research time and blunts their creative potential. Good science has an inherent potential for self-organization.

Let's have non-bureaucratic, generous research support for the broadest possible palette of basic science problems. That is what would change Europe's research system for the better.

Theo Wallimann Institute of Cell Biology, ETH Zurich, Hoenggerberg, HPM D24.1, Schafmattstrasse 18, 8093 Zurich, Switzerland

COMMENTARY

B. MELLOR

Follow the leader

If Japan is to become a front-runner in pharmaceutical development, it must not only speed up its approval of new drugs, but also enhance its own research capabilities, argue **Kaori Tsuji** and **Kiichiro Tsutani**.

How long should a therapy take to journey from lab bench to bedside? The obvious need to prove the safety and efficacy of an investigational therapeutic product has notoriously led to delays that that can be aggravating for both drug developers and patients. The time taken between general acceptance of a new drug and its official approval, the 'drug lag', has been particularly pronounced in Japan, where drugs approved elsewhere in the world must undergo additional regulatory review, often requiring clinical trials performed in Japanese patients to supplement safety and efficacy data that are deemed sufficient for approval in the United States or by countries in Europe.

Japan's procedures for approving drugs can add years of costly research and development. Unsurprisingly this has historically discouraged investment in the introduction of new drugs to Japan despite the country being the second largest pharmaceutical market in the world. With a dearth of new drugs, public opinion has begun to turn against the policies and practices that create Japan's drug lag.

The Japanese Ministry of Health, Labour and Welfare (MHLW) has been struggling with the issue for several years. More than three years ago it established a special committee to review promising drugs that were being approved elsewhere in the world and foster fast turnaround in Japan. Additionally, in April 2007, the government launched a 5-year plan aimed at cutting the approval lag by 2.5 years, supposedly bringing the system in line with those of US and European regulators by 2011. The ¥1.75-billion (US$16.6-million) investment will be used to develop 40 facilities to perform or expedite clinical trials and add 240 new drug reviewers to the Pharmaceuticals and Medical Devices Agency, which handles drug approval[1].

> "More proactive measures must be considered to solve the drug-lag situation in Japan."

All of this represents a laudable effort to bring Japanese clinical science up to speed with its impressive basic research efforts, but if the past three years of activity by the MHLW special committee are any indication, governmental measures will not be enough to close the gap.

Few studies have quantified the Japanese drug lag. A 2006 review carried out by the Japan Pharmaceutical Manufacturers Association found that only 60 of the 88 top-selling pharmaceutical products in 2004 were available in Japan. The study showed that the relative delay in approval of these drugs was 2.5 years when compared with the United States or the United Kingdom[2]. These estimates, on which the MHLW targets were developed, only cover top-selling products, however.

The real picture

We conducted a comprehensive study covering 370 new chemical entities (NCEs) that were approved either in the United States, European Union (EU) or Japan between 1999 and 2006. The study showed the whole picture of drug lag in Japan including important but small-market drugs such as 'orphan drugs' for rare diseases. Of 370 NCEs, 304 (82%) were approved in the United States, 280 (76%) in the EU and 192 (52%) in Japan. The mean approval lag for Japan during the studied period was 3.1 years after approval in the United States or the EU. An analysis of annual change showed that the gap widened to 3.9 years in 2005 and 6.2 years in 2006 compared with the United States[3]. Lag was even longer for new biopharmaceuticals — therapeutics derived using living

organisms or cells. Top-selling drugs such as pegfilgrastim, an immune-stimulating factor, and cetuximab, an anticancer antibody, aren't yet available in Japan. Adalimumab, a popular antitumour necrosis factor antibody for rheumatoid arthritis was approved in April in Japan. The US Food and Drug Administration approved adalimumab at the end of 2002. As of 2006, only 22 of 65 new biopharmaceuticals were available in Japan and the lag behind the United States was approximately four years[4].

What three years has done

The aim of the Committee for the Use of Unapproved Drugs, established by the MHLW in January 2005, is to accelerate the development of promising new candidates and carry them to early approval in Japan. The committee reviews drugs that have been approved either in the United States, the United Kingdom, Germany or France. It also reviews petitions for early drug approval submitted to the MHLW by patient or academic groups. The committee identifies drugs that are needed in Japan and urges the company holding the rights to develop the drug to start or accelerate the development in Japan.

The efforts had a slight effect in speeding drugs to approval in the three years we observed. Between January 2005 and December 2007, the committee reviewed 169 unapproved drugs. It recommended 42 for accelerated development.

The petitions seemed to have only a small effect on approval. Of the 169 drugs reviewed, more than half (89) were submitted by petition and 30 of these were recommended for accelerated development. Of these, 18 drugs reached approval or pre-registration by the end of 2007, but ten had already been in development before the review. For the eight not already under development, the MHLW approached the companies involved and requested that they initiate action for getting approval, welcoming the new drug applications even with limited Japanese clinical data. The remaining 12 drugs petitioned and recommended by the committee were not in development as of 31 December, 2007 (see graphic). For some of these drugs, the MHLW are still looking for a company that will take on development.

The committee's action was effective for the eight drugs that were not in development at the time of the committee review. However, these drugs represent only a part of what is needed in Japan. It is noteworthy that no action was taken for 59 of the 89 drugs for which petitions were submitted. By the end of 2007, 44 of these still hadn't made it into development.

The recent policies and measures undertaken by the Japanese government might reduce drug lag, but considering the significant backlog of unapproved drugs that aren't even in development, the situation isn't likely to improve quickly. The MHLW target of reducing drug lag by 2.5 years seems to have been based on analyses of top-selling drugs rather than a comprehensive review of the problem. This target should be re-evaluated post-haste.

Further measures

Even if more ambitious targets are met, other measures are needed. The processes established by the Committee for the Use of Unapproved Drugs react to the approval of drugs elsewhere in the world. These measures maintain Japan's position as a follower of pharmaceutical development rather than a leader. More proactive measures must be considered to solve the drug-lag situation in Japan, such as developing Japan's own clinical-research capabilities and promoting licensing and partnering opportunities between Japanese companies and those based elsewhere in the world. If worldwide licensors could find partners in Japan earlier on, drug development could start earlier. Ideally, Japan might be involved in multinational clinical trials and simultaneous new drug applications. This cannot be done through political measures alone; Japanese pharmaceutical companies need to be more sensitive to information about promising new drug candidates around the world. A kind of public–private partnership to share information about such candidates might be useful.

Moreover, measures should enhance the attractiveness of the Japanese pharmaceutical market. Although Japan is the second largest pharmaceutical market in the world, the labour-intensive nature of the clinical development and of the creation of the new-drug-applications data package cuts deep into potential profits. Efforts should be made to reduce time-consuming paperwork and to translate English documents. Finally, regulators must approach the sensitive matter of price control with new analytical rigour. Although keeping the cost of drugs down is a worthy cause, the price the Japanese pay for not having access to certain drugs could be greater.

The drug lag in Japan is a result of complex factors and cannot be solved easily. Continuous efforts to correct the situation are necessary for the sake of the nation's health. ■

Kaori Tsuji is at the Health Care Science Institute, Akasaka NOA 5F, 3-2-12 Akasaka, Minato-ku, Tokyo 107-0052, Japan. Kiichiro Tsutani is at the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

1. McCurry, J. *Lancet* **369,** 1333–1336 (2007).
2. Fukuhara, H. *Iyakusangyoseisaku Kenkyujo* (Office of Pharmaceutical Industry Research, Research Paper No. 31, 2006) [in Japanese].
3. Tsuji, K. & Tsutani, K. *Jpn. J. Clin. Pharmacol. Therapeut.* **38,** S198 (2007) [in Japanese].
4. Tsuji, K. & Tsutani, K. *Eur. J. Pharm. Biopharm.* **68,** 496–502 (2008).

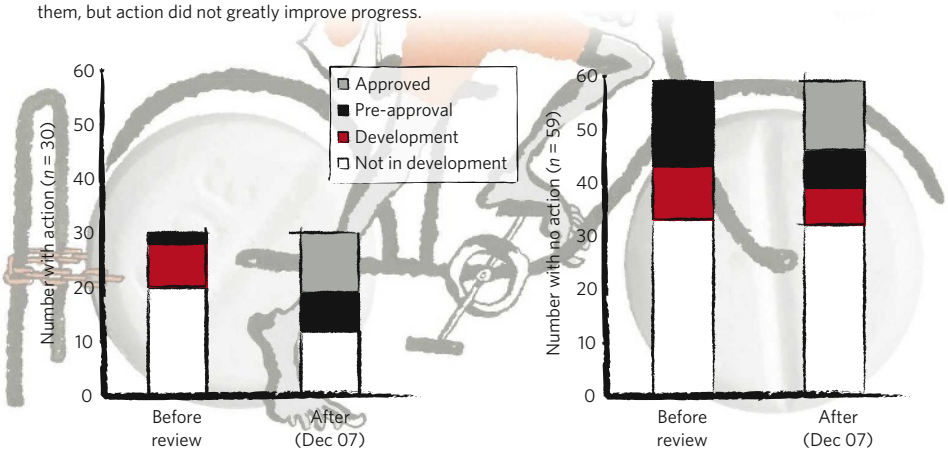
Kaori Tsuji consults for RHC USA Corporation, a US-based drug distributor in Japan.

See Editorial, page 823, and online at http://tinyurl.com/3tt3y3.

"If worldwide licensors could find partners in Japan earlier on, drug development could start earlier."

DID PETITIONING HELP CLOSE THE GAP?

Patient and academic groups submitted petitions to a government-sanctioned committee asking them to accelerate the approval of 89 drugs. The committee took action on 30 of them, but action did not greatly improve progress.



COMMENTARY

A new relationship

Fuelling innovation requires a different kind of collaboration between industrial and academic researchers, argues **Bill Destler**.

A. MEEHAN

The standard of living in Western Europe and the United States has been sustained for several decades by new products, services and businesses, the result of leading-edge research and development (R&D). Nevertheless, competitive cost cutting has forced the elimination of all but the shortest-term private-sector R&D programmes. Major corporate laboratories, once remarkable catalysts for economic development, have all but vanished. RCA Laboratories, for example, no longer exists. Bell Labs in Murray Hill, New Jersey, once the foremost corporate research laboratory in the world, saw its budget reduced from US$3 billion in 1982 to $1.3 billion in 2005 (actual dollars not adjusted for inflation).

Meanwhile, corporate competitors in Asia and elsewhere, taking advantage of lower labour costs in many cases, have increased their R&D efforts to the point at which many are now superior to companies in the industrialized West in both product quality and productivity[1]. As a result, both Europe and the United States may soon find that the majority of the new ideas for products and services that result in economic growth are, for the first time in modern history, generated elsewhere.

B. MELLOR

Competitive advantage

Does the West have any competitive advantages that it could exploit to correct this trend? Well, the institutions of higher education in the United States and in Western Europe are still regarded as the finest in the world. The 2007 World University Rankings[2] by Shanghai Jiao Tong University in China, for example, list 48 North American and European institutions in the top 50, and together they possess an unmatched reservoir of intellectual talent and creativity. Graduate students, moreover, are still the most cost-effective R&D labour force anywhere, and with most faculty member salaries typically paid by colleges and universities for instructional services, the costs of adding this intellectual firepower to R&D projects is minimal. In addition, many colleges and universities have laboratory assets that would be prohibitively expensive for most companies to reproduce.

> "Graduate students are still the most cost-effective R&D labour force anywhere."

Academic obstacles

So why haven't US and European corporations adopted colleges and universities as their corporate R&D centres? Why aren't more technology-based companies incubating their new product concepts at universities? Both parties are to blame. On the academic side, a number of factors work against the best interests of all involved.

First there is the 'Gatorade' factor: a small number of colleges and universities, mostly in the United States, have reaped financial windfalls from intellectual property developed by their faculty members. The University of Florida, for example, has received more than $150 million in royalties for licensing its Gatorade drink formula since creating it in 1965. This dream of significant financial return has led colleges and universities to demand intellectual-property rights and subsequent royalty payments from the corporate sector with such vigour that many collaborative projects are terminated by lawyers before they even begin. University faculty members and administrators typically have no idea of what it takes to move an idea to the point at which it can be turned into a new product or service for which there is commercial demand, and this naivety can make negotiations with academia over intellectual-property and

royalty issues an exercise in frustration.

Then there's the 'give us the money and we'll work on something related to your interest' factor: faculty members are usually looking for support for their own ideas, not those of others, and that causes many corporate executives to wonder what it is they are funding.

Finally there's the 'fund me for three years and I'll give you a progress report' factor: academic timescales are typically much longer than corporations can tolerate, especially when they are under competitive pressure from abroad. Companies are not in business to fund PhD dissertations. They have specific questions that need answers and they need them on a timescale of a year or less.

"Companies have specific questions that need answers and they need them within a year or less."

Corporate clogging

On the corporate side, there are also negative factors at work.

One of those is the 'next quarter's bottom line' factor: the increased expectations of corporate boards and shareholders for short-term profit growth has made it difficult for managers to justify significant investments in medium- and long-term R&D projects. It is amazing that people still question Toyota's multi-year, billion-dollar commitment to hybrid-vehicle development, a commitment US and European car companies were unwilling to make because of the long lead-time that was necessary before any profits could be realized. This was a technology, by the way, that was pioneered in the United States.

There is also the 'we'll buy any new technology we need' factor: corporate acquisitions and mergers are consuming untold dollars and euros in legal costs without adding any really new intellectual assets to US and Western European inventories. These funds could be better used to fund technology R&D to help maintain a position of technological leadership.

Additionally, there is the 'we won't pay overheads' factor: many companies refuse to acknowledge the very real costs that universities must bear to support R&D projects. They often baulk at paying overhead rates at academic institutions that are typically 50% or less of direct cost (such as salaries to researchers, materials and equipment), despite the fact that many companies have internal overhead rates of more than 100% of direct cost on internal corporate R&D projects.

Brave new world

There is a potential solution. Imagine, if you will, a group of colleges and universities that decide to make their faculty members and staff, graduate and undergraduate students, and facilities available to companies to carry out short- and medium-term corporate R&D projects at low cost and without the intellectual-property fights that usually derail such efforts. Imagine a new relationship between business and academia in which hundreds of companies discover that they can once again afford to do new product R&D, while identifying future employees at the same time.

Imagine if companies were able to submit candidate short-term (about 1 year) R&D projects to a central receiving point where they could be matched up with qualified faculty members and students at participating colleges and universities. Graduate students, under the supervision of a faculty member and a corporate representative, might be assigned to work on each project as a thesis problem. Faculty members and students who sign up for such projects would therefore have a real interest in the problem, and the ongoing interaction between the faculty members, students and corporate representatives would be of benefit to all three groups. Suppose, in addition, that the participating colleges and universities agree to accept a modest up-front payment, to be shared by the students, faculty and the institution, in return for relinquishing any intellectual-property rights associated with the work to the sponsoring company.

Such an arrangement could re-energize corporate R&D, but only if business and academia join together to craft template agreements in which all parties see real benefits to their organizations as a result of the collaboration. At the Rochester Institute of Technology (RIT) in New York, we are working with several corporate representatives, and have developed such an agreement in which the university receives a flat fee for a year-long project. The fee includes the costs of faculty-member and student time, access to university facilities, university overheads, and a modest payment in lieu of any future university intellectual-property claims associated with the work. Funding for supplies and equipment not already in hand is added to the flat fee. Proposed projects come from either corporations or university faculty members, and projects are accepted when a match between the company's research needs and university capabilities is evident. Participation by all parties (companies, universities, faculty members and students) is voluntary, and the university reserves the right to publish the results of the research after an appropriate delay (usually not to exceed one year) to enable the company to pursue intellectual-property protection.

"The West's future economic prosperity may depend on its success in exploiting institutions of higher education."

Such a programme assumes that the university does not already own significant intellectual property in the area of the participating company's interests. If previous work by university faculty members and students has resulted in university-owned patents or copyrighted work prior to the agreement with the corporation, a more traditional licensing agreement could cover that part of the work and any new intellectual property generated could be covered by the template agreement. The agreement could be readily adapted to different circumstances.

Since the RIT programme was announced in March 2008, three companies have already signed on to fund projects at the institute under the template agreement, with one (PAETEC, a telecommunications firm based in Rochester) committing to a total funding level of approximately $1 million over the next three years. Discussions with seven additional companies are currently under way.

Clearly, such arrangements would benefit both industry and academia. Although it is not a good thing in general to have the private sector dictate the intellectual directions that colleges and universities take, much university research today is of a very basic nature and there are benefits to connecting more of this activity with real-world needs. If universities and colleges are to become the economic engines in their communities that they aspire to be, then their R&D activities need to be focused both on traditional long-term fundamental research that advances human understanding and on shorter-term projects that have the potential to lead to new products and services.

The West must get on with it. Its future economic prosperity may well depend on its success in exploiting one of its last competitive advantages — the institutions of higher education and the extraordinary research and development assets that they represent. ■

Bill Destler is the president of the Rochester Institute of Technology, One Lomb Memorial Drive, Rochester, New York 14623-5603, USA.

1. Okimoto, D. I. & Nishi, Y. in *The Japanese Firm: Sources of Competitive Strength* (eds Aoki, M & Dore, R.) 178–207 (Clarendon, Oxford & New York, 1994).
2. http://ed.sjtu.edu.cn/rank/2007/ARWU2007_Top100.htm

For more details and to see the template agreement used in negotiation of these projects, visit http://www.rit.edu/research/srs/formsagreements/#agreements.

See Editorial, page 823 and online at http://tinyurl.com/3tt3y3.

BOOKS & ARTS

A prescription for public health

Will high drug prices and a lack of new medicines force the pharmaceutical industry to restructure and take a more personalized approach to research, asks **Merrill Goozner**.

Pharmaceutical executives will find nothing reasonable in the provocative view offered in *Reasonable Rx*. In this discussion of prescription drug (or Rx) policy, medical researcher Stan Finkelstein and economist Peter Temin of the Massachusetts Institute of Technology propose the break-up of pharmaceutical companies into two separate entities: one devoted to research and development (R&D) and the other to manufacturing and marketing. To broker between them, the authors would create a non-profit agency to establish research priorities, offer prizes to R&D winners and auction off the intellectual property to manufacturers, who would operate as generic companies.

Why is radical restructuring necessary? The first two-thirds of this short and accessible book takes readers through the standard critique of contemporary drug-industry practices. For years, the industry has justified its high prescription-drug prices — at least in the United States, where there are no price controls and purchasers have little bargaining power — by claiming they are a prerequisite for innovation. That policy has been an abject failure. The number of new drugs coming out of industry labs is declining, and high prices either deny patients the drugs they need or turn their lives into a constant scramble to come up with the money to pay for them.

The core of the problem, assert Finkelstein and Temin, is that perverse market signals are channelling most of the pharmaceutical industry's research skills into areas that improve health only marginally. In its search for blockbuster drugs, defined as those that make more than US$1 billion in annual sales, industry deploys marketers to medicalize trivial conditions such as enlarged prostate and heartburn. It pursues copy-cat drugs for chronic conditions such as high blood pressure and high cholesterol so it can gain a share in well-established markets. Meanwhile, its R&D managers ignore the infectious diseases that are ravaging the developing world, where tens of millions of poor patients cannot support research.

BAYER AG/AP

Drug manufacturers' focus on finding blockbuster drugs often provides only small gains in health.

Reasonable Rx: Solving the Drug Price Crisis
by Stan Finkelstein and Peter Temin
FT Press: 2008. 208 pp. $27.99

The authors point out that the current R&D model undermines one of the most promising areas of contemporary research: personalized medicine, a term that conjures up images of individualized treatments using promising technologies such as stem cells. But the more immediate opportunity lies in the growing recognition by researchers that many of the most intractable diseases facing ageing societies, such as many cancers and dementia, have multiple causes. The process of discovering drugs for their cure will divide a disease into subsets according to cause. Each subset will be treated more like a rare disease, an area that also lacks the attention of drug companies because of market failure.

Reasonable Rx highlights another personalized medicine opportunity that has been lost to the blockbuster-drug mentality. Industry has skimped on promising techniques for identifying patients who might experience side effects from drugs or who could benefit from drugs that help only a small fraction of people, such as the lung-cancer drug Iressa. It makes little sense to invest in tools or pursue strategies that will only shrink your market.

In their search for an answer to the conundrum of high drug prices and declining research productivity, the authors reject price control, stating that it "kills innovation". Thankfully, this point is not crucial to their case because it makes no sense. They are right that more than a third of the top-selling drugs originated from US-based companies, and that foreign firms have gravitated to the United States to take advantage of its excellent biomedical research infrastructure, which is mainly financed by US taxpayers. Although some of that R&D is innovative, a lot more of it, from a medical standpoint, is duplicative or wasteful. And even without price controls, the number of new drug approvals is taking a nosedive. Health-care systems around the world will always pay for unique medicines, so how much would the United States really lose by adopting the same price-control strategies used in the European Union, Canada and Japan?

The solution to getting medical innovation back on track, the book argues, is to transfer the risk of drug development from "sick people to society". The authors propose to do this by

establishing an independent, non-profit drug-development agency to acquire new drugs from private or public entities after they get regulatory approval. The agency would then auction the drug rights to manufacturing firms, who would bid based on prices it set. The auction revenue, supplemented by government funds generated by the lower cost of medicines for the health-care system, would cover the rewards given to R&D companies. To enforce the separation of the research and manufacturing arms, the authors propose that the agency would refuse applications from companies engaged in drug production.

Finkelstein and Temin argue that their proposal preserves free enterprise. They also tip their hats to Senator Bernie Sanders (Independent, Vermont), the socialist maverick who introduced legislation to establish a prize fund to spur drug innovation. They claim their scheme is better because of its auction component, but like Sanders they propose to replace exclusive marketing rights derived from patent monopolies with a prize system. The real innovation is their insistence on also changing the pharmaceutical industry's structure.

The authors should have spent more time wrestling with the finer points of their proposal instead of dwelling on the failures of the current system. For instance, to determine research priorities and set prize levels once market failure has been ushered off stage, their solution is to give the task to multidisciplinary committees set up by the new drug-development agency, which they compare with the grant-approval sections of the National Institutes of Health. The work would be peer-reviewed by groups such as the National Academy of Sciences or the US Pharmacopeia, a 188-year-old non-profit organization responsible for establishing formularies as part of the Medicare Prescription Drug, Improvement and Modernization Act of 2003. It added a prescription-drug benefit to Medicare, the US health-care programme for senior citizens.

This solution is curious given that Finkelstein, as revealed in the book, sat on the US Pharmacopeia–Medicare committee that set formularies under the act; drug firms successfully lobbied to influence the committee's decisions. It is unlikely to be any different when a panel is determining, for example, the relative value of a cure for Alzheimer's disease versus an incremental advance in treating dyspepsia. In the end, the perceived values of such cures determine the willingness of the private sector to pour resources into the hunt.

These quibbles could be thrashed out in the legislative arena if the book's ideas ever gain traction. The hard facts remain: drug prices are unsustainably high, new drug approvals are declining, and promising approaches are being ignored. By suggesting a way for public-health objectives to drive private biomedical research investment, Finkelstein and Temin offer the drug industry a path out of its current predicament. ■

Merrill Goozner is at the Center for Science in the Public Interest, 1875 Connecticut Avenue NW, Washington DC 20009, USA. He is author of *The $800 Million Pill* and writes at www.gooznews.com.

See Editorial, page 823 and online at http://tinyurl.com/3tt3y3.

Complementary cures tested

Trick or Treatment? Alternative Medicine on Trial
by Simon Singh and Edzard Ernst
Random House/Norton: 2008. 352 pp. £16.99/$25.95

The international market for alternative therapies is estimated at US$40 billion. Because so many people use alternative medicine, it provides an excellent vehicle for discussing the nature of scientific research. Yet explaining the evaluation of evidence, balance of probabilities and risk is not easy.

Combining their communication skills and knowledge, writer Simon Singh and professor of complementary medicine Edzard Ernst set out for the lay person the scientific approach to testing alternative medical treatments. *Trick or Treatment?* starts by detailing the development and evolution of the double-blind, placebo-controlled, randomized controlled trial and its role in evidence-based medicine. The authors evaluate the evidence for four common alternative therapies — acupuncture, homeopathy, herbal medicine and chiropractic. They discuss the pitfalls of placebo-based medicine and ask who is to be blamed for spreading misinformation about unproven treatments. The book concludes with a manifesto for better regulation of alternative medicine and reliance on properly tested therapies.

Trick or Treatment? is thoroughly researched and clearly written. Historical descriptions sit beside detailed and lucid evaluations of the research evidence. Some stories are well known, such as how naval surgeon James Lind developed the first clinical trial to test the effectiveness of lemons for treating scurvy. Others are less familiar, such as Florence Nightingale's aptitude for statistics and her development of a variant of the pie chart, the polar area chart, to support the case that good sanitation dramatically reduced deaths in military hospitals. The description of the Nazis' adoption of homeopathy is particularly compelling and sobering. These tales make the book entertaining as well as informative.

In the discussions of the four therapies, the authors' combined strengths shine through. The examination of the evidence is comprehensive, forensic and, for champions of these therapies, damning. For each treatment, Singh and Ernst present the available randomized controlled trials. They describe and dissect good-quality evidence and dismiss the poor-quality stuff, giving their reasons why it should be discounted. The authors conclude that acupuncture works as a short-term analgesic and can relieve nausea but not much else; that some herbs such as Devil's Claw for musculoskeletal pain or garlic for high cholesterol are effective; chiropractic can improve back pain but less well than conventional treatments; and that homeopathy is no better than placebo. They summarize evidence for a further 30 therapies, most of which they find wanting.

Singh and Ernst base their evaluations solely on results from randomized controlled trials. Many advocates of alternative treatments argue

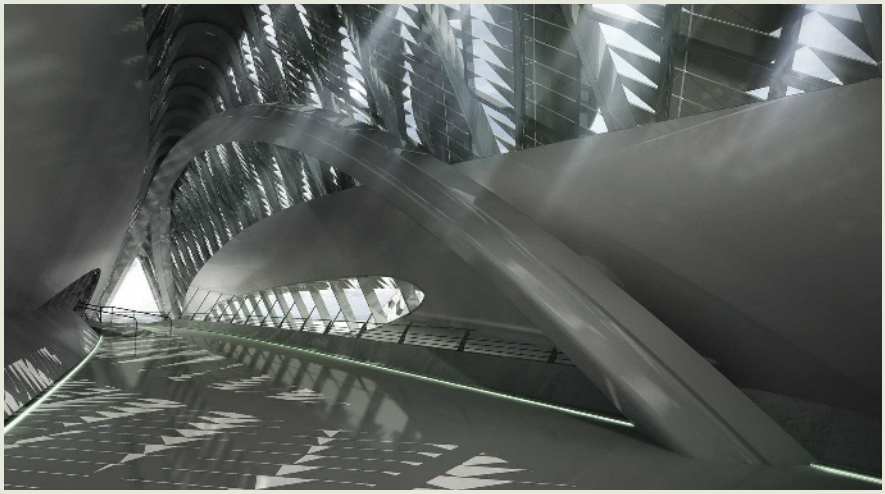
Acupuncture: pain relief or placebo?

J. FEINGERSH/GETTY IMAGES

that these trials are unsuitable. Some of these practitioners' arguments are easily dismissed, for example, the idea that alternative treatments are beyond science. Other criticisms come from respectable commentators and are harder to ignore; for instance, the difficulties of designing trials to investigate complex treatments with multiple variables, or whether these trials use test conditions that differ from a treatment as practised. Randomized controlled trials are powerful tools, but they are imperfect and it would have strengthened the argument of *Trick or Treatment?* had the book discussed these downsides.

Scientific research is intrinsically provisional; it may asymptotically approach a truth, but it is never unequivocal. Singh and Ernst, however, make repeated claims that they provide the truth, and have even included this word in the title of every chapter. The balance of evidence from randomized controlled trials supports their arguments, but the authors are not tendering a disprovable hypothesis. Many science communicators argue that to present science as the only truth does it a disservice. For now, the certainty expressed in *Trick or Treatment?* mirrors that of the proponents of alternative therapies, leaving each position as entrenched as ever. ■

Toby Murcott is a freelance science writer and broadcaster based in the UK. He is author of *The Whole Story: Alternative Medicine on Trial?*



M. GATTO

Saving water

Water and sustainable development is the theme of Expo Zaragoza 2008, a biennial international festival of culture to be held in Saragossa, Spain, from 14 June to 14 September. Inside a meander of the river Ebro, a park has been built displaying ecological materials, renewable energy and sustainable water management. The festival's 140 pavilions — including architect Zaha Hadid's sinuous bridge (pictured), clad in steel scales that mimic shark skin, and a glass tower shaped like a water droplet — will house exhibitions and events.

A platform for technical, scientific and social debate, the Expo will host nine weeks of themed seminars on water conservation, climate change and development. A series of essays has been commissioned by the Expo from global figures, including ex-president of the former Soviet Union Mikhail Gorbachev, EU High Representative Javier Solana and Nobel Peace prize winners Rigoberta Menchú and Wangari Maathai. **J.B.**

www.expozaragoza2008.es

Suppressing science

Doubt Is Their Product: How Industry's Assault on Science Threatens Your Health
by David Michaels
Oxford University Press: 2008. 384 pp. $27.95, £14.99

David Michaels has written a powerful, thorough indictment of the way big business has ignored, suppressed or distorted vital scientific evidence to the detriment of the public's health. *Doubt Is Their Product* catalogues numerous corporate misdemeanours, especially in the United States, from the criminal neglect of the dangerous nature of asbestos and the lies told by the tobacco industry, to the suppression of adverse findings of deaths caused by the anti-inflammatory drug Vioxx and the increased risk of suicide among teenagers taking selective serotonin re-uptake inhibitors for depression. The book concludes with a list of prescriptions for securing better regulation and greater protection for the public, mainly through increased public disclosure of vested interests.

The central question Michaels raises is whether our dependence on corporate funding in Western society can be reconciled with the integrity of scientific research and, if so, how. It can be argued that the importance of the motivation of a company or a scientist tends to be exaggerated. Our present system contains a strong element of self-regulation through self-interest. Companies make profits by manufacturing successful products that are useful to the public, and are damaged if their products are shown to be ineffective or harmful. They may face ruin if they cause disaster, as in the case of thalidomide. From time to time they err, but regulation keeps aberrations to a minimum. Big pharmaceutical companies, for example, have served public interest by producing a stream of drugs that has greatly improved the quality and length of our lives, as Michaels acknowledges.

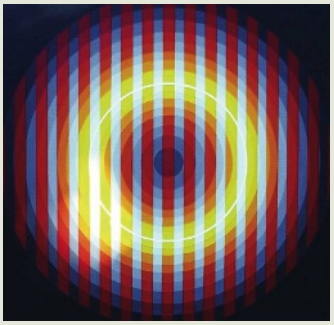
Individual scientists have reason to avoid dishonesty and incentives to ensure that their research stands up to scrutiny. Their reputations — and careers — depend on doing good science and suffer if findings are discredited. Corporate research is peer reviewed and results are accepted only when shown to be reproducible. Whatever its limitations, peer review is the best guarantee we have of research quality. These incentives and safeguards apply whether scientists work for companies, universities or the government. If the science is good, it survives; if not, it does not, whatever the funding source or the scientist's personal motive.

Yet, as Michaels demonstrates, motivation cannot be ignored. Canadian scientists examined papers on the controversial question of whether calcium-channel blockers used to treat high blood pressure increased the risk of heart attack. They found that, of those who supported the use of such blockers, 96% had a financial connection with the manufacturers. This compared with 60% of those who were neutral and 37% of those who were critical. Many studies of other drugs have found similar correlations between sponsorship and conclusions.

CULTURE DISH

PLAYING THE BUILDING
Artist and musician David Byrne, lead singer of the rock group Talking Heads, has transformed New York's Battery Maritime Building into a giant musical instrument. By tapping on the keys of an antique organ, visitors to the historic disused ferry terminal — open Fridays to Sundays until 10 August — can make its pipes, beams and spaces sing out harmoniously as installed devices strike, vibrate and blow through the building's structure.
www.davidbyrne.com

EYEING THE UNIVERSE
An inspirational array of astronomy photographs graces the Albert Dock in Liverpool, UK, this month. The exhibit will tour the world as part of the International Year of Astronomy 2009 celebrations after marking Liverpool's status as this year's European Capital of Culture. Pictures of planets, stars, nebulae and galaxies from observatories on Earth and in orbit, including the Hubble Space Telescope, were chosen by astronomers to convey the beauty of the Universe.
http://tinyurl.com/6meufd



H. GARCIA ROSSI

PAINTING WITH PERCEPTION
French optical artists from the Group of Visual Arts Research, who explored visual effects in painting in the 1960s, are the subject of an exhibition at the State Hermitage Museum in St Petersburg, Russia, running until 14 September. The artists painted optical effects and illusions on canvas, such as Horacio Garcia Rossi's *Electric Colour Light* (above), using knowledge of how the human eye works. By focusing on physiology and physics, their aim was to produce art that anyone can experience, regardless of cultural origin, status or education.
www.hermitagemuseum.org

In an ideal world, profit would not influence research. As a banner carried by an anti-globalization demonstrator at an international trade conference in Genoa proclaimed: "Capitalism should be replaced by something nicer." However, the socialist alternative, as exemplified by the former Soviet Union, did not prove to be 'nicer'. By stifling opposition to his ideas, Stalin's director of biology, Trofim Lysenko, did not promote good science, nor did he benefit the Soviet public.

Profit-pursuing corporations are not the only sinners to disrespect the integrity of science. Polls show that whereas the public distrusts scientists funded by industry, it respects those who work for environmental non-governmental organizations such as Greenpeace and Friends of the Earth. They are assumed to be objective and public spirited. Yet those bodies also have their own agendas, namely to promote their causes by increasing membership; they know that the most effective appeal is through scare stories. They too ignore and distort evidence, but in their case there is no self-regulatory mechanism. The more sensational the scare — for example, "Frankenfoods!" — the greater the publicity.

Michaels describes how defenders of the tobacco industry exploited the uncertainties of science by promoting doubt: they demanded that tobacco should be positively proved to cause harm. Green lobbyists make the equally unjustifiable demand that genetically modified crops should be positively proved to be safe. Reports on transgenic crops by the World Health Organization, by every national academy of science and their worldwide cultivation for more than ten years provide no evidence of harm to human health. Yet environmental organizations ignore this. They constantly recycle discredited findings by Árpád Pusztai in 1999 and Irina Ermakova in 2007 that transgenic crops cause harm to rats, and continue to make long-disproved claims that transgenic maize is destroying monarch butterflies.

The motives of these green activists are ideological not financial, based on fears that science has gone too far and we must go 'back to nature', and that transgenic crops benefit only big corporations. More than ten million small-scale farmers have benefited from transgenic crops, mainly farmers of genetically modified cotton, who saw improvements in health and income from the reduced need to spray pesticides. Confident of the virtue of their cause, Greenpeace, Friends of the Earth and the Soil Association feel no need for peer-reviewed papers and show less regard for evidence than the large corporations they denounce.

Doubt Is Their Product underlines the need for independent regulation. Michaels records how George W. Bush's administration applied political pressure to undermine the independence of regulators in the United States, and demonstrates the perils of trusting corporations themselves to protect the public interest. In Europe, the threat to science comes from another quarter: the excessive influence of over-zealous green campaigners, who have virtually driven agricultural biotechnology into exile.

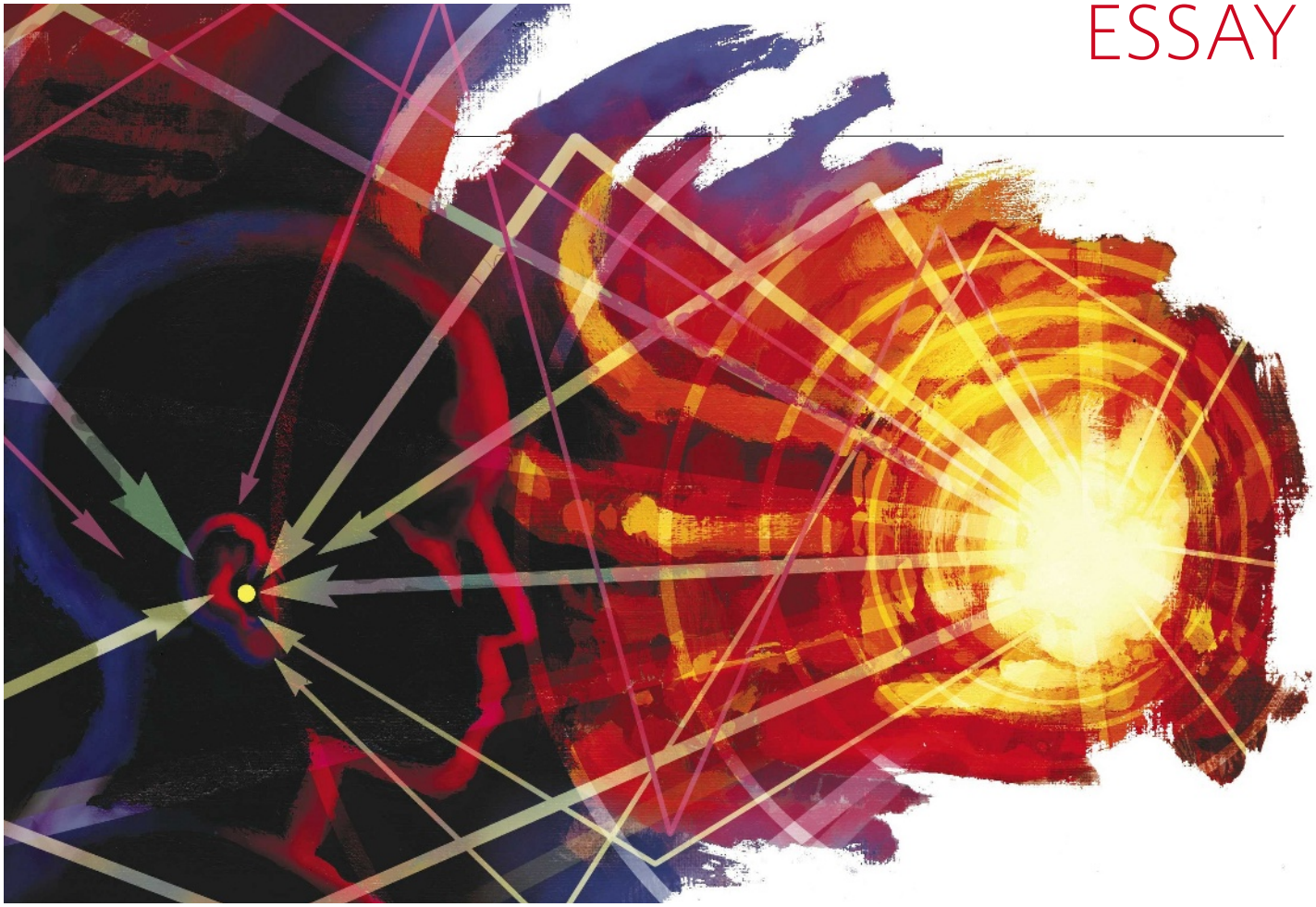
Despite scientific evidence, environmentalists oppose transgenic crops such as modified rice.

O. TORRES/AFP/GETTY IMAGES

The moral is that we must all recognize our tendency to judge evidence with a bias towards our own interests and beliefs. This makes it especially incumbent on those with corporate connections to ensure that respect for evidence predominates in industry-financed research. Equally, those of us who care passionately about the environment must be on our guard to ensure that green causes do not ignore or distort the scientific evidence on which their success depends. ■

Dick Taverne is chairman of Sense About Science, a member of the UK House of Lords Committee on Science and Technology and author of *The March of Unreason — Science, Democracy and the New Fundamentalism*.

ESSAY



D. PARKINS

Raising the roof

Michael Barron explores how physics, psychology and fashion have influenced concert hall acoustics.

A concert hall is a hall of mirrors — acoustic mirrors. When sound hits a hard surface, it is reflected just as light is. Masonry, plaster, timber and glass all reflect sound with very little energy loss. Sound in rooms is predominantly a question of thousands of reflections. In a typical concert hall, only the audience and the seating absorb sound as a black surface absorbs light.

The science of concert hall acoustics is founded on our understanding of the physical behaviour of sound and how our ears interpret it. But concert halls are, of course, more than just scientifically designed spaces. The raised status of acousticians within design teams since the mid-1980s has resulted in less risk-taking and more conservative designs.

Much remains to be discovered about how our ears and brains process sound reflections. Understanding this has been complicated, for instance, by our remarkable ability to work out where a sound is coming from. This ability, called localization, works even when the sound arriving directly from the source represents only a small proportion of the total sound we receive, perhaps only 5% at the back of a concert hall. Usually we are listening to speech or music, which have short elements such as syllables or notes that vary with time. Our brains use this time-varying information to extract where the initial sound comes from. Continuous sound from a fan, for instance, is much more difficult to localize.

The downside of this localization is that, in effect, our hearing suppresses awareness of sound reflections. We notice early sound reflections but are often not conscious of their effects — such as making sound seem clearer than it would be otherwise. In a cathedral-type space, sound persists for several seconds, reflecting back and forth between walls, pillars and ceiling. We can hear this process in a large space because it happens slowly. The sound we hear is known as reverberation, which occurs in virtually all rooms. In smaller spaces, our brains cannot unscramble it because it happens too quickly.

Between 1850 and 1900, concert halls and theatres were built following earlier precedents, with acoustics handled on a trial-and-error basis. Large theatres that provide intelligible speech are a testament to designers from this period, but sadly they left no written record. The importance of reverberation for musical performances was probably appreciated by some, who realized that concert halls need high ceilings; again we have only the buildings as evidence.

The science of room acoustics began around 1900. As in architecture, radical new approaches followed periods of consolidation.

Taming reverberation

In 1895, physicist Wallace Clement Sabine of Harvard University, Massachusetts, was asked to investigate the disastrous acoustics of a new lecture theatre in the Fogg Art Museum at the university. Sabine must have suspected that reverberation was the key. Rather than just solve the problem of one room, he chose to make a fundamental study. He discovered that the duration of sound persistence, or reverberation time, is proportional to the volume of the space divided by the amount of sound-absorbing material. This relationship, published in 1900 and now known as the Sabine equation, is still the basis of room acoustic design. The

lecture theatre itself was eventually demolished, although the museum still stands.

For speech to be intelligible, a short reverberation time is needed, typically between 0.8 and 1.0 seconds. For a symphony concert hall, a long reverberation time of 2 seconds is the optimum. Musical clarity is equivalent to speech intelligibility, but for music we prefer to hear more reverberation from the room, which adds a bloom to the experience. Sabine's equation leads to the first rule of concert hall design: a large room volume is needed, with ceilings that are higher than would be selected on purely visual grounds.

Sabine was acoustic consultant for the Boston Symphony Hall, Massachusetts, which opened in 1900 and is world-renowned for its good acoustics. That the reverberation time in Boston is shorter than predicted suggests an equation is not the whole answer. The sound absorption by objects and building materials also had to be determined, and assessing the absorption by seating and audiences was (and still is) particularly difficult.

During the first half of the twentieth century, seat area gradually increased for reasons of comfort, but it took some time for acousticians to realize that the absorption of each person also increased with seat area. This affected many auditoria of the time, such as the 1940 Kleinhans Music Hall in Buffalo, New York, and London's Royal Festival Hall, completed in 1951, which has a reverberation time shorter than the preferred value because audience absorption was underestimated. A solution would be to raise the roof, but during the hall's refurbishment, completed last year, this was deemed too expensive and damaging to architectural heritage. Some sound-absorbing material was removed and small increases in volume were made, both of which increased the reverberation time. But the change remains too small. Experienced listeners comment that the sound is still 'drier' than in the best halls.

Sound psychology

By the 1950s, it was apparent that there was more to concert hall acoustics than just reverberation time. Research moved into the realm of experimental psychology and away from the area of physics, where acoustics had traditionally been studied. Scientists explored the subjective significance of early sound reflections. They simulated concert hall conditions in anechoic chambers (where the walls, floor and ceiling are covered with sound-absorbing material) to ensure that the listener received only sound direct from individual loudspeakers. For reflections, modified tape-recorders produced the delay, and the direction of simulated reflections from loudspeakers could also be varied.

This research revealed that musical clarity is related to the ratio of sounds that arrive early to those that arrive late, and that listeners prefer to receive a significant proportion of sound from the side. Key studies at the universities of Göttingen and Berlin in Germany during the 1970s used a 'dummy head', a solid artificial head bearing anatomically accurate ear lobes with microphones mounted in each ear canal. A recording made with a dummy head allows for accurate reproduction of sound as heard at each location. By listening to recordings of music made at different seats in actual concert halls, the acoustics could then be rated.

The current consensus is that there are five subjective dimensions for concert hall listening. These are clarity, or ability to hear musical detail; reverberance, or being able to hear reverberations; acoustic intimacy, which describes how involved we seem to be in the performance; envelopment, the extent to which we feel surrounded by sound, and loudness. A range of measures has been developed that allows concert hall designs to be tested before building starts, using computer programs or acoustic scale models. As a result, acoustic disasters are now much less likely.

"Acoustic disasters are now much less likely."

Music box

That's the science, what of the reality? Concert halls represent a civic statement by the client, and are designed by architects advised by acoustic consultants. Many pragmatic aspects come into play, not least architectural fashions. One of the most famous buildings in the world, the Sydney Opera House, was designed back-to-front with the concert hall and opera theatre having to fit within those famous shells.

Sabine's formula says nothing about the appropriate shape for a concert hall. Between 1960 and 1990 there was virtually a free-for-all in auditoria shapes. One success from this period is the striking design of the 1972 Christchurch Town Hall auditorium in New Zealand. Its acoustician, Harold Marshall, used his theory about the importance of sound reflections from the side to develop its acoustic design. It has an oval plan with large tilted reflectors above and behind gallery seating.

In the mid-1980s, clients realized that a hall with poor acoustics was likely to have financial problems and to reflect badly on the city or local area. Once a concert hall is built, acoustic faults are often difficult to rectify, as in London's Royal Festival Hall. This realization shifted the balance within design teams away from the architect and towards the acoustician. It also introduced a period of conservatism, with just two forms predominating for symphony concert spaces: the rectangular shoe-box hall and the terraced hall. Two important parallel-sided auditoria from this period are the 1989 Eugene McDermott Concert Hall in Dallas, Texas, and the 1991 Birmingham Symphony Hall in the United Kingdom. The rectangular form was resurrected from more than 100 years ago. Three such halls from the nineteenth century are regularly mentioned as having the best acoustics in the world: the Vienna Musikvereinssaal, Amsterdam's Concertgebouw and Sabine's own Boston Symphony Hall.

The terraced hall subdivides the audience into different seating levels with useful sound-reflecting surfaces in between. These are often referred to as vineyard terraces, as on a hillside. The first example of this form, the Berlin Philharmonie of 1963, remains the most impressive. Two more recent examples are the 1982 St David's Hall in Cardiff, Wales, and the 2003 Walt Disney Concert Hall in Los Angeles.

This then is the current state of the art and practice of concert hall design. The era of intense research is probably over, although optimizing acoustic conditions for orchestral players remains to be resolved. The period of ultra-conservatism in design is also waning and tentative steps are being made towards new forms. The new Philharmonie concert hall for Paris, due for completion in 2012, promises to be an exciting development. The aim in Paris, as in Christchurch, is to provide musical clarity with a rich sense of reverberation by suspending seating areas in a larger surrounding volume.

Some acoustic designers have offered variable acoustics in their halls, which has met with mixed enthusiasm. Tools include variable volume, variable additional absorption and electronic enhancement. These need to be used in the correct way and should be obvious to the ear. Perhaps such features will allow acoustics to be tailored to the music of different eras. The technology is there, less so the acceptance by musicians and hall managers.

Most large cities in the developed world now have concert halls, and they are unlikely to be demolished because of disappointing acoustics. Maybe we will have to look to other countries such as India and China to find the wild boy or girl of acoustic design for the next generation of concert halls. ■

Michael Barron is reader in the Department of Architecture and Civil Engineering, University of Bath, BA2 7AY, UK, and partner in Fleming & Barron, acoustic consultants. He is author of *Auditorium Acoustics and Architectural Design*.

See other essays in the Science & Music series at www.nature.com/nature/focus/scienceandmusic.

NEWS & VIEWS

BIOCHEMISTRY

Molecular cloaking devices

Thomas Kodadek

Protease enzymes cut other proteins into pieces, but some can be blocked by inhibitors. One such inhibitor binds to the substrate rather than the enzyme, suggesting a new tactic for drug discovery.

Most drug developers try to identify small molecules that bind to a target enzyme with high affinity and specificity. Binding usually occurs at the enzyme's active site; by nestling into this molecular pocket, the drug blocks binding of the enzyme's natural substrate. Alternatively, the drug binds to the enzyme at a location that is distinct from the active site, thus locking the enzyme into a conformation that either strongly inhibits or activates substrate processing in the active site itself. These two models of binding explain the behaviour of most drugs.

There is, however, a third option: the drug could bind to the enzyme's substrate. This is almost unheard of, because enzyme substrates are thought to lack drug-friendly binding pockets, and thus represent slippery targets for small molecules. But on page 925 of this issue, Kukar *et al.*[1] report the remarkable finding that certain small molecules — promising compounds for the development of drugs to treat Alzheimer's disease — operate by just such a substrate-targeted mechanism.

The enzymatic reaction studied by Kukar *et al.*[1] is the cleavage of a protein substrate into pieces by a protease enzyme. Proteases are increasingly attracting attention as targets for drug development. This is because their substrate proteins often have little or no biological activity before they are processed by the enzyme, and so the cleavage reaction represents a convenient choke point at which to block their latent function. Protease inhibitors have already proved their worth as drugs in spectacular fashion — the development of potent inhibitors of HIV protease that block the conversion of precursor molecules to viral proteins has transformed the treatment of AIDS[2].

Kukar and colleagues[1] focus on the cleavage of amyloid precursor protein (APP) by the enzyme γ-secretase. APP is produced by nerve cells, and is cut at various positions along its amino-acid chain by γ-secretase. Some of the cleavage products can assume a misfolded conformation that causes them to aggregate, forming oligomers and long fibril-like structures that might be responsible for the devastating effects on brain function in patients with Alzheimer's disease. Certain small molecules, including some simple anti-inflammatory drugs, delay the onset of the symptoms of Alzheimer's[3] and inhibit the formation of the dangerous cleavage products[4]. But the mechanism of inhibition is unknown, hindering the development of clinically useful drugs.

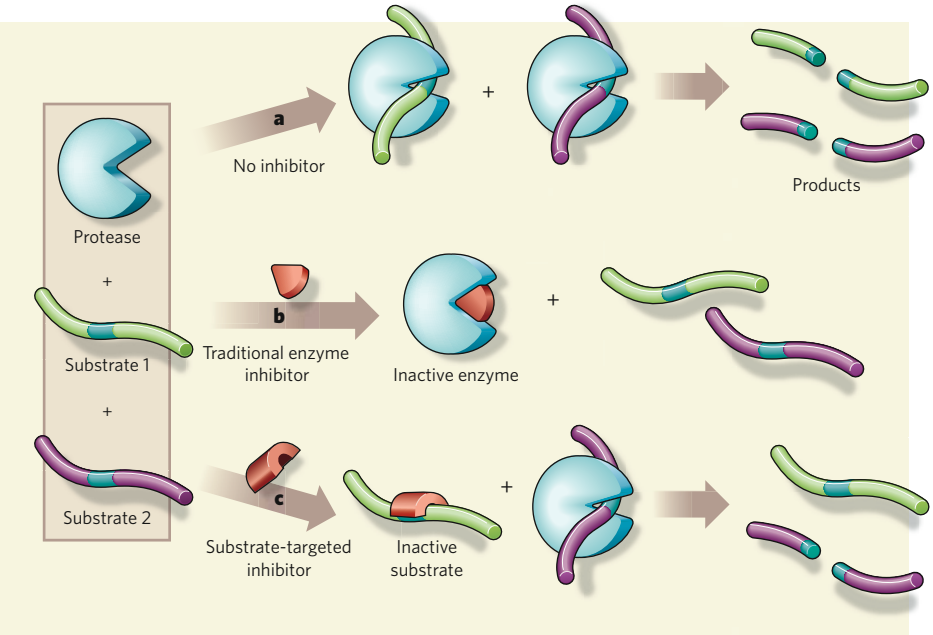


Figure 1 | Targets for enzyme inhibition. Protease enzymes cleave other proteins into fragments. The protease shown has two substrates that share a common cleavage site (blue area) but have different flanking regions (green or pink). **a**, In the absence of an inhibitor, both substrates are cleaved by the protease. **b**, An inhibitor that targets the protease prevents the cleavage of both substrates. **c**, Inhibitors that target substrates can be designed to recognize the unique arrangement of amino-acid residues that flank the cleavage site of a particular substrate. This allows them to specifically protect that substrate from cleavage, as shown for substrate 1. Kukar *et al.*[1] have identified substrate-targeting inhibitors that prevent the cleavage of amyloid precursor protein (APP) by the γ-secretase enzyme, but do not protect similar proteins that have the same cleavage site as APP.

To address this issue, Kukar *et al.*[1] used a method by which an inhibitor drug could be attached covalently to the protein (or proteins) with which it associates. The authors then identified the resulting protein–inhibitor products. To their surprise, they found no evidence for association of the inhibitor with the γ-secretase, which had always been presumed to be the drug's target. Instead, the authors' results, from several different experiments and in various milieux — using either purified proteins or cruder cell extracts — consistently indicated that the binding partner of the drug was the APP substrate. The authors narrowed down the location of the drug's binding site to a short peptide sequence in APP that includes a γ-secretase cleavage site, and that is also known to be involved in the aggregation of the cleaved products.

This is an astonishing result. There are precious few examples of substrate-targeted enzyme inhibitors in the literature, all of which are peptides rather than drug-like small molecules[5,6]. If Kukar and colleagues' findings[1] prove to be relevant to other proteases and their substrates, it would suggest that a completely new set of drug targets exists for the treatment of a variety of diseases. At the very least, it argues that drug-screening efforts should not just seek compounds that bind to

proteases themselves, but also those that bind to their substrates.

Substrate-targeted drugs are an attractive prospect, because they are expected to be much more selective for their proteins than protease-targeted inhibitors could be. Proteases invariably have many different substrates, so that even the most specific protease inhibitors will affect dozens, if not hundreds, of different reactions. The potential for unwanted side effects *in vivo* is therefore significant. But a substrate-targeted inhibitor could, in theory, specifically block the cleavage of the targeted substrate by binding not only to the substrate's cleavage site, but also to surfaces flanking the site[7]. Such surfaces provide unique 'signatures' to distinguish the protein from the many other competing substrates (Fig. 1). The cleavage site would thus effectively be cloaked from the protease by the inhibitor. This scenario might well apply in Kukar and colleagues' study[1], as some of the molecules under investigation inhibited the cleavage of APP by γ-secretase, but not that of other substrates for the enzyme.

The elucidation of the structural basis of the drug–APP interactions will be eagerly awaited. One would assume that the cleavage-site region in APP assumes some three-dimensional structure that is recognized by the drug. But more surprises may be in store, because the cleavage sites in protease substrates are generally found in loops that extend away from the body of the protein, rather than forming a pocket. In the longer term, any information that aids the rational design of substrate-targeted inhibitors of γ-secretase will be invaluable for investigating these compounds as possible drugs in the fight against Alzheimer's disease. ■

Thomas Kodadek is in the Departments of Internal Medicine and Molecular Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9185, USA.
e-mail: thomas.kodadek@utsouthwestern.edu

1. Kukar, T. L. *et al. Nature* **453,** 925–929 (2008).
2. Flexner, C. *N. Engl. J. Med.* **338,** 1281–1292 (1998).
3. Rich, J. B. *et al. Neurology* **45,** 51–55 (1995).
4. Beher, D. *et al. J. Biol. Chem.* **279,** 43419–43426 (2004).
5. Dong, D. L., Liu, R., Sherlock, R., Wigler, M. H. & Nestler, H. P. *Chem. Biol.* **6,** 133–141 (1999).
6. Zhang, Z., Ly, T. & Kodadek, T. *Chem. Biol.* **8,** 391–397 (2001).
7. Kodadek, T. *Biopolym. Peptide Sci.* **66,** 134–140 (2002).

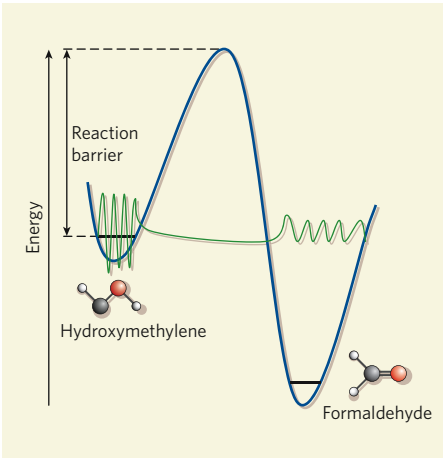


Figure 1 | Quantum tunnelling of hydroxymethylene. Schreiner *et al.*[1] have made and characterized hydroxymethylene (HCOH), an unstable organic molecule, for the first time. They observe that HCOH decomposes to form formaldehyde. The graph (blue line) shows the calculated total energy of the atoms in HCOH as the reaction proceeds. The energy reaches a peak mid-way through the process; the difference in energy between the ground state of HCOH and the peak is known as the reaction barrier. At the low temperatures used in Schreiner and colleagues' study, HCOH molecules have insufficient thermal energy to overcome the reaction barrier. Instead, the molecules 'tunnel' through. This behaviour results from the wave–particle duality of the molecule: the waveform of HCOH (green) extends into the energy well of formaldehyde, allowing molecules to bypass the barrier.

PHYSICAL CHEMISTRY

Cool it, baby

Markku Räsänen

A long-sought but short-lived molecule has been made and characterized for the first time. This compound decays at low temperatures using an unusual trick — a mechanism known as quantum tunnelling.

Chemists are now so skilled at making organic compounds that you might think that all the smallest molecules (containing fewer than 10 atoms) have been made and characterized. This might be true for stable molecules, but some compounds that have weak bonding or high reactivity are short-lived, and have remained elusive. On page 906 of this issue, Schreiner *et al.*[1] report the formation and detection of one of these bashful beasts — hydroxymethylene (HCOH). This molecule is one of the simplest members of a family of compounds known as singlet carbenes, which have recently gained prominence as molecular components of catalysts for organic synthesis. Schreiner and colleagues' findings might cast fresh light on the chemistry of these fascinating compounds, and could even lead to the discovery of new reactions.

All previous attempts to conclusively identify HCOH have failed. This is puzzling, as theoretical studies[2] suggest that it should be a relatively stable molecule. Until Schreiner and colleagues' report[1], the best evidence for the molecule's existence came from studies[3] of a light-induced reaction of isotopically labelled formaldehyde (a stable isomer of HCOH). The incorporation of isotopes from formaldehyde into the products suggested that HCOH is an intermediate in these reactions, but this evidence is indirect.

Despite the lack of convincing evidence, chemists have always been confident that HCOH exists. Perhaps most intriguingly, some have speculated that it might form in interstellar clouds, where it could participate in the formation of larger organic molecules[4]. What is known for sure is that certain larger carbenes, in which the hydrogen atoms of HCOH are replaced with chemical groups, are perfectly stable. Indeed, such carbenes are the basis of a rich branch of chemistry and have many applications in organic and organometallic syntheses[5]. They also have a central role in combustion and atmospheric chemistry.

So how does one go about identifying an ephemeral molecule? Several methods have been developed. Spectroscopic measurements can now be taken with femtosecond resolution (1 femtosecond = 10^{-15} seconds), allowing the detection of extremely short-lived molecules. Alternatively, the lifetime of such molecules can be increased by lowering the temperature of the experimental sample, so providing time for a more leisurely investigation using slower spectroscopic techniques. For gas-phase experiments, cooling is achieved by rapidly expanding the volume of the gas being studied. Molecules can also be trapped in an inert medium at low temperatures, allowing them to be studied in the solid phase. The second method is called matrix isolation[6], and was pioneered by the chemist George Pimentel, whose slogan was "cool it, baby".

Schreiner *et al.*[1] used matrix isolation in their identification of HCOH. They prepared the target compound by heating glyoxylic acid ($HCOCO_2H$) in a high vacuum; under these conditions, carbon dioxide is eliminated from the starting material, yielding HCOH directly. The authors isolated the product in solid argon at a temperature of about 10 kelvin, and characterized it using a combination of spectroscopic techniques, supporting their interpretation of the results with predictions from quantum-chemical calculations. In this way, they provided the first unambiguous proof for the formation of HCOH.

The reason most short-lived molecules have such a fleeting existence is because they easily react to form something else. To put it another way, the amount of energy required by the molecules for an onward reaction — the reaction barrier — is low. The lifetime of a short-lived molecule is lengthened by cooling because it

reduces the amount of thermal energy available to the molecule to overcome its reaction barrier. But some remarkable reactions occur without the need for thermal excitation, through a process known as quantum tunnelling. Schreiner *et al.*[1] report that just such a process occurs for HCOH — when the molecule adopts a certain conformation, it decays to produce formaldehyde in a matter of minutes, even though the calculated barrier for the reaction is too high to be overcome at the temperatures used in the experiments. Instead, the hydrogen atom in the hydroxyl group (OH) of HCOH 'tunnels' through the barrier, in a process that relies on the wave–particle duality of the atom (Fig. 1). This unexpected reaction might explain why previous attempts to detect HCOH failed — the product decayed before it was found and analysed.

The authors[1] investigated the quantum tunnelling process by replacing the hydrogen atom of the hydroxyl group with deuterium (a heavy isotope of hydrogen). They found that the deuterated form of HCOH is essentially stable under their experimental conditions. This stabilizing effect occurs because the wave associated with the deuterium atom decays before reaching the exit of the barrier (unlike the wave for the hydrogen atom), and provides direct evidence of quantum-tunnelling. Turning again to theoretical predictions to validate their results, the authors modelled the tunnelling process computationally. The half-life for HCOH at close to zero kelvin was predicted to be 122 minutes, whereas that of deuterated HCOH was predicted to be more than 1,200 years, in good agreement with their experimental observations.

The data provided by this study[1] will be invaluable for those searching for HCOH under different conditions, for example in the gas phase, where it is expected to participate in many reactions. For such studies, it will be necessary to find out if the tunnelling process is similar in gases and solids — that is, does the matrix surrounding the HCOH molecules in Schreiner and colleagues' experiments have some role in the mechanism? The findings also provide a well-defined system of small molecules that will be a perfect model for studying the mechanisms of tunnelling reactions in solids. Far from being yesterday's news, small molecules still have much to teach us. ■

Markku Räsänen is in the Department of Chemistry, PO Box 55, FIN-00014, University of Helsinki, Finland.
e-mail: markku.rasanen@helsinki.fi

1. Schreiner, P. R. *et al. Nature* **453,** 906–909 (2008).
2. Pau, C. F. & Hehre, W. J. *J. Phys. Chem.* **86,** 1252–1253 (1982).
3. Sodeau, J. R. & Lee, E. K. C. *Chem. Phys. Lett.* **57,** 71–74 (1978).
4. Baly, E. C. C., Heilbron, I. M. & Barker, W. F. C. *J. Chem. Soc. Trans.* **119,** 1025–1035 (1921).
5. Bourissou, D., Guerret, O., Gabbai, F. P. & Bertrand, G. *Chem. Rev.* **100,** 39–91 (2000).
6. Whittle, E., Dows, D. A. & Pimentel, G. C. *J. Chem. Phys.* **22,** 1943 (1954).

ANIMAL BEHAVIOUR

Guardian caterpillars

Parasites are in the business of hijacking their hosts for their own purposes. A dramatic example is described by Amir Grosman and colleagues who studied behavioural changes induced in the Brazilian geometrid moth *Thyrinteina leucocerae* by a braconid parasitic wasp of the genus *Glyptapanteles* (A. H. Grosman *et al. PLoS ONE* **3,** e2276; 2008).

The wasp lays up to 80 eggs in a moth caterpillar, where they hatch, grow and then exit to form pupae on nearby leaves three weeks later. At this point the caterpillar ceases its previous activity, and instead stands guard over the wasp pupae (pictured). While 'guarding' the pupae, caterpillars make violent swings of their heads towards any passing insect or other potential threat.

Grosman *et al.* show that this behaviour is a direct result of parasite infestation and profits the wasp. Unlike the parasitized caterpillars, their unparasitized siblings failed to take up sentry duty when placed by wasp pupae, continuing to feed as normal. Also, the head butts of parasitized caterpillars proved effective protection for the pupae, repelling nearly half of the attacks by predatory stinkbugs (*Supputius cincticeps*) in the laboratory.

Such behaviour has no fitness advantage for the caterpillars, who die soon after the adult wasps emerge from the pupae. But the wasps profit hugely, the presence of a guarding caterpillar almost halving the mortality of pupae in field experiments.


J. LINO-NETO

Although the mechanism by which these parasites influence their hosts is not clear, live wasp larvae were found in the bodies of caterpillars even after their broodmates had departed. Such 'brain worms' have also been seen with trematode and liver fluke parasites that modify the behaviour of their ant hosts.

Christopher Surridge

HUNTINGTON'S DISEASE

Genetics lends a hand

Stéphane Palfi and Bechir Jarraya

A monkey model of Huntington's disease created by gene transfer is only a work in progress. But as a technological feat it offers great promise for fathoming this devastating condition.

Huntington's disease is a heritable disorder that affects more than 1 in 10,000 people. Its associated neurological symptoms are severe, and there is no therapy to halt or slow its progress. To understand its pathology, and with the ultimate hope of finding a treatment, researchers have generated several experimental models of this disease[1], in organisms such as flies and rodents. Although these models have led to tremendous progress in understanding the pathogenic mechanisms of Huntington's disease, none of them can replicate all the behavioural changes seen in the human disease or the changes that occur at the tissue level. The need for a primate model of the disease is thus clear.

On page 921 of this issue, Yang *et al.*[2] describe their attempt to generate a transgenic monkey model of Huntington's*. Although only a 'proof-of-principle', their achievement is a step forward, and will undoubtedly be welcomed by those involved in developing a cure for this distressing condition.

Initial symptoms of Huntington's disease, which generally appear between the ages of 30 and 40, include changes in personality, a progressive cognitive decline and a spectrum of motor disturbances ranging from abrupt to slow and sustained involuntary movements. At a cellular level, features of the human disease include severe neuronal loss in the striatum, a deep brain structure that regulates movement and the cognitive and emotional aspects of behaviour. At the molecular level, the disease is caused by mutation in the *HTT* gene, which encodes the huntingtin protein. This protein normally contains a chain of 6–35 glutamine amino-acid residues. But when *HTT* is mutated, the number of repeats of the CAG trinucleotide, which encodes glutamine, is greatly expanded, leading to extension of the glutamine chain in huntingtin by anything from 36 to 180 repeats[3].

The technique of transgenesis is used to generate organisms carrying a specific mutation, and involves the introduction of a foreign gene into an organism. The first transgenic primate,

*This News & Views article and the paper concerned[2] were published online on 18 May 2008.

a rhesus monkey[4], was generated in 2001. This animal carried a marker gene encoding green fluorescent protein (GFP). Yang *et al.*[2] faced a bigger challenge in trying to introduce a disease gene into rhesus macaques.

The authors inserted a virus vector carrying part of the mutated human *HTT* gene, with 84 CAG repeats, into unfertilized monkey egg cells. The virus, which became integrated into the egg's genome, also contained a gene for GFP, the fluorescent properties of which served as a marker for the success of the transgenic procedure. Of the six resulting pregnancies, five were born at full term. Of these, three newborns carried between two and four copies of mutated *HTT*. They all developed severe symptoms of Huntington's disease and died within the first month. Of the two surviving monkeys, each of which carries a single copy of the mutated gene, one has only 29 CAG repeats and so shows no disease symptoms, and the other has 83 CAG repeats and has developed mild symptoms, which began within 1 week of birth. The authors suggest that the discrepancy in the number of CAG repeats is due to their instability during integration into the viral vector. These two macaques are still under detailed investigation for disease progression as they get older.

Previous studies[5,6] of patients with Huntington's disease have shown that the expanded stretch of glutamine residues forms aggregates in the nucleus and cytoplasm of affected neurons. Yang *et al.* found that only monkeys expressing mutated *HTT* show the anatomical changes (aggregates in neurons in both the striatum and cortex) and movement abnormalities characteristic of the disease. But the authors' transgenic monkeys show only modest signs of neurodegeneration at the tissue level, suggesting that the neuronal degeneration they observed is at a very early stage.

Researchers working on other experimental models of Huntington's disease are divided over both the exact effect of the *HTT* mutation and the temporal relationship between protein aggregation and neuronal death. Current hypotheses predict that abnormalities may occur in neurotransmission, the regulation of gene transcription, the function of mitochondria (the cell's powerhouses), and transport within axons (neuronal extensions)[7,8]. So far, the features of Yang and colleagues' transgenic primates do not favour any one hypothesis. But because of the fundamental physiological and genetic similarities between monkeys and humans (including lifespan, cellular metabolism, and endocrine and reproductive function), a primate model is more likely than other animal models to provide insights into the mechanisms of cell death associated with Huntington's disease.

In contrast to rodents, non-human primates show neuroanatomical and behavioural characteristics, both motor and cognitive, that closely resemble those of humans. For instance, primate models of Parkinson's and Huntington's diseases[9] that were generated chemically have improved our understanding of the behavioural characteristics of these diseases. But the role of genetic factors in the development of neurodegenerative diseases cannot be determined in such models. Thus, transgenic approaches (for example, local transfer of mutated *HTT* into the monkey striatum[10] or gene introduction into oocytes, as Yang *et al.* have done) seem to be the way forward in addressing the remaining crucial questions.

The primate model generated by Yang *et al.*[2] takes us another step on the long road towards developing a treatment for Huntington's disease. The history of translational (bench-to-bedside) research in neuroscience has shown a strong correlation between the availability of relevant primate models for a given disease and the development of new treatments[11]. But we must retain a healthy caution, as there is a need to determine how closely the characteristics of these authors' transgenic monkeys[2] match the spectrum of symptoms of Huntington's disease.

Yang and colleagues' approach to generating transgenic monkeys should also facilitate the development of a more elaborate model for the common form of Huntington's disease, which is characterized by a delayed clinical onset of motor and cognitive dysfunction. What's more, it opens the way to generating non-human primate models for other neurodegenerative diseases that are caused by single-gene mutations, such as familial forms of Parkinson's disease, Alzheimer's disease and amyotrophic lateral sclerosis. ■

Stéphane Palfi and Bechir Jarraya are in the CEA/DSV/I²BM–MIRCen, Fontenay-aux-Roses, and at the Assistance Publique-Hôpitaux de Paris, l'Hôpital Henri Mondor, Service de Neurochirurgie, Université Paris 12, Faculté de Médecine, Créteil F-94010, France.
e-mail: stephane.palfi@hmn.aphp.fr

1. Sipione, S. & Cattaneo, E. *Mol. Neurobiol.* **23,** 21–51 (2001).
2. Yang, S.-H. *et al. Nature* **453,** 921–924 (2008).
3. MacDonald, M. E. *et al. Cell* **72,** 971–983 (1993).
4. Chan, A. W. S., Chong, K. Y., Martinovich, C., Simerly, C. & Schatten, G. *Science* **291,** 309–312 (2001).
5. DiFiglia, M. *et al. Science* **277,** 1990–1993 (1997).
6. Gutekunst, C.-A. *et al. J. Neurosci.* **19,** 2522–2534 (1999).
7. Romero, E. *et al. Neuron* **57,** 27–40 (2008).
8. Li, S. & Li, X. J. *Mol. Neurodegen.* **1,** 19 (2006).
9. Brouillet, E. *et al. Proc. Natl Acad. Sci. USA* **92,** 7105–7109 (1995).
10. Palfi, S. *et al. Mol. Ther.* **15,** 1444–1451 (2007).
11. Capitanio, J. P. & Emborg, M. P. *Lancet* **371,** 1126–1135 (2008).

NUCLEAR PHYSICS

A neutrino's wobble?

Philip M. Walker

Periodic oscillations have been observed in what should be straightforward exponential decay curves of two radioactive isotopes. An entirely mysterious phenomenon, its proposed cause seems equally exotic.

It is a well-established fact that the rate at which a collection of radioactive atoms decays itself decays exponentially over time. It's easy to see why: the number of decays is directly proportional to the number of radioactive atoms remaining in the sample; so the fewer active atoms there are left, the fewer decays will occur. That makes observations from Litvinov *et al.*[1], in a paper in *Physics Letters B*, all the more surprising. Investigating β-decays of highly positive ions of heavy elements, these authors see a decay rate whose decline is not purely exponential, but which seems to be modulated up and down over time. They propose that the modulation arises from the 'mixing' of different types of neutrino — chargeless and almost massless particles produced in β-decays. That explanation would itself raise a host of further questions.

The observations were made at the Experimental Storage Ring (ESR) of the GSI nuclear physics laboratory in Darmstadt, Germany. The 108-metre-circumference ESR has been the scene of many experiments with highly charged radioactive ions since the early 1990s. Most of these have been precise measurements of ion masses made possible by fastidious determinations of the time ions take to circumnavigate the ring — half a microsecond, give or take. Individual ion decays have also been monitored, and their half-lives deduced. This has yielded a crop of new decay processes: an example is bound-state β-decay[2], in which the β-particle (electron) produced by the ion in its decay is captured in one of its many vacant electron orbitals.

Litvinov *et al.*[1] studied the decay of two heavy radioactive nuclides: praesodymium-140 (^{140}Pr) and promethium-142 (^{142}Pm), both from the lanthanide series of the periodic table. The nuclides, with half-lives of 3 minutes 23 seconds and 40.5 seconds respectively, were produced by smashing stable samarium (^{152}Sm) atoms into a beryllium target at more than two-thirds the speed of light. The violence of this impact strips the samarium atoms not only of almost all of their electrons, but also of several protons and neutrons — knocking off a proton and nine neutrons makes ^{142}Pm, and the loss of three protons and nine neutrons creates ^{140}Pr. With the help of magnets, the isotopes can be selected from others produced. Injected into the ESR, the precise revolution frequency for each individual ion can be obtained, so that the

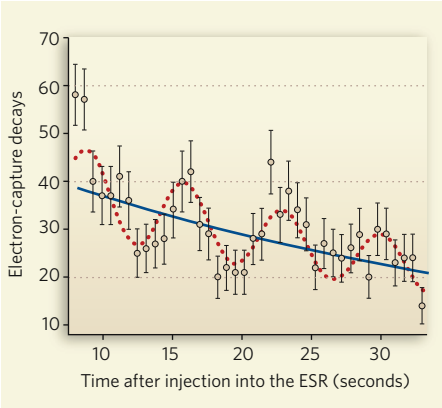


Figure 1 | Oscillating decay. Litvinov and colleagues[1] observe decay-rate oscillations with a period of about 7 seconds, here for decays of promethium-142 through electron capture, but similarly for praesodymium-140 — a phenomenon they attribute to the effect of neutrino oscillations. There are no observations for the first few seconds after formation while the ions are cooled to reduce their velocity spread.

ion's identity (mass) and decay (mass change) can be monitored.

The authors measured the process of electron capture on $^{140}Pr^{58+}$ and $^{142}Pm^{60+}$ — 'hydrogen-like' ions in which all but one of the electrons orbiting the nucleus have been removed. Electron capture is a kind of reverse β-decay in which a nucleus captures an electron from an atomic orbital, resulting in the conversion of a proton to a neutron in the nucleus and the emission of an electron neutrino. The decay products of ^{140}Pr and ^{142}Pm are thus the (stable) isotopes cerium-140 and neodymium-142; decays are signalled through a change in the ion's revolution frequency caused by a small drop in its mass. By observing many such individual decays, the decay rate is obtained, which should, as ever, fall over time as a simple exponential. What Litvinov *et al.*[1] discovered was an unexpected modulating oscillation with a period of about 7 seconds for both ions (Fig. 1).

This behaviour is perhaps all the more curious for popping up in an unusually 'clean' experimental environment. Because the radioactive ions being studied had only a single atomic electron, which went on to be captured during the decay, there are few confounding effects such as Coulomb interactions to take into account. By being confined to the high-vacuum storage ring, the ions were also effectively isolated from outside influences. In addition, only a very small number of ions — three or fewer — were allowed in the storage ring at any one time, limiting their interactions with each other.

So how are we to interpret these results? Litvinov *et al.* go to considerable lengths to rule out spurious causes such as a regular instability in the storage ring or the detection apparatus. They discuss several possible physical origins, such as the quantum-mechanical oscillation between two spin states, one of which is 'sterile', in that it is forbidden by angular-momentum conservation from decaying through electron capture. This possibility seems discounted by the fact that the overall decay rate agrees with the predictions that do not countenance a sterile state.

The authors thus argue by a process of elimination, and in agreement with a recent theoretical suggestion[3], that the modulations are due to the oscillation of neutrinos between two different mass states: that of an electron neutrino, emitted in the original decay; and that of a muon neutrino, which is observed in decays of the electron's 200-fold-heavier sibling, the muon. The generalized phenomenon of neutrino oscillation is now well documented in several contexts in which neutrinos arise — in radiation given off from the Sun, in cosmic rays, and in neutrinos produced in nuclear reactors for energy generation. In the case of heavy-ion experiments, a crucial feature is the minimal recoil energy of the ion as it decays emitting the neutrino, which should make the period of the interference oscillation dependent on the ion's mass — potentially an unequivocal experimental signature.

A back-of-the-envelope calculation shows that Litvinov and colleagues' conjecture[1] that what they see is the expression of neutrino oscillation could be well-founded — for values of the difference in neutrino masses in the middle of the presumed range, an oscillation period of the order of 10 seconds would be expected. If the conclusion did prove to be right, it could represent a sea change in neutrino physics. Neutrinos probably make up a substantial fraction of the mass of the Universe, so it is well worth our while getting to know them better. Yet they are notoriously aloof, generally passing straight through Earth without interacting. Vast underground detectors have to be built to stand a chance of detaining a few of them. One thing the ESR findings would produce would be a way of testing the properties of neutrinos purely through the decay characteristics of heavy ions that are much more amenable to investigation — without the bother of detecting neutrinos at all.

Caution is due, as it is far from simple to get to grips with the underlying cause of the decay-rate oscillations. Experimentally, the next step must be to study another example of electron capture involving a nuclide of substantially different mass that generates a different oscillation period. The authors are planning just such a test — it will be intriguing to see whether this putative new neutrino oscillation proves to be a robust phenomenon, or as elusive as the neutrinos themselves. ■

Philip M. Walker is in the Department of Physics, University of Surrey, Guildford GU2 7XH, UK.
e-mail: p.walker@surrey.ac.uk

1. Litvinov, Yu. A. *et al. Phys. Lett. B* **664,** 162–168 (2008).
2. Jung, M. *et al. Phys. Rev. Lett.* **69,** 2164–2167 (1992).
3. Lipkin, H. J. preprint at http://arxiv.org/abs/0801.1465 (2008).

50 YEARS AGO

Prior to a recent television series on evolution, the producer asked the audience research department of the B. B. C. to equip him with information about the knowledgeability of potential viewers [and] their attitude towards evolution ... Viewers were asked whom they associated with evolution. One in three could give no name; the name given by far the most (by one-third of the total) was that of Darwin. A few mentioned Huxley — but as many named Einstein. Other suggestions ranged from Aristotle to Attenborough, or Marconi to Mortimer Wheeler ... About two-thirds of the sample of viewers said they themselves believed in evolution; just over a tenth disbelieved the theory, the remainder having no firm opinions. Of those who said they believed in evolution, almost half were unable to advance a reason for doing so ... Asked what would ultimately happen to man ...in descending order of frequency the forecasts were: (1) that man would suffer destruction at his own hands; (2) that he would increase his power and conquer space; (3) that there would be development of brain power, or that man would lose certain parts, such as his toes. A few thought man to be "at his peak" and not likely to change further.
From *Nature* 14 June 1958.

100 YEARS AGO

Thomas Alva Edison: Sixty Years of an Inventor's Life. By Francis Arthur Jones.— This biography should do much to disillusion the impressions which are so commonly formed about successful men, that they only have to invent something to make a fortune. It shows clearly that the only road to success is through failure. His career as a telegraph operator was most precarious, and one of his first inventions—a vote-recording machine for election purposes—was refused, really because it was too ingenious and perfect; in fact, it could not be tampered with.
From *Nature* 11 June 1908.

50 & 100 YEARS AGO

CONDENSED-MATTER PHYSICS

Paralysed by disorder

Daniel A. Steck

In a disordered medium, a quantum particle can literally stop itself in its tracks. This localization phenomenon can be observed directly using the coldest known form of matter, caught in a laser trap.

Those familiar with the game of pachinko — in which, in its simplest incarnation, little balls bounce and cascade chaotically through a regular lattice of pin-like obstacles — have some idea of what a current of electrons goes through when it flows through the ordered crystal lattice of a metallic wire. But now imagine that you are playing quantum pachinko. Instead of little balls, wavelets shoot around the atomic obstacle course. The rules of the game are suddenly very difficult: if the waves have the right energy, they complete the course with aplomb, moving around the lattice essentially unimpeded. But you only need to change things slightly, by putting the obstacles in a marginally more disordered arrangement, to stop those quantum waves in their tracks.

This phenomenon is known as Anderson localization, and influences many physical properties of disordered and amorphous materials[1]. Two papers in this issue[2,3] do their bit for a better understanding of it. They detail how the first direct images of this process of 'localization through disorder' were obtained.

But let's return first to our electrons in a wire. An electron in a conducting material is stuck in one of a series of potential wells, each representing an atom in the crystal lattice of the conductor. Quantum mechanics says that in each well, the electron can only occupy certain energy levels. The positioning of these levels depends on the depth and width of the well. Suppose, for simplicity, that there is only a single possible level in each well (Fig. 1a). In a perfect crystal, with a perfectly regular lattice of identical wells, the levels in each well will then line up exactly, and an electron can thus easily tunnel between the wells. The result: electrons are freely mobile, and the material conducts.

Consider now what happens when the lattice is disordered. This might be because of chemical impurities, defects or an innately amorphous structure. The energy levels no longer line up nicely in neighbouring wells, and the electron is blocked from burrowing through the well wall when it finds an energy mismatch that is too large (Fig. 1b). When a determined electron encounters such an obstacle, it can try to change path, and tunnel into a different well. But at some point, it is always frustrated, coming across a barrier that it cannot surmount. Returning to a quantum-wave picture, this acts as an interference effect: an electron tries to take different paths of effectively random lengths, and so on average it destructively interferes with itself, no matter where in the crystal it tries to go. This was the phenomenon that Philip Anderson originally predicted[4] in 1958.

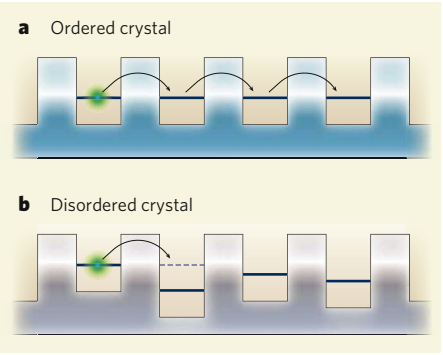


Figure 1 | Well, well, well. a, In an ordered crystal lattice, electron energy levels are the same in each atomic potential well. Electrons can easily tunnel between wells, resulting in conduction. **b**, As the two new papers[2,3] demonstrate, in a disordered lattice, energy mismatches between adjacent wells inhibit electron motion and produce localization.

A perhaps more familiar manifestation of Anderson localization comes when one shines light through a stack of glass slides of ever so slightly varying thickness — the transmission of light decreases exponentially as the slides are piled on[5]. For matter waves such as electrons, a direct observation of Anderson localization has been more elusive: the delicate quantum interferences required are destroyed by influences such as the vibrations in a crystal lattice; and it's difficult to image an electron in a crystal directly anyway.

In the new experiments[2,3], ultracold atoms replace the electrons, and the crystal lattice is an optical lattice — one supplied by the forces due to laser light. This set-up allows the localization of the atoms to be seen directly, by illuminating the atoms with laser light and taking their picture. Both groups studied the localization of atoms in a Bose–Einstein condensate, a vapour of atoms so cold (at a temperature of a few nanokelvin above absolute zero) that the atoms move in perfect lock-step. Such a condensate has two agreeable properties: it has a small extent in space, so it is obvious whether or not the atoms are stuck in place; and it has a small spread in momentum, so the wavelengths of the atoms are large enough to get an overview of the surrounding landscape and so 'sense' its disorder.

Bose–Einstein condensates in disordered potentials had already been studied, but two experimental challenges have impeded localization being observed. First, at the densities required to produce the condensate, collisions between atoms smooth out any disorder. Second, disorder has to be produced on length scales smaller than the typical atomic wavelengths. Both groups solved the first problem by allowing the condensate to expand, thus lowering its density, before looking for localization; but they took different approaches to the second problem. Billy *et al.* (page 891)[2] studied atoms moving in a fine-grained 'laser speckle field', such as is produced when a laser beam is scattered by a finely roughened surface. By contrast, Roati *et al.* (page 895)[3] used a superposition of optical lattices made from lasers of unequal wavelengths. This creates a 'quasicrystal' that is a halfway house between a perfectly regular crystal and an entirely amorphous one — with an intermediate degree of disorder that is still sufficient for localization.

Both groups observed the freezing of atomic motion as they introduced disorder in their respective, carefully controlled ways. They also both observed the incontrovertible signature for Anderson localization — that the probability of finding an atom at a particular point in space diminishes exponentially the farther one moves from the original centre of the condensate. But the two experiments also had interesting twists of their own. Billy *et al.* observed, for example, that a fraction of higher-energy atoms partially ignored the disorder and moved around unimpeded. Roati *et al.* also observed a smoother momentum distribution in the localized state — the signature of localization in momentum space — as well as interference patterns due to atoms localized near several sites.

These impressive experiments are the latest in an exciting new thrust of research: the use of ultracold atoms to study condensed-matter phenomena in precise and controlled experiments[6–8], and thus bridge the gap between theory and experiment in condensed-matter physics (for an overview of current efforts in this direction, see the Q&A feature[9] in last week's issue). And on a more applied note, effects such as Anderson localization are gaining technological importance as a way to confine and control optical waves in, for example, engineered optical media[10]. ■

Daniel A. Steck is in the Oregon Center for Optics and the Department of Physics, University of Oregon, Eugene, Oregon 97403-1274, USA.
e-mail: dsteck@uoregon.edu

1. Lee, P. A. & Ramakrishnan, T. V. *Rev. Mod. Phys.* **57,** 287–337 (1985).
2. Billy, J. *et al. Nature* **453,** 891–894 (2008).
3. Roati, G. *et al. Nature* **453,** 895–898 (2008).
4. Anderson, P. W. *Phys. Rev.* **109,** 1492–1505 (1958).
5. Berry, M. V. & Klein, S. *Eur. J. Phys.* **18,** 222–228 (1997).
6. Raizen, M., Salomon, C. & Niu, Q. *Phys. Today* **50,** 30–34 (1997).
7. Bloch, I., Dalibard, J. & Zwerger, W. *Rev. Mod. Phys.* (in the press); preprint at http://arxiv.org/abs/0704.3011 (2007).
8. Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W. & Bloch, I. *Nature* **415,** 39–44 (2002).
9. Greiner, M. & Fölling, S. *Nature* **453,** 736–738 (2008).
10. Istrate, E. & Sargent, E. H. *Rev. Mod. Phys.* **78,** 455 (2006).

OBITUARY

Willis E. Lamb Jr (1913–2008)

Meticulous physicist and discoverer of the Lamb shift.

Willis Lamb, who died on 15 May aged 94, received his highest recognition in 1955, when he was awarded one half of the Nobel Prize in Physics for his discovery of the subtle quantum-mechanical shift in energy levels of the hydrogen atom that now bears his name. But in truth, he might have been nominated twice more. He produced a theory of the resonant absorption of γ-rays by solid-state atoms — the Mößbauer effect — 19 years before Rudolf Mößbauer himself published the observations that won him a share of the 1961 prize. And in the early 1960s, Lamb developed one of the first, semi-classical theories of how lasers work.

Lamb was born on 12 July 1913 in Los Angeles, and as an undergraduate at the University of California, Berkeley, studied chemistry. But he became increasingly fascinated with physics, and his presaging of the Mößbauer effect was part of a doctoral dissertation on the electromagnetic properties of nuclear systems, completed in 1938 under the guidance of J. Robert Oppenheimer.

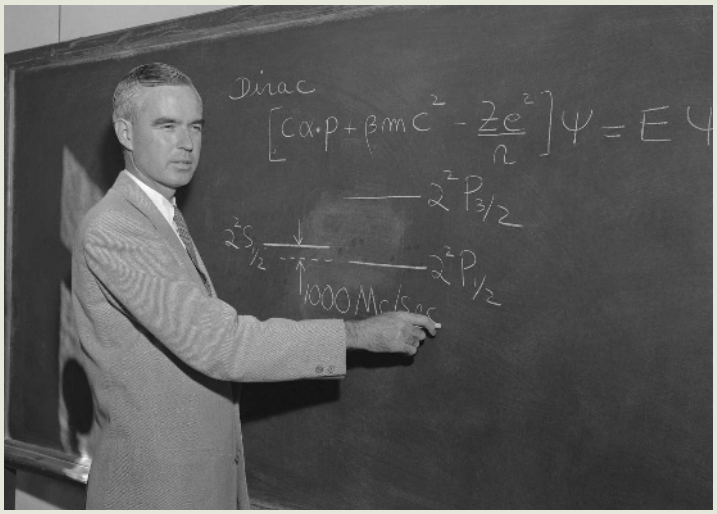
Unlike many of his generation of physicists, Lamb did not follow Oppenheimer into the wartime atom-bomb project. Instead, he concentrated on his specialisms — microwaves and radar — at Columbia University in New York, performing the experiments that culminated in the observation of the Lamb shift (W. E. Lamb Jr and R. C. Retherford *Phys. Rev.* **72,** 241–243; 1947).

This shift is a tiny difference in energy between two atomic orbitals in hydrogen, denoted 2*s* and 2*p*, distinguished only by their angular momenta. Quantum theories of the time predicted that these levels should have identical energies. The discovery that they did not, demanded a fundamental theoretical rethink — one that was initiated almost immediately by Hans Bethe.

The Lamb shift thus became a cornerstone of the modern edifice of quantum electrodynamics (QED). This, the quantum field theory of the electromagnetic interaction, explains the shift as resulting from energy fluctuations in the vacuum that smear out the position of the electron in a hydrogen atom. This process has a greater effect on the Coulomb energy of the electron's binding to the central proton at smaller radii (where the 2*s* state is most likely) than at larger radii (where the 2*p* state dominates). Today, precise measurements of the Lamb shift have tested QED to an accuracy of better than one part in a million.

Lamb's disposition made him entirely suited to the difficult measurements of this tiny shift. A perfectionist, his approach to physics embodied the principle that physical intuition is a necessary, but not sufficient, condition for success. A sound conclusion demanded a solid calculation or experiment.

That perfectionism extended to Lamb's writing: he would spend considerable time revising papers, or having his students revise them, pondering long and hard over subtle refinements such as the insertion of a comma. That honed his students' ability to write and led to unusually readable papers. But his meticulousness was never overbearing, and was always tempered by a wry, self-deprecating sense of humour. He was an outstanding supervisor, inspiring many students and postdocs with sharp insight and sure guidance. I recall trying to convince him to take me on as a graduate student, presumptuously claiming expertise in his field of interest. His response was that he was always interested in expertise, as he had so little of his own. It was the start of a long and fertile collaboration on laser theory.



BETTMANN/CORBIS

In his later career, Lamb's refusal to generalize conclusions purely on intuition prevented him on occasion from publishing important work that didn't meet his own high standards. Perfection for him was a quality to be strived for, and perhaps not to be attained. He once told me that he regretted having been so close to so many great discoveries without realizing it. I replied that I wished I had that problem.

Lamb's contributions to laser physics were a case in point. The first microwave-frequency equivalent of the laser, the maser, had been constructed by Charles Townes and collaborators at Columbia in 1953. But like others working on the extension of the maser to higher, optical frequencies, Lamb was convinced that a closed cavity, such as that in which light was amplified in a maser, would be needed to create a laser. It was left to Gordon Gould to predict that an open cavity could lase, and for Theodore Maiman to implement the idea in 1960.

Lamb's part was nevertheless substantial. His theory of laser action (W. E. Lamb Jr *Phys. Rev.* **134,** A1429–A1450; 1964) predicted the 'Lamb dip' in the output intensity of a gas laser. This spectral feature is a consequence of the Doppler effect, occurring when, owing to thermal motions of the gas, an atomic absorption line becomes smeared out in frequency. In this case, two oppositely directed light waves will interact with atoms of the same velocity if tuned to the centre of the Doppler-broadened line, but with atoms of different velocities when tuned off-resonance. The laser gain thus saturates at a lower intensity on resonance. This Lamb dip in lasers, and its counterpart in absorptive media, enabled the creation of incredibly stable devices that redefined the way we measure time and distance.

The fact that Lamb's 1964 laser theory was semi-classical was again entirely typical: Lamb was fascinated by the extent to which nature was necessarily quantum-mechanical. In particular, he insisted that the original photoelectric effect — the proof of whose quantum nature had garnered Albert Einstein his Nobel in 1921 — could be accounted for using a classical electromagnetic field. He had little patience for those who misused the term 'photon', particularly in situations where a classical electromagnetic wave would suffice, even once saying that he should be given responsibility for issuing licences for its use.

Tall and handsome, at first sight Willis Lamb seemed more Hollywood actor or high-flying businessman than scientist. But his dedication to physics was complete, and did not diminish in a 65-year career that encompassed Bell Labs (as a consultant on laser theory) and Stanford, Harvard, Oxford and Yale universities. He retired from the faculty of the University of Arizona only in 2002, at the age of 89.

Murray Sargent
Murray Sargent is at the Microsoft Corporation, 1 Microsoft Way, Redmond, Washington 98052, USA.
e-mail: murrays@microsoft.com

REVIEWS

What we can do and what we cannot do with fMRI

Nikos K. Logothetis[1]

Functional magnetic resonance imaging (fMRI) is currently the mainstay of neuroimaging in cognitive neuroscience. Advances in scanner technology, image acquisition protocols, experimental design, and analysis methods promise to push forward fMRI from mere cartography to the true study of brain organization. However, fundamental questions concerning the interpretation of fMRI data abound, as the conclusions drawn often ignore the actual limitations of the methodology. Here I give an overview of the current state of fMRI, and draw on neuroimaging and physiological data to present the current understanding of the haemodynamic signals and the constraints they impose on neuroimaging data interpretation.

Magnetic resonance imaging (MRI) is the most important imaging advance since the introduction of X-rays by Conrad Röntgen in 1895. Since its introduction in the clinic in the 1980s, it has assumed a role of unparalleled importance in diagnostic medicine and more recently in basic research. In medicine, MRI is primarily used to produce structural images of organs, including the central nervous system, but it can also provide information on the physico-chemical state of tissues, their vascularization, and perfusion. Although all of these capacities have long been widely appreciated, it was the emergence of functional MRI (fMRI)—a technique for measuring haemodynamic changes after enhanced neural activity—in the early 1990s that had a real impact on basic cognitive neuroscience research. A recent database (ISI/Web of Science) query using the keywords 'fMRI' or 'functional MRI' or 'functional magnetic resonance imaging' returned over 19,000 peer-reviewed articles. Given that the first fMRI study without exogenous contrast agents was published in 1991, this corresponds to approximately 1,100 papers per year, or over 3 papers per day. This average obscures the actual rate of publications, as in 1992 there were four publications in total, increasing to about eight per day by 2007. About 43% of papers explore functional localization and/or cognitive anatomy associated with some cognitive task or stimulus—constructing statistical parametric maps from changes in haemodynamic responses from every point in the brain. Another 22% are region of interest studies examining the physiological properties of different brain structures, analogous to single-unit recordings; 8% are on neuropsychology; 5% on the properties of the fMRI signal; and the rest is on a variety of other topics including plasticity, drug action, experimental designs and analysis methods.

In humans, fMRI is used routinely not just to study sensory processing or control of action, but also to draw provocative conclusions about the neural mechanisms of cognitive capacities, ranging from recognition and memory to pondering ethical dilemmas. Its popular fascination is reflected in countless articles in the press speculating on potential applications, and seeming to indicate that with fMRI we can read minds better than direct tests of behaviour itself. Unsurprisingly, criticism has been just as vigorous, both among scientists and the public. In fact, fMRI is not and will never be a mind reader, as some of the proponents of decoding-based methods suggest, nor is it a worthless and non-informative 'neophrenology' that is condemned to fail, as has been occasionally argued.

Perhaps the extreme positions on both sides result from a poor understanding of the actual capacities and limitations of this technology, as well as, frequently, a confusion between fMRI shortcomings and potential flaws in modelling the organizational principles of the faculties under investigation. For example, a frequently made assumption is that the mind can be subdivided into modules or parts whose activity can then be studied with fMRI. If this assumption is false, then even if the brain's architecture is modular, we would never be able to map mind modules onto brain structures, because a unified mind has no components to speak of. Even if true, the challenge remains in coming up with the correct recursive decompositions—in each of which any given cognitive capacity, however abstract, is divided into increasingly smaller functional units that are localized to specific brain parts, which in turn can be detected and studied with fMRI. This is not a neuroimaging problem but a cognitive one. Hierarchical decompositions are clearly possible within different sensory modalities and motor systems. Their mapping, which reflects the brain's functional organization, is evidently possible and certainly meaningful beyond any reasonable doubt[1].

Here, I offer an assessment of fMRI methodology itself, leaving aside such epistemological and ontological issues. I take the modular organization of many brain systems as a well established fact, and discuss only how far fMRI can go in revealing the neuronal mechanisms of behaviour by mapping different system modules and their dynamic inter-relationships. In this context the term module captures the classical local neuronal circuits repeated iteratively within a structure (for example, the columns or swirling, slab-like tangential arrangements of the neocortex), as well as the entities within which modules might be grouped by sets of dominating external connections. The often used term functional segregation refers to such specialized and spatially separated modules. Segregated entities that are interconnected might further result in nested distributed systems, the activity of which, often termed functional integration, can only be visualized by large-scale neuroimaging.

The principal advantages of fMRI lie in its noninvasive nature, ever-increasing availability, relatively high spatiotemporal resolution, and its capacity to demonstrate the entire network of brain areas engaged when subjects undertake particular tasks. One disadvantage is that, like all haemodynamic-based modalities, it measures a surrogate signal whose spatial specificity and temporal response are subject to both physical and biological constraints. A more important

[1]Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany, and Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PL, UK.

shortcoming is that this surrogate signal reflects neuronal mass activity. Although this fact is acknowledged by the vast majority of investigators, its implications for drawing judicious conclusions from fMRI data are most frequently ignored. The aim of this review is first to describe briefly the fMRI technology used in cognitive neuroscience, and then discuss its neurobiological principles that very often limit data interpretation. I hope to point out that the ultimate limitations of fMRI are mainly due to the very fact that it reflects mass action, and much less to limitations imposed by the existing hardware or the acquisition methods. Functional MRI is an excellent tool for formulating intelligent, data-based hypotheses, but only in certain special cases can it be really useful for unambiguously selecting one of them, or for explaining the detailed neural mechanisms underlying the studied cognitive capacities. In the vast majority of cases, it is the combination of fMRI with other techniques and the parallel use of animal models that will be the most effective strategy for understanding brain function.

A brief overview of fMRI

The beautiful graphics MRI and fMRI produce, and the excitement about what they imply, often mask the immense complexity of the physical, biophysical and engineering procedures generating them. The actual details of MRI can only be correctly described via quantum mechanics, but a glimpse of the method's foundation can be also afforded with the tools of classical physics using a few simple equations. (See refs 2 and 3 for a comprehensive account of the theoretical and practical aspects of MRI, and ref. 4 for its functional variants.) Here I offer a brief overview that permits an understandable definition of the terms and parameters commonly used in magnetic resonance imaging (see 'MRI and fMRI principles' in the Supplementary Information for a description of the principles and terms of anatomical and functional MRI). Functional activation of the brain can be detected with MRI via direct measurements of tissue perfusion, blood-volume changes, or changes in the concentration of oxygen. The blood-oxygen-level-dependent (BOLD) contrast mechanism[5,6] is currently the mainstay of human neuroimaging.

Critical factors determining the utility of fMRI for drawing conclusions in brain research are signal specificity and spatial and temporal resolution. Signal specificity ensures that the generated maps reflect actual neural changes, whereas spatial and temporal resolution determine our ability to discern the elementary units of the activated networks and the time course of various neural events, respectively. The interpretability of BOLD fMRI data also depends critically on the experimental design used.

Spatiotemporal properties of BOLD fMRI. The spatiotemporal properties of fMRI are covered in some detail in the Supplementary Information. Briefly, spatial specificity increases with increasing magnetic field strength and for a given magnetic field can be optimized by using pulse sequences that are less sensitive to signals from within and around large vessels (see Fig. 1 and 'Spatial and temporal specificity' in the Supplementary Information). Spatiotemporal resolution is likely to increase with the optimization of pulse sequences, the improvement of resonators, the application of high magnetic fields, and the invention of intelligent strategies such as parallel imaging, for example, sensitivity encoding (SENSE) method (see 'Spatial resolution' section in the Supplementary Information).

Human fMRI can profit a great deal from the use of high-field scanners and by the optimization of the pulse sequences used. Surprisingly, only a minority of the studies in the cognitive sciences seem to exploit the technical innovations reported from laboratories working on magnetic resonance methodologies. Most of the top-cited cognitive neuroscience studies (approximately 70%) were carried out at 1.5 T scanners, 20% were carried out at 3 T scanners, and very few at 2 T or 4 T field strengths. About 87% of all studies used the conventional gradient-echo echoplanar imaging (GE-EPI), whereas the rest used different variants of the spin-echo echoplanar imaging (SE-EPI) sequence. This combination of low magnetic field and traditional GE-EPI is prone to many localization errors. However, as of the beginning of the twenty-first century the percentage of middle-field (3 T) studies has increased, to reach about 56% in 2007. High magnetic fields are likely to dominate magnetic resonance research

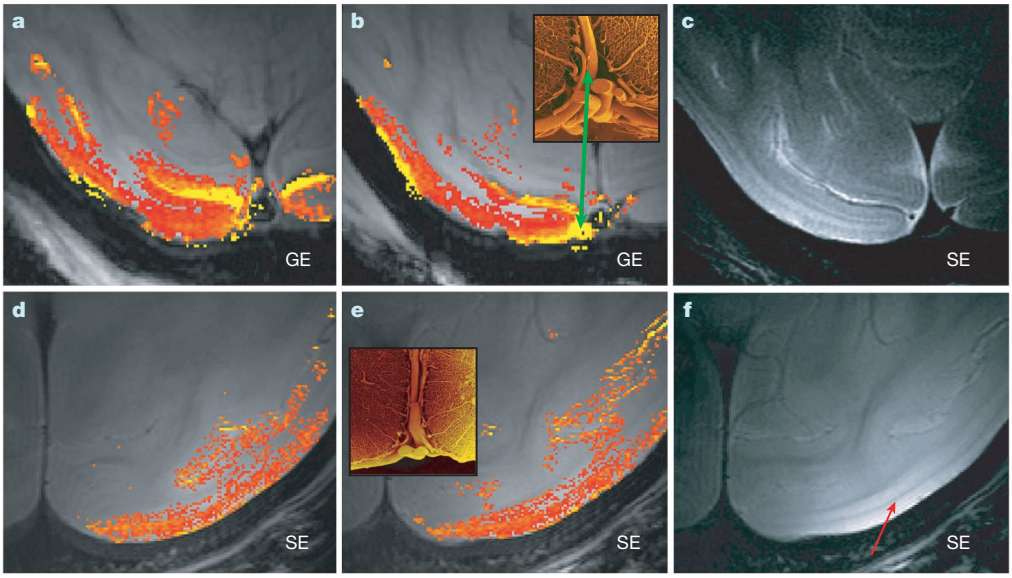


Figure 1 | Specificity of GE-EPI and SE-EPI. Examples of high-resolution GE-EPI and SE-EPI (courtesy J. Goense, MPI for Biological Cybernetics). **a**, **b**, Two slices of GE-EPI demonstrating the high functional signal-to-noise ratio (SNR) of the images, but also the strong contribution of macrovessels. The yellow areas (indicated with the green arrows) are pia vessels, an example of which is shown in the inset scanning electron microscopy image (total width of inset, 2 mm). For the functional images red indicates low and yellow indicates high. In-plane resolution 333 × 333 μm^2; slice thickness 2 mm. **c**, Anatomical scan, SE-EPI, 250 × 188 μm^2, 2 mm slice, with time to echo (TE) and repetition time (TR) 70 and 3,000 ms respectively. **d**, **e**, Two slices of SE-EPI showing the reduction of vascular contribution at the pial side of the cortex. In-plane resolution 250 × 175 μm^2, slice thickness 2 mm. **f**, The anatomical scan is the SE-EPI used for obtaining the functional scans (TE/TR = 48/2,000 ms) but at different greyscale and contrast. The resolution of the anatomical scan permits the clear visualization of the Gennari line (red arrow), the characteristic striation of the primary visual

facilities in the future, and this should definitely improve the quality of data obtained in human magnetic resonance studies. At the same time, high magnetic field scanners are likely to require even tighter interaction between magnetic resonance physicists and application scientists, as the much larger inhomogeneity of both B0 (main static field) and B1 (the field generated by the excitation pulses) at high field will demand a great deal of expertise and experimental skill to achieve the desired image quality.

All in all, MRI may soon provide us with images of a fraction of a millimetre (for example, $300 \times 300\ \mu m^2$ with a couple of millimetres slice thickness or $500 \times 500 \times 500\ \mu m^3$ isotropic), which amount to voxel volumes of about two–three orders of magnitude smaller than those currently used in human imaging (see 'Developments and perspectives' in the Supplementary Information). With an increasing number of acquisition channels such resolution may ultimately be attained in whole-head imaging protocols, yielding unparalleled maps of distributed brain activity in great regional detail and with reasonable—a couple of seconds—temporal resolution. Would that be enough for using fMRI to understand brain function?

The answer obviously depends on the scientific question and the spatial scale at which this question could be addressed—"it makes no sense to read a newspaper with a microscope", as neuroanatomist Valentino Braitenberg once pointed out. To understand the functioning of the microcircuits in cortical columns or of the cell assemblies in the striosomes of basal ganglia, one must know a great deal about synapses, neurons and their interconnections. To understand the functioning of a distributed large-scale system, such as that underlying our memory or linguistic capacities, one must first know the architectural units that organize neural populations of similar properties, and the interconnections of such units. With 10^{10} neurons and 10^{14} connections in the cortex alone, attempting to study dynamic interactions between subsystems at the level of single neurons would probably make little sense, even if it were technically feasible. It is probably much more important to understand better the differential activity of functional subunits—whether subcortical nuclei, or cortical columns, blobs and laminae—and the instances of their joint or conditional activation. If so, whole-head imaging with a spatial resolution, say, of $0.7 \times 0.7\ mm^2$ in slices of 1-mm thickness, and a sampling time of a couple of seconds, might prove optimal for the vast majority of questions in basic and clinical research. More so, because of the great sensitivity of the fMRI signal to neuromodulation (see below and Supplementary Information). Neuromodulatory effects, such as those effected by arousal, attention, memory, and so on, are slow and have reduced spatiotemporal resolution and specificity[7,8].

Designs and analyses. Many studies initially used block designs, reminiscent of earlier positron emission tomography (PET) paradigms. These designs use time-integrated averaging procedures, and usually analyse the data by means of subtraction methods. The central idea is to compare a task state designed to place specific demands on the brain with an investigator-defined control state. Under these conditions, both enhancements and reductions of the fMRI signal are observed. In the early cognitive fMRI studies the prevailing block design was cognitive subtraction, with an emphasis on serial subtraction designs[9]. Such designs rely strictly on pure insertion, which asserts that a single cognitive process can be inserted into a task without affecting the remainder, an assumption that all too often is not tenable (see 'On pure insertion' in the Supplementary Information). Even if an experimental design could satisfy this assumption at the cognitive level, the assumption would be condemned to fail at the level of its neuronal instantiation[10] owing to the highly nonlinear nature of most brain processes. To overcome this kind of problem and ensure better interpretation of the neuroimaging data it is necessary to perform a detailed task analysis to determine subtraction components and their interactions. Yet most neuroimaging studies provide no formal task analysis that would ensure that the particular cognitive process of interest is indeed being isolated by the subtraction[11]. Traditional block designs have excellent functional contrast-to-noise ratio (that is, signal difference between test and control epochs, normalized to the mean signal of all epochs), but they are usually long (from 20 to 60 s), and may be confounded by the general state of arousal of the subject. High-speed fMRI methods, capable of whole-brain imaging with a temporal resolution of a few seconds, enabled the employment of so-called event-related designs[12]. The time course of the response in such experiments is closer to the underlying neural activity.

The block designs discussed so far may reveal differential patterns of activation only in those cases in which different stimulus attributes or different cognitive processes have distinct, non-overlapping spatial organizations. Overlapping networks of neurons subserving different functions are likely to go unnoticed owing to the spatial averaging that characterizes the blocked subtraction paradigms. Functional MRI adaptation designs were conceived as tools that might, at least to some extent, tackle the problem of spatially overlapping neural networks[13]. In this experimental design, a stimulus is presented repeatedly with the expectation that it will eventually induce response adaptation in neurons selective for its various properties. In general, repetition of an identical stimulus does indeed produce a reduction in the fMRI signal. After adaptation, the subject is presented with a stimulus that is varied along one dimension (for example, the direction of a moving pattern or the view of a human face) and the possibility of a response rebound is examined. If the underlying neural representation is insensitive to the changes in the stimulus then the fMRI signal will be reduced, similar to the reduction produced by the repetition of identical stimuli. Alternatively, if the neurons are sensitive to the transformation, the signal will show a clear rebound to its original, pre-adaptation level.

Functional MRI adaptation designs have been widely used in cognitive neuroscience, but they also have shortcomings, as any area receiving input from another region may reveal adaptation effects that actually occurred in that other region, even if the receiving area itself has no neuronal specificity for the adapted property[13]. Moreover, the conclusions of experiments relying on adaptation designs strongly rely on existing electrophysiological evidence, which itself may hold true for one area and not for another[72].

Finally, clever analysis is required to exploit clever design. Most studies so far have used voxel-based conventional analyses of MRI time series from one or more subjects[14]. The approach is predicated on an extension of the general linear model that allows for correlations between error terms owing to physiological noise or correlations that ensue after temporal smoothing. The method is reliable and, when well implemented, offers the best analysis strategy for most studies. Another approach is to take into account the full spatial pattern of brain activity, measured simultaneously at many locations[15]. Such multivariate analyses or pattern-classification-based techniques (decoding techniques) can often detect small differences between two task or stimulus conditions—differences that are not picked up by conventional univariate methods. However, this is not equivalent to saying that they unequivocally reveal the neural mechanisms underlying the activation patterns. The presence, for instance, of voxels selective to two different stimulus attributes could be potentially detected by modern classifiers, yet the existence of two types of patterns does not necessarily imply the existence of two different types of neural populations[72].

What do activation maps represent?

Does the activation of an area mean that it is truly involved in the task at hand? This question implies that we understand what neural activity in a given area would unequivocally show its participation in the studied behaviour. But do we? It is usually alleged that cognitive capacities reflect the 'local processing of inputs' or the 'output' of a region, instantiated in the patterns of action potentials, with their characteristic frequency and timing. In principle, brain structures can be conceptualized as information processing entities, with an input, a local-processing capacity, and an output. Yet, although such a scheme may describe the function of subcortical nuclei, its implementation

in different areas of cortex is anything but straightforward. In fact, we now know that the traditional cortical input–elaboration–output scheme, commonly presented as an instantiation of the tripartite perception–cognition–action model, is probably a misleading oversimplification[16]. Research shows that the subcortical input to cortex is weak; the feedback is massive, the local connectivity reveals strong excitatory and inhibitory recurrence, and the output reflects changes in the balance between excitation and inhibition, rather than simple feedforward integration of subcortical inputs[17]. In the context of this review, the properties of these excitation–inhibition networks (EIN) deserve special attention, and are briefly discussed below.

Feedforward and feedback cortical processing. Brain connectivity is mostly bidirectional. To the extent that different brain regions can be thought of as hierarchically organized processing steps, connections are often described as feedforward and feedback, forward and backward, ascending and descending, or bottom-up and top-down[18]. Although all terms agree on processing direction, endowing backward connections with a role of engineering-type or functional 'feedback' might occasionally be misleading, as under a theoretical generative model perspective on brain function, it is the backward connections that generate predictions and the forward connections that convey the traditional feedback, in terms of mismatch or prediction error signals[19].

In the sensory systems, patterns of long-range cortical connectivity to some extent define feedforward and feedback pathways[20]. The main thalamic input mainly goes to middle layers, whereas second-order thalamic afferents and the nonspecific diffuse afferents from basal forebrain and brain-stem are, respectively, distributed diffusely regionally or over many cortical areas, making synapses mainly in superficial and/or deep layers. Cortical output has thalamic and other subcortical projections originating in layers VI and V, respectively, and corticocortical projections mostly from supragranular layers. The primary thalamic input innervates both excitatory and inhibitory neurons, and communication between all cell types includes horizontal and vertical connections within and between cortical layers. Such connections are divergent and convergent, so that the final response of each neuron is determined by all feedforward, feedback and modulatory synapses[17].

Very few of the pyramid synapses are thalamocortical (less than 10–20% in the input layers of cortex, and less than 5% across its entire depth; in the primary visual cortex the numbers are even lower, with the thalamocortical synapses on stellate cells being about 5%[21]), with the rest originating from other cortical pyramidal cells. Pyramidal axon collateral branches ascend back to and synapse in superficial layers, whereas others distribute excitation in the horizontal plane, forming a strongly recurrent excitatory network[17].

The strong amplification of the input signal caused by this kind of positive feedback loop is set under tight control by an inhibitory network interposed among pyramidal cells and consisting of a variety of GABAergic interneurons[22,23]. These can receive both excitatory and inhibitory synapses on to their somata, and have only local connections. About 85% of them in turn innervate the local pyramidal cells. Different GABAergic cells target different subdomains of neurons[22,24]. Some (for example, basket cells) target somata and proximal dendrites, and are excellent candidates for the role of gain adjustment of the integrated synaptic response; others (for example, chandelier cells) target directly the axons of nearby pyramidal neurons, and appear to have a context-dependent role[25]—they can facilitate spiking during low activity periods, or act like gatekeepers that shunt most complex somatodendritic integrative processes during high activity periods (for example, see up- and down states below). Such nonlinearities might generate substantial dissociations between subthreshold population activity and its concomitant metabolic demand and the spiking of pyramidal cells.

Modules and their microcircuits. A large number of structural, immunochemical and physiological studies, in all cortical areas examined so far, suggested that the functional characteristics of a cortical module are instantiated in a simple basic EIN, referred to as a canonical microcircuit[17] (see also Fig. 2a). Activation of a microcircuit sets in motion a sequence of excitation and inhibition in every neuron of the module, rather than initiating a sequential activation of separate neurons at different hypothetical processing stages. Re-excitation is tightly controlled by local inhibition, and the time evolution of excitation–inhibition is far longer than the synaptic delays of the circuits involved. This means the magnitude and timing of any local mass activation arise as properties of the microcircuits.

Computational modelling suggested that EIN microcircuits, containing such a precisely balanced excitation and inhibition, can account for a large variety of observations of cortical activity, including amplification of sensory input, noise reduction, gain control[26], stochastic properties of discharge rates[27], modulation of excitability with attention[28], or even generation of persisting activity during the delay periods of working memory tasks[29].

The principle of excitation–inhibition balance implies that microcircuits are capable of large changes in activity while maintaining proportionality in their excitatory and inhibitory synaptic conductances. This hypothesis has been tested directly in experiments examining conductance changes during periods of high (up) and low (down) cortical activity. Alternating up states and down states can be readily observed in cerebral cortex during natural sleep or anaesthesia[30], but they can be also induced *in vitro* by manipulating the ionic concentrations in a preparation so that they match those found *in situ*. Research showed that the up state is characterized by persisting synaptically mediated depolarization of the cell membranes owing to strong barrages of synaptic potentials, and a concomitant increase in spiking rate, whereas the down state is marked by membrane hyperpolarization and reduction or cessation of firing[31,32]. Most importantly, the excitation–inhibition conductances indeed changed proportionally throughout the duration of the up state despite large changes in membrane conductance[31,32].

Microcircuits therefore have the following distinct features: (1) the final response of each neuron is determined by all feedforward, feedback and modulatory synapses; (2) transient excitatory responses may result from leading excitation, for example, due to small synaptic delays or differences in signal propagation speed, whereupon inhibition is rapidly engaged, followed by balanced activity[31,32]; (3) net excitation or inhibition might occur when the afferents drive the overall excitation–inhibition balance in opposite directions; and (4) responses to large sustained input changes may occur while maintaining a well balanced excitation–inhibition. In the latter case, experimentally induced hyperpolarization of pyramidal cells may abolish their spiking without affecting the barrages of postsynaptic potentials (see ref. 31 and references therein). It is reasonable to assume that any similar hyperpolarization under normal conditions would decrease spiking of stimulus-selective neurons without affecting presynaptic activity. In visual cortex, recurrent connections among spiny stellate cells in the input layers can provide a significant source of recurrent excitation[26]. If driven by proportional excitation–inhibition synaptic currents, the impact of their sustained activity might, once again, minimally change the spiking of the pyramidal cells. This last property of microcircuits suggests that changes with balanced excitation–inhibition are good candidates for mechanisms adjusting the overall excitability and the signal-to-noise ratio (SNR) of the cortical output. Thus microcircuits—depending on their mode of operation—can, in principle, act either as drivers, faithfully transmitting stimulus-related information, or as modulators, adjusting the overall sensitivity and context-specificity of the responses[28]. Figure 2b summarizes the different types of excitation-inhibition changes and their potential effect on the haemodynamic responses.

This interesting and important driver/modulator distinction was initially drawn in the thalamus[33], in which the afferents in the major sensory thalamic relays were assigned to one of two major classes on the basis of the morphological characteristics of the axon terminals, the synaptic relationships and the type of activated receptors, the

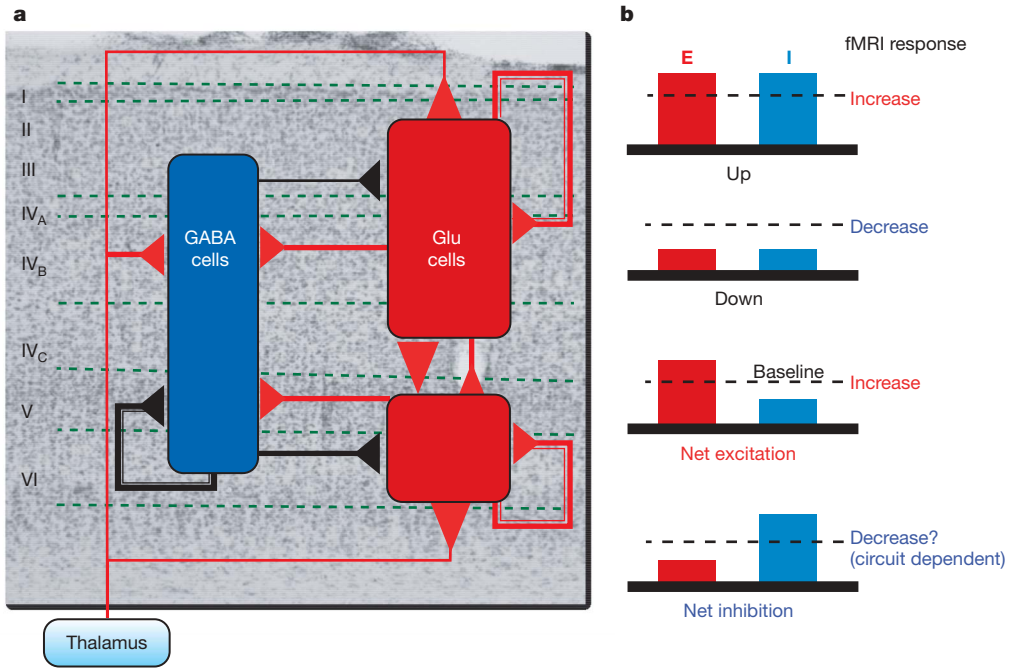


Figure 2 | Principles of excitation–inhibition circuits. a, Model of a canonical cerebral microcircuit (adapted from ref. 71). Three neuronal populations interact with each other: supragranular–granular and infragranular glutamatergic spiny neurons, and GABAergic cells. Excitatory synapses are shown in red and inhibitory synapses in black. All groups receive excitatory thalamic input. The line width indicates the strength of connection. The circuit is characterized by the presence of weak thalamic input and strong recurrence (see text for details). Glu, glutamatergic. **b**, Potential proportional and opposite-direction changes of cortical excitation (E) and inhibition (I). Responses to large sustained input changes may occur while maintaining a well balanced excitation–inhibition (up and down). The commonly assumed net excitation or inhibition might occur when the afferents drive the overall excitation–inhibition balance in opposite directions. The balanced proportional changes in excitation–inhibition activity, which occur as a result of neuromodulatory input, are likely to strongly drive the haemodynamic responses.

degree of input convergence, and the activity patterns of postsynaptic neurons. The same concept also broadly applies to the afferents of the cerebral cortex[34], wherein the thalamic or corticocortical axons terminating in layer IV can be envisaged as drivers, and other feedback afferents terminating in the superficial layers as modulators. It can also be applied to the cortical output, whereby the projections of layer VI back to the primary relays of the thalamus are modulatory, whereas the cortico-thalamo-cortical paths originating in layer V of cortex, reaching higher-order thalamic nuclei (for example, pulvinar), and then re-entering cortex via layer IV, are drivers[33].

The initial information reaching a cortical region is elaborated and evaluated in a context-dependent manner, under the influence of strong intra- and cross-regional cortical interactions. The cortical output reflects ascending input but also cortico-thalamo-cortical pathways, whereas its responsiveness and SNR reflect the activity of feedback, and likely input from the ascending diffuse systems of the brain-stem. The neuromodulation (see 'Neurotransmission and neuromodulation' in Supplementary Information) afforded by these systems, which is thought to underlie the altered states of cognitive capacities, such as motivation, attention, learning and memory, is likely to affect large masses of cells, and potentially induce larger changes in the fMRI signal than the sensory signals themselves.

Excitation–inhibition networks and fMRI. The organization discussed above evidently complicates both the precise definition of the conditions that would justify the assignment of a functional role to an 'active' area, and interpretation of the fMRI maps. Changes in excitation–inhibition balance—whether they lead to net excitation, inhibition, or simple sensitivity adjustment—inevitably and strongly affect the regional metabolic energy demands and the concomitant regulation of cerebral blood flow (CBF) (that is, they significantly alter the fMRI signal). A frequent explanation of the fMRI data simply assumes an increase in the spiking of many task- or stimulus-specific neurons. This might be correct in some cases, but increases of the BOLD signal may also occur as a result of balanced proportional increases in the excitatory and inhibitory conductances, potential concomitant increases in spontaneous spiking, but still without a net excitatory activity in stimulus-related cortical output. In the same vein, an increase in recurrent inhibition with concomitant decreases in excitation may result in reduction of an area's net spiking output, but would the latter decrease the fMRI signal? The answer to this question seems to depend on the brain region that is inhibited, as well as on experimental conditions.

Direct haemodynamic measurements with autoradiography suggested that metabolism increases with increased inhibition[35]. An exquisite example is the inhibition-induced increase in metabolism in the cat lateral superior olive (LSO). This nucleus, which contains the representations of low-, middle- and high-tone frequencies, receives afferents from both ears: over a two-neuron pathway from the ipsilateral ear and over a three-neuron pathway from the contralateral ear. Furthermore, it has no presynaptic axo-axonic endings that might mediate presynaptic inhibition via excitatory terminals. Electrophysiology showed that the LSO afferents from the ipsilateral ear are excitatory whereas the afferents from the contralateral ear are inhibitory. This unusual combination of anatomical and physiological features suggests that if one ear is surgically deafened and the animal is exposed to a high-frequency pure tone, a band of tissue in the LSO on the side opposite to the remaining active ear is subjected to strictly inhibitory synaptic activity without complications by presynaptic inhibition, concurrent lateral excitation, disinhibition/excitation, or other kinds of possibly excitatory action. Under these conditions, maps obtained with [^{14}C]2-deoxyglucose (2DG) autoradiography[36] demonstrated clear increases in metabolism in the contralateral LSO[37], suggesting that the presynaptic activity in that area is sufficient to show strong energy consumption despite the ensuing spiking reduction. Similar increases in metabolism during the reduction of spike rates were observed during long-lasting microstimulation of the fornix, which induces sustained suppression of pyramidal cell firing in hippocampus[38].

In contrast, human fMRI studies reported haemodynamic and metabolic downregulation accompanying neuronal inhibition in motor[39] and visual cortices[40], suggesting that the sustained negative BOLD response (NBR) is a marker of neuronal deactivation. Similarly, combined fMRI and electrophysiological experiments showed a clear correspondence of NBR and decreased population spiking in haemodynamically 'negative' areas in the monkey primary visual cortex[41]. Decreases in blood oxygenation and volume were also found to be co-localized with predominant neuronal inhibition and arteriolar vasoconstriction during somatosensory stimulation in rats[42]. Thus, without understanding the intrinsic correlation between direct or indirect inhibitory activity and concomitant changes in energy metabolism in a given situation, conclusions cannot be drawn. Unfortunately, the few published theoretical estimates of energy budget have not considered the metabolic costs of spikes in interneurons and of the inhibitory postsynaptic potentials (IPSPs) they produce[43]. Modelling of inhibition is unlikely to be straightforward. On the one hand, the density of cortical inhibitory neurons is 10–15 times lower than excitatory neurons[16], and for each one of them the electrochemical gradient, down which Cl^- moves postsynaptically at inhibitory synapses, is weaker than that of Na^+ at excitatory synapses, requiring less energy to pump Cl^- back. In fact, the transport cycles of the cation–chloride co-transporters, which have a key role in intracellular Cl^- regulation, are driven without the direct hydrolysis of ATP, by using the energy from the cation gradients generated by the Na,K-ATPase[44]. On the other hand, inhibitory interneurons are fast spiking[45,46]. For example, the firing of pyramidal cells in hippocampus is 1.4 Hz, whereas that of interneurons in the strata pyramidale and oriens is 15 Hz and 10 Hz, respectively. Similarly, cortical inhibitory interneurons may discharge 2–3 times faster than pyramidal cells[47]. In principle, inhibition may increase or decrease energy consumption depending on the contribution of the aforementioned factors (for a recent comprehensive review on inhibitory neurons and brain metabolism, see ref. 48). Last but not least, neurons directly affect microvessels. Pericytes, the flat, contractile connective-tissue cells, often attached to the abluminal surface of the capillary endothelial cells, might directly alter CBF in response to changes in neural activity[49]. Moreover, a body of evidence suggests that increased activity of single inhibitory interneurons results in precise vasomotor responses in neighbouring brain microvessels, and these contractile or dilatory responses were attributed to arteriole smooth muscle[50].

The diversity of the haemodynamic responses to neural inhibition obtained in different types of experiments is therefore hardly surprising: it is primarily due to the fact that regional inhibition itself might have a number of different causes, including early shunting of the weak cortical input, leading to a reduction of recurrent excitation rather than an increase in summed inhibition; increased synaptic inhibition; shunting of the cortical output through the axo-axonic connections of the chandelier cells; or any combination thereof. In the first case inhibition might result in a clear NBR; in the other two it might reflect the local metabolism increases induced by the unaffected input and its ongoing processing, resulting in fMRI activations. The fMRI responses might further blur the origin of inhibition owing to the direct effects of the latter on the arterioles and microvessels.

Evidently much research is needed to characterize the actual state of an area and its participation in behaviour, but quite independent of this fact, the nature of the EIN suggests that mass action and its surrogate haemodynamics are ambiguous signals, the interpretation of which must be constrained by the concurrent use of other methodologies.

Neurophysiological correlates of the BOLD signal

EIN and mesoscopic neural signals. The active regions of the membrane of a discharging neuron at any given time are considered to act as a current sink, whereas the inactive ones act as a current source for the active regions (see 'Neural signals' in Supplementary Information). The linear superposition of currents from all sinks and sources forms the extracellular field potential measured by microelectrodes. The extracellular field potential captures at least three different types of EIN activity: single-unit activity representing the action potentials of well isolated neurons next to the electrode tip, multiple unit activity reflecting the spiking of small neural populations in a sphere of 100–300 μm radius, and perisynaptic activity of a neural population within 0.5–3 mm of the electrode tip, which is reflected in the variation of the low-frequency components of the extracellular field potential. Multiple unit activity and local field potentials (LFPs) can be reliably segregated by frequency band separation. A high-pass filter cutoff in the range of 500–1,000 Hz is used in most recordings to obtain the multiple unit activity, and a low-pass filter cutoff of approximately 250 Hz to obtain LFP. A large number of experiments have presented data indicating that such a band separation does indeed underlie different neural events (see 'Neural signals' in Supplementary Information).

LFP signals and their different band-limited components (alpha, beta, gamma, and so on) are invaluable for understanding cortical processing, as they are the only signs of integrative EIN processes. In fact, LFPs do not, as initially thought, solely reflect population postsynaptic potentials, but also integrative soma–dendritic processes—including voltage-dependent membrane oscillations and afterpotentials following soma–dendritic spikes—that all together represent the local (perisynaptic) activity in a region (see 'Neural signals' in Supplementary Information). A shortcoming of the LFP is its ambiguity. A change in the power of LFP in a particular frequency band most likely occurs for any mode of operations of the EIN. As most of the excitatory input into an area is local, LFPs will also indirectly reflect some of the postsynaptic effects of pyramidal cell activity. In addition, LFPs have a certain neural-class bias, which in this case is determined by geometry and regional architecture. The arrangement of the pyramidal and Purkinje cells will give rise to large LFP modulations; in contrast, interneurons will contribute only weakly because of their star-shaped dendrites and their geometrical disorder. Finally, inhibitory synapses may occasionally act as 'shunts' for the excitatory currents through low-resistance channels, in which case large synaptic conductance changes may produce little effect in the membrane potential, and result in weak and hard-to-measure multiple unit activity and LFPs.

When individual LFP bands are examined separately, local spiking activity is occasionally found to affect certain frequency bands, whereas that of neuromodulation affects others[51–53]. It is evident that the most useful information will not be derived by one type of signal alone, but rather by the study of relative changes in one signal or the other. Electrophysiological studies examining the individual contributions of different LFP frequency bands, multiple unit activity, and spiking of individual neurons are probably our only realistic chance of gaining insights into the neural mechanisms of haemodynamic responses and their meaning in the context of different cognitive tasks.

Mesoscopic signals and the BOLD signal. The relationship of neocortical LFPs and spiking activity to the BOLD signal itself was examined directly in concurrent electrophysiology and fMRI experiments in the visual system of anaesthetized[54] and alert[55] monkeys. These studies found that the BOLD responses reflect input and intracortical processing rather than pyramidal cell output activity. Initially, both LFPs and spiking seemed to be correlated with the BOLD response, although quantitative analysis indicated that LFPs are better predictors of the BOLD response than multiple-unit or single-unit spiking. The decisive finding leading to the papers' conclusion, however, was not the degree of correlation between the neural and the fMRI responses or the differential contribution of any type of signal into the BOLD responses[55], but rather the striking, undiminished haemodynamic responses in cases where spiking was entirely absent despite a clear and strong stimulus-induced modulation of the field

potentials[54,55]. Similar dissociations between spikes and CBF had been demonstrated earlier and very recently in a number of studies using other techniques[56–58].

The findings are in close agreement with a number of older autoradiography studies, also showing that regional glucose utilization is directly related to neuronal synaptic activity[35]. For example, the greatest 2-DG uptake occurs in the neuropil (that is, in areas rich in synapses, dendrites and axons, rather than in cell bodies). During orthodromic and antidromic electrical microstimulation, only orthodromic microstimulation, which involves presynaptic terminals, increases glucose consumption. Similarly, the highest density of cytochrome oxidase (an enzyme of the respiratory chain) is found in somato-dendritic regions that are adjacent to axon terminals. Finally, as mentioned earlier, presynaptic activity increases metabolism even if the output is inhibited (that is, the spiking activity is abolished).

Despite all this evidence, some discussion still concentrates on the importance of the firing rate of action potentials of projection neurons in the generation of the haemodynamic responses, perhaps stemming from the fact that important early studies of neural correlates of behaviour took the mean spiking rate to be the gold standard for quantifying neuronal activation. These discussions, however, often suffer from a certain amount of contention seeking where none is warranted. In many cases, spikes do indeed correlate with LFPs, and they will also correlate with the BOLD signal. In addition, unusually high correlations between multiple unit activity and BOLD signal (or LFP and multiple unit activity) may result from excessive signal-smoothing owing to sampling rates of several seconds rather than a fraction of a second, as well as inter-subject averaging when simultaneous physiology and fMRI measurements are not possible (see ref. 55 for discussion).

Predicting neural activity from the fMRI signals. Functional MRI signals are presumed to result from changes in the activity of the neuronal populations responsible for the functions in question (for example, stimulus- or task-selective neurons). This assumption is mainly based on decades of electrophysiology research with recordings from isolated single neurons in experimental animals, in which particular sensory stimuli that the animal perceives or tasks that it performs were found to increase the firing rate of certain cells but not of others. The psychologist or cognitive neuroscientist who finds cortical area X to be activated by the task at hand implicitly or explicitly assumes that—if an electrode were placed in the subject's brain—an increase in the spiking rate of those specialized neurons underlying the subject's behaviour would be observed. This might well be true in some cases, but not in all. When attempting to interpret the fMRI signal by modelling, or when comparing the results of human neuroimaging to those obtained in monkey physiology experiments, it is useful to take the following facts into consideration.

In humans, there are about 90,000–100,000 neurons under 1 mm^2 of cortical surface. This number is relatively constant for all structurally and functionally distinct areas, including the somatosensory, temporal, parietal, frontal and motor cortical areas[16,59]. An exception is the primary visual cortex of certain primates, including monkey and human, which has approximately twice as many neurons. The number of cortical neurons under unitary cortical surface is also similar across many species, including mouse, rat, cat, monkey and human. Its small variability is the result of a trade-off between cortical thickness and neural density. The former varies from area to area and from species to species (for example, from mouse to human the cortex becomes approximately three times thicker). Neural density varies inversely to cortical thickness. On average, density is 20,000 to 30,000 neurons per mm^3; it peaks in the primary visual cortex by a factor of 4, and it is minimal in the motor cortex[59,60]. Synaptic density ranges from 0.4 to 1×10^9 per mm^3. Depending on the thickness of the cortex (2–4 mm), the number of synapses beneath 1 mm^2 surface is around 10^9 ($0.8–4 \times 10^9$). Although the number of synapses and the axonal length per neuron increases with increasing cortical thickness[61], the overall length of neuronal processes remains relatively constant, with axonal length being approximately 4 km mm^{-3} and dendrite length 0.4 km mm^{-3}. Overall, synaptic density and the ratio of excitatory to inhibitory synapses also remain constant.

Given these neuro-statistical data, what are the actual contents of a neuroimaging voxel? An examination of the 300 top-cited cognitive fMRI studies suggests that the commonly used in-plane resolution is 9–16 mm^2, for slice thicknesses of 5–7 mm. The average voxel size before any pre-processing of the data is thus 55 μl (or 55 mm^3). Often the effective size is 2–3 times larger due to the spatial filtering that most investigators apply to improve the functional SNR. Less than 3% of this volume is occupied by vessels and the rest by neural elements (see Fig. 3) A typical unfiltered fMRI voxel of 55 μl in size thus contains 5.5 million neurons, $2.2–5.5 \times 10^{10}$ synapses, 22 km of dendrites and 220 km of axons.

This 'large population view' is in contrast to the scope of the traditional microelectrode recordings. It would be nice if we could monitor every relevant neuron in the cortex during intracortical microelectrode recordings, but this is practically impossible. Instead, the typical electrophysiological measurements in behaving animals report only on the properties of most active large neurons that constitute a minority. The strong selection bias during extracellular recordings is partly due to practical limitations (for example, injury or simply size bias[62]) and partly to the physiological properties of neurons and/or the organizational principles of neural networks. In fact, many different types of electrical and optical measurements provide evidence that a substantial proportion of neurons, including the cortical pyramidal cells, might be silent[63]. Their silence might reflect unusually high input selectivity or the existence of decoding schemes relying on infrequent co-spiking of neuronal subsets. Most important for the comparison of neuroimaging and electrophysiology results is the fact that lack of measurable neuronal spiking may not necessarily imply lack of input and subthreshold processing.

A direct analogy between neuronal spiking as measured in animal experiments and the fMRI signal obtained in human recording is thus simply unrealistic and might often lead to incorrect conclusions. It is hardly surprising that most studies so far relying purely on BOLD fMRI have failed to reveal the actual neural properties of the studied area, at least those properties (for example, selectivity to various visual features) that were previously established in electrophysiological studies.

An example is cortical area V5 (or MT) that has been extensively studied in the context of motion processing and perception[64,65]. Electrophysiology has shown that the vast majority of the V5 neurons in monkeys are direction and speed selective. Neuroimaging localized the homologue of area V5 in humans as an area responding stronger to moving than to stationary stimuli. Later studies suggested that human V5 is sensitive to motion direction, and that it may be thought of as containing large populations of directionally selective units, just like its monkey homologue. The studies of directional specificity exploited the phenomenon of motion after-effect induced by motion adaptation. After prolonged exposure to a stimulus moving in one direction, subjects perceive a subsequent static stimulus to move in the opposite direction. It is assumed that motion after-effect is due to the fact that the balance of mutual inhibition (opponency) between detectors for opposite directions of movement is distorted after adaptation. The sensitivity of the detectors selective for the adapting direction is reduced, which in turn releases from inhibition the neurons selective for the opposite direction[66]. Using this phenomenon, human studies demonstrated that the fMRI response to a stationary stimulus was greater when the stimulus was preceded by a motion-after-effect-inducing, unidirectional adaptation, than when preceded by bidirectional adaptation[67]. Given the existing physiology data in the monkey V5, these findings were interpreted as demonstrating that the BOLD signal directly reflects direction-selective spiking activity of the area.

Yet, as I have indicated above, the BOLD signal is primarily affected by changes in excitation–inhibition balance, and this balance may be controlled by neuromodulation more than by the changes in spiking rate of a small set of neurons. In fact, the BOLD signal is strongly modulated by attention[68], and the results of the motion after-effect experiments could, in principle, be due to the fact that a stimulus with illusory motion automatically draws the attention of a subject more compared to a situation in which there is no motion after-effect. This hypothesis turned out to be correct, as a later study—in which balance in attentional load was accomplished by having the subjects perform a concurrent visual task—found no signal differences between the motion after-effect and no motion after-effect conditions[69].

A similar example pertains to the differences in neurophysiological and fMRI responses in the primary visual cortex during different perceptual states. It is known that physiological signals are in general stronger when stimuli are perceived as opposed to when they are not. Intriguingly, in some regions the BOLD response seems to reflect this even more sensitively than physiological measures like spikes and multi-unit activity[70]. An example is the pattern of fMRI activation changes in V1 during binocular rivalry (that is, the perceptual alternations experienced when the two eyes view different stimuli). This phenomenon has been studied extensively psychophysically and also over the last two decades in a series of electrophysiology studies in monkeys[70]. These studies showed that only a small fraction of V1 cells modulate their spiking during the perceptual changes; neuroimaging, on the other hand, demonstrated fMRI-signal modulations that were nearly as large as those obtained during the physical alternation of stimuli[70]. The difference, once again, reflects the fact that neuromodulatory feedback from higher areas can be easily detected by means of fMRI, but not through the measurement of single-unit activity. Interestingly, measurements of subthreshold activity in another study of perceptual multistability revealed perception-related modulations in LFP, despite the unaltered spike rates[53]. Such clear spiking and BOLD signal mismatches appear even in simple experiments probing sensory processing. Simple stimuli, such as those used in the aforementioned studies, are most likely to generate a proportional enhancement in both the afferent and efferent activity of any sensory area. The activation of high-level association areas related to cognitive processing might be more sensitive or even dominated by feedback and neuromodulation, whose differential effect on spiking and haemodynamic responses is utterly unknown.

Conclusions and perspectives

The limitations of fMRI are not related to physics or poor engineering, and are unlikely to be resolved by increasing the sophistication and power of the scanners; they are instead due to the circuitry and

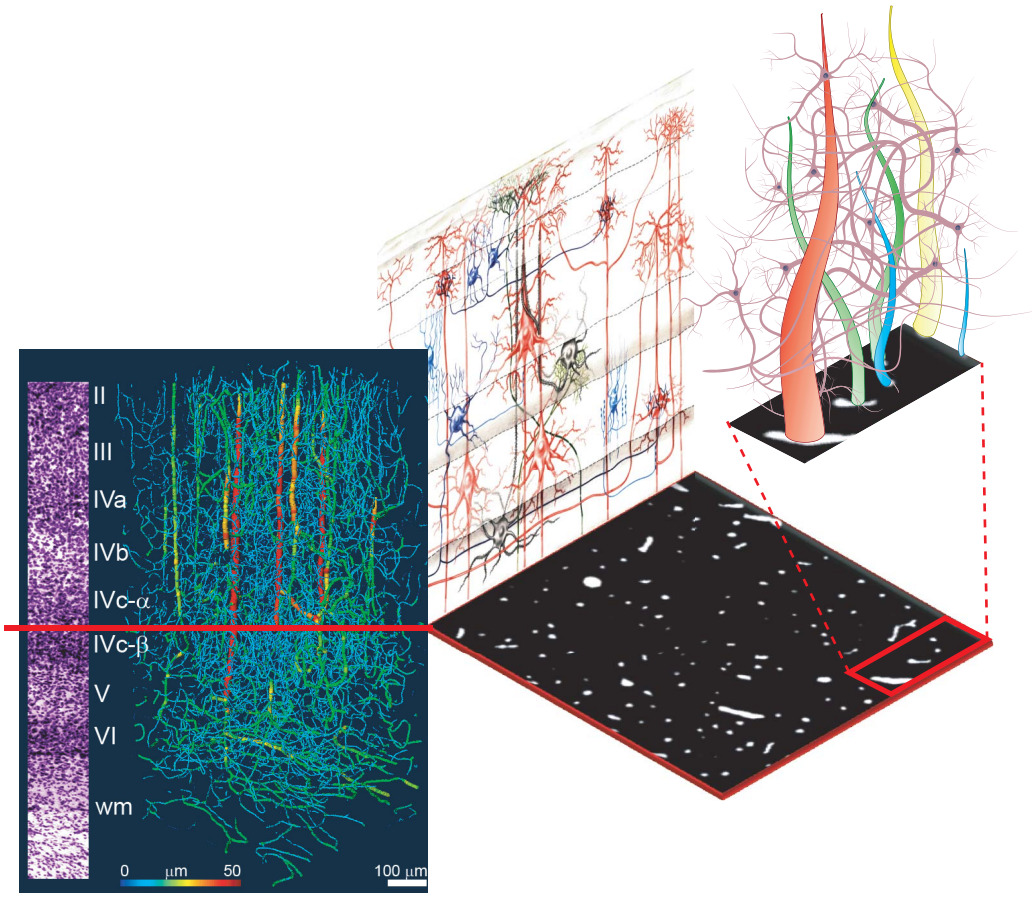


Figure 3 | Neural and vascular contents of a voxel. The left panel demonstrates the relative density of vessels in the visual cortex of monkeys. The dense vascular mesh is displayed by perfusing the tissue with barium sulphate and imaging it with synchrotron-based X-ray microtomography (courtesy B. Weber, MPI for Biological Cybernetics). The vessel diameter is colour coded. Cortical surface without pial vessels is displayed at the top; white matter at the bottom. At the left of the panel is a Nissl slice from the same area, showing the neural density for layers II through to the white matter (wm). Although the density of the vessels appears to be high in this three-dimensional representation, it is actually less the 3% (see section at the right; white spots are cross-sections of vessels). The average distance between the small vessels (capillaries) is about 50 µm. This is approximately the distance that oxygen molecules travel by diffusion within the limited transit time of the blood. The dense population of neurons, synapses and glia occupy the intervascular space, as depicted in the drawing at the top right—a hypothetical distribution of vascular and neural elements in a small section (red rectangle). The drawing in the background shows some of the typical neuronal types (for example, red, large pyramidal cell; dark blue, inhibitory basket cells; light blue, chandelier inhibitory neurons; and grey, stellate cells) and their processes.

functional organization of the brain, as well as to inappropriate experimental protocols that ignore this organization. The fMRI signal cannot easily differentiate between function-specific processing and neuromodulation, between bottom-up and top-down signals, and it may potentially confuse excitation and inhibition. The magnitude of the fMRI signal cannot be quantified to reflect accurately differences between brain regions, or between tasks within the same region. The origin of the latter problem is not due to our current inability to estimate accurately cerebral metabolic rate of oxygen ($CMRO_2$) from the BOLD signal, but to the fact that haemodynamic responses are sensitive to the size of the activated population, which may change as the sparsity of neural representations varies spatially and temporally. In cortical regions in which stimulus- or task-related perceptual or cognitive capacities are sparsely represented (for example, instantiated in the activity of a very small number of neurons), volume transmission (see Supplementary Information)—which probably underlies the altered states of motivation, attention, learning and memory—may dominate haemodynamic responses and make it impossible to deduce the exact role of the area in the task at hand. Neuromodulation is also likely to affect the ultimate spatiotemporal resolution of the signal.

This having been said, and despite its shortcomings, fMRI is currently the best tool we have for gaining insights into brain function and formulating interesting and eventually testable hypotheses, even though the plausibility of these hypotheses critically depends on used magnetic resonance technology, experimental protocol, statistical analysis and insightful modelling. Theories on the brain's functional organization (not just modelling of data) will probably be the best strategy for optimizing all of the above. Hypotheses formulated on the basis of fMRI experiments are unlikely to be analytically tested with fMRI itself in terms of neural mechanisms, and this is unlikely to change any time in the near future.

Of course, fMRI is not the only methodology that has clear and serious limitations. Electrical measurements of brain activity, including invasive techniques with single or multiple electrodes, also fall short of affording real answers about network activity. Single-unit recordings and firing rates are better suited to the study of cellular properties than of neuronal assemblies, and field potentials share much of the ambiguity discussed in the context of the fMRI signal. None of the above techniques is a substitute for the others. Today, a multimodal approach is more necessary than ever for the study of the brain's function and dysfunction. Such an approach must include further improvements to MRI technology and its combination with other non-invasive techniques that directly assess the brain's electrical activity, but it also requires a profound understanding of the neural basis of haemodynamic responses and a tight coupling of human and animal experimentation that will allow us to fathom the homologies between humans and other primates that are amenable to invasive electrophysiological and pharmacological testing. Claims that computational methods and non-invasive neuroimaging (that is, excluding animal experimentation) should be sufficient to understand brain function and disorders are, in my opinion, naive and utterly incorrect. If we really wish to understand how our brain functions, we cannot afford to discard any relevant methodology, much less one providing direct information from the actual neural elements that underlie all our cognitive capacities.

1. Wandell, B. A., Brewer, A. A. & Dougherty, R. F. Visual field map clusters in human cortex. *Phil. Trans. R. Soc. Lond. B* **360,** 693–707 (2005).
   This paper provides a description of human visual field maps and the rationale generating and naming them.
2. Haacke, E. M. *et al. Magnetic Resonance Imaging: Principles and Sequence Design* (John Wiley & Son, New York, 1999).
3. Wood, M. L. & Wehrli, F. W. Principles of magnetic resonance imaging. In *Magnetic Resonance Imaging* 3rd edn (eds Stark, D. D. & Bradley, W.) 1–14 (Mosby, St Louis/Baltimore/Boston/London/Tokyo, 1999).
4. Buxton, R. B. *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques* (Cambridge Univ. Press, Cambridge, UK, 2002).
5. Ogawa, S. & Lee, T. M. Magnetic resonance imaging of blood vessels at high fields: *in vivo* and *in vitro* measurements and image simulation. *Magn. Reson. Med.* **16,** 9–18 (1990).
6. Ogawa, S. *et al.* Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14,** 68–78 (1990).
7. Motter, B. C. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70,** 909–919 (1993).
8. Luck, S. J., Chelazzi, L., Hillyard, S. A. & Desimone, R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* **77,** 24–42 (1997).
9. Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M. & Raichle, M. E. Positron emission tomographic studies of the processing of single words. *J. Cogn. Neurosci.* **1,** 153–170 (1989).
10. Friston, K. J. *et al.* The trouble with cognitive subtraction. *Neuroimage* **4,** 97–104 (1996).
11. Poeppel, D. A critical review of PET studies of phonological processing. *Brain Lang.* **55,** 317–351 (1996).
12. Buckner, R. L. *et al.* Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc. Natl Acad. Sci. USA* **93,** 14878–14883 (1996).
13. Krekelberg, B., Boynton, G. M. & van Wezel, R. J. Adaptation: from single cells to BOLD signals. *Trends Neurosci.* 29, 250–256 (2006).
    This paper discusses fMRI adaptation designs.
14. Friston, K. J. *et al.* Analysis of fMRI time-series revisited. *Neuroimage* **2,** 45–53 (1995).
15. Haynes, J. D. & Rees, G. Decoding mental states from brain activity in humans. *Nature Rev. Neurosci.* **7,** 523–534 (2006).
16. Braitenberg, V. & Schuez, A. *Cortex: Statistics and Geometry of Neuronal Connectivity* 2nd edn (Springer, Berlin, 1998).
17. Douglas, R. J. & Martin, K. A. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27,** 419–451 (2004).
    This paper provides a review of cortical microcircuits.
18. Ullman, S. Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cereb. Cortex* **5,** 1–11 (1995).
19. Friston, K. A theory of cortical responses. *Phil. Trans. R. Soc. B* **360,** 815–836 (2005).
20. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in primate cerebral cortex. *Cereb. Cortex* **1,** 1–47 (1991).
21. Douglas, R. J. & Martin, K. A. Mapping the matrix: the ways of neocortex. *Neuron* **56,** 226–238 (2007).
22. Freund, T. F. Interneuron diversity series: Rhythm and mood in perisomatic inhibition. *Trends Neurosci.* **26,** 489–495 (2003).
23. Markram, H. *et al.* Interneurons of the neocortical inhibitory system. *Nature Rev. Neurosci.* **5,** 793–807 (2004).
    This paper is a review on the various types of interneuron.
24. DeFelipe, J. Types of neurons, synaptic connections and chemical characteristics of cells immunoreactive for calbindin-D28K, parvalbumin and calretinin in the neocortex. *J. Chem. Neuroanat.* **14,** 1–19 (1997).
25. Szabadics, J. *et al.* Excitatory effect of GABAergic axo-axonic cells in cortical microcircuits. *Science* **311,** 233–235 (2006).
26. Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. & Suarez, H. H. Recurrent excitation in neocortical circuits. *Science* **269,** 981–985 (1995).
27. Shadlen, M. N. & Newsome, W. T. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.* **4,** 569–579 (1994).
28. Chance, F. S., Abbott, L. F. & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* **35,** 773–782 (2002).
29. Brunel, N. & Wang, X. J. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* **11,** 63–85 (2001).
30. Steriade, M., Timofeev, I. & Grenier, F. Natural waking and sleep states: a view from inside neocortical neurons. *J. Neurophysiol.* **85,** 1969–1985 (2001).
31. McCormick, D. A., Shu, Y. S. & Hasenstaub, A. Balanced recurrent excitation and inhibition in local cortical networks. in *Excitatory-Inhibitory Balance: Synapses, Circuits, Systems* (ed. Hensch, T.) (Kluver Academic Press, New York, 2003).
32. Haider, B., Duque, A., Hasenstaub, A. R. & McCormick, D. A. Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* 26, 4535–4545 (2006).
    This paper provides a demonstration of the regulation of excitation–inhibition balance changes *in vivo*.
33. Sherman, S. M. & Guillery, R. W. *Exploring the Thalamus and its Role in Cortical Function* 2nd edn (MIT Press, Cambridge, Massachusetts, 2006).
34. Crick, F. & Koch, C. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* **391,** 245–250 (1998).
35. Jueptner, M. & Weiller, C. Review: does measurement of regional cerebral blood flow reflect synaptic activity? Implications for PET and fMRI. *Neuroimage* **2,** 148–156 (1995).
36. Sokoloff, L. *et al.* The [^{14}C]deoxyglucose method for the measurement of local cerebral glucose utilization: Theory, procedure and normal values in the conscious and anesthetized albino rat. *J. Neurochem.* **28,** 897–916 (1977).
37. Nudo, R. J. & Masterton, R. B. Stimulation-induced [^{14}C]2-deoxyglucose labeling of synaptic activity in the central auditory system. *J. Comp. Neurol.* **245,** 553–565 (1986).

38. Ackermann, R. F., Finch, D. M., Babb, T. L. & Engel, J. Jr. Increased glucose metabolism during long-duration recurrent inhibition of hippocampal pyramidal cells. *J. Neurosci.* **4,** 251–264 (1984).
39. Stefanovic, B., Warnking, J. M. & Pike, G. B. Hemodynamic and metabolic responses to neuronal inhibition. *Neuroimage* **22,** 771–778 (2004).
40. Shmuel, A. *et al.* Sustained negative BOLD, blood flow and oxygen consumption response and its coupling to the positive response in the human brain. *Neuron* **36,** 1195–1210 (2002).
41. Shmuel, A., Augath, M., Oeltermann, A. & Logothetis, N. K. Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. *Nature Neurosci.* **9,** 569–577 (2006).
42. Devor, A. *et al.* Suppressed neuronal activity and concurrent arteriolar vasoconstriction may explain negative blood oxygenation level-dependent signal. *J. Neurosci.* **27,** 4452–4459 (2007).
43. Attwell, D. & Gibb, A. Neuroenergetics and the kinetic design of excitatory synapses. *Nature Rev. Neurosci.* **6,** 841–849 (2005).
44. Payne, J. A., Rivera, C., Voipio, J. & Kaila, K. Cation-chloride co-transporters in neuronal communication, development and trauma. *Trends Neurosci.* **26,** 199–206 (2003).
45. McCormick, D. A., Connors, B. W., Lighthall, J. W. & Prince, D. A. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.* **54,** 782–806 (1985).
46. Buzsaki, G., Geisler, C., Henze, D. A. & Wang, X. J. Interneuron diversity series: Circuit complexity and axon wiring economy of cortical interneurons. *Trends Neurosci.* **27,** 186–193 (2004).
47. Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C. Z. & Markram, H. Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cereb. Cortex* **12,** 395–410 (2002).
48. Buzsaki, G., Kaila, K. & Raichle, M. Inhibition and brain work. *Neuron* **56,** 771–783 (2007).
49. Peppiatt, C. M. *et al.* Bidirectional control of CNS capillary diameter by pericytes. *Nature* **443,** 700–704 (2006).
50. Hamel, E. Perivascular nerves and the regulation of cerebrovascular tone. *J. Appl. Physiol.* **100,** 1059–1064 (2006).
51. Kayser, C. & Konig, P. Stimulus locking and feature selectivity prevail in complementary frequency ranges of V1 local field potentials. *Eur. J. Neurosci.* **19,** 485–489 (2004).
52. Liu, J. & Newsome, W. T. Local field potential in cortical area MT: Stimulus tuning and behavioral correlations. *J. Neurosci.* **26,** 7779–7790 (2006).
53. Wilke, M., Logothetis, N. K. & Leopold, D. A. Local field potential reflects perceptual suppression in monkey visual cortex. *Proc. Natl Acad. Sci. USA* **103,** 17507–17512 (2006).
54. Logothetis, N. K. *et al.* Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412,** 150–157 (2001).
55. Goense, J. B. M. & Logothetis, N. K. Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current Biol.* **18,** 631–640 (2008).
56. Mathiesen, C., Caesar, K., Akgoren, N. & Lauritzen, M. Modification of activity-dependent increases of cerebral blood flow by excitatory synaptic activity and spikes in rat cerebellar cortex. *J. Physiol.* **512,** 555–566 (1998).
57. Viswanathan, A. & Freeman, R. D. Neurometabolic coupling in cerebral cortex reflects synaptic more than spiking activity. *Nature Neurosci.* 10, 1308–1312 (2007).
This paper provides a demonstration of the coupling between $CMRO_2$ and the LFP.
58. Rauch, A., Rainer, G. & Logothetis, N. K. The effect of a serotonin-induced dissociation between spiking and perisynaptic activity on BOLD functional MRI. *Proc. Natl Acad. Sci. USA* **105,** 6759–6764 (2008).
59. Rockel, A. J., Hiorns, R. W. & Powell, T. P. The basic uniformity in structure of the neocortex. *Brain* **103,** 221–244 (1980).
60. Cragg, B. G. The density of synapses and neurones in the motor and visual areas of the cerebral cortex. *J. Anat.* **101,** 639–654 (1967).
61. Schuz, A. & Demianenko, G. P. Constancy and variability in cortical structure. A study on synapses and dendritic spines in hedgehog and monkey. *J. Hirnforsch.* **36,** 113–122 (1995).
62. Logothetis, N. K. & Wandell, B. A. Interpreting the BOLD signal. *Annu. Rev. Physiol.* **66,** 735–769 (2004).
63. Shoham, S., O'Connor, D. H. & Segev, R. How silent is the brain: is there a "dark matter" problem in neuroscience? *J. Comp. Physiol. A* **192,** 777–784 (2006).
64. Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu. Rev. Neurosci.* **28,** 157–189 (2005).
65. Zeki, S. Thirty years of a very special visual area, area V5. *J. Physiol.* **557,** 1–2 (2004).
66. Mather, G., Verstraten, F. A. & Anstis, S. M. *The Motion Aftereffect: a Modern Perspective* (MIT Press, Cambridge, Massachusetts, 1998).
67. Tootell, R. B. H. *et al.* Visual motion aftereffect in human cortical area MT revealed by functional magnetic-resonance-imaging. *Nature* **375,** 139–141 (1995).
68. Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L. & Petersen, S. E. Attentional modulation of neural processing of shape, color, and velocity in humans. *Science* **248,** 1556–1559 (1990).
69. Huk, A. C., Ress, D. & Heeger, D. J. Neuronal basis of the motion aftereffect reconsidered. *Neuron* **32,** 161–172 (2001).
70. Blake, R. & Logothetis, N. K. Visual competition. *Nature Rev. Neurosci.* **3,** 13–21 (2002).
71. Douglas, R. J., Martin, K. A. C. & Whitteridge, D. A canonical microcircuit for neocortex. *Neural Comput.* **1,** 480–488 (1989).
72. Bartels, A., Logothetis, N. K. & Moutoussis, K. FMRI and its interpretations: An illustration on directional sensitivity in area V5/MT. *Trends Neurosci.* (in the press).

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

Acknowledgements I thank my co-workers A. Bartels, J. Goense, M. Munk and A.-C. Zappe for discussions; my colleagues P. Hoffmann, C. Koch, K. Martin, A. Schüz, C. Kayser and R. Turner for their insightful comments and suggestions on the latest version of the article; J. Goense, B. Weber and A. L. Keller for providing graphics; and D. Blaurock for language corrections. The work is supported by the Max Planck Society.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to N.K.L. (nikos.logothetis@tuebingen.mpg.de).

ARTICLES

A phosphatase cascade by which rewarding stimuli control nucleosomal response

Alexandre Stipanovich[1,2,3]*, Emmanuel Valjent[1,2,3]*, Miriam Matamales[1,2,3]*, Akinori Nishi[4,5], Jung-Hyuck Ahn[5], Matthieu Maroteaux[1,2,3], Jesus Bertran-Gonzalez[1,2,3], Karen Brami-Cherrier[1,2,3], Hervé Enslen[1,2,3], Anne-Gaëlle Corbillé[1,2,3], Odile Filhol[6], Angus C. Nairn[5,7], Paul Greengard[5], Denis Hervé[1,2,3] & Jean-Antoine Girault[1,2,3]

Dopamine orchestrates motor behaviour and reward-driven learning. Perturbations of dopamine signalling have been implicated in several neurological and psychiatric disorders, and in drug addiction. The actions of dopamine are mediated in part by the regulation of gene expression in the striatum, through mechanisms that are not fully understood. Here we show that drugs of abuse, as well as food reinforcement learning, promote the nuclear accumulation of 32-kDa dopamine-regulated and cyclic-AMP-regulated phosphoprotein (DARPP-32). This accumulation is mediated through a signalling cascade involving dopamine D1 receptors, cAMP-dependent activation of protein phosphatase-2A, dephosphorylation of DARPP-32 at Ser 97 and inhibition of its nuclear export. The nuclear accumulation of DARPP-32, a potent inhibitor of protein phosphatase-1, increases the phosphorylation of histone H3, an important component of nucleosomal response. Mutation of Ser 97 profoundly alters behavioural effects of drugs of abuse and decreases motivation for food, underlining the functional importance of this signalling cascade.

Midbrain dopamine neurons, activated after unexpected rewarding stimuli, are essential in reinforcement learning[1]. Drugs of abuse mimic the physiological action of dopamine neurons by increasing their firing rate or preventing the uptake of dopamine. Thus, they enhance extracellular dopamine levels in the forebrain, especially in the nucleus accumbens (NAc), a key structure required for the reinforcing effects of addictive drugs[2–4]. To understand how dopamine mediates natural or pathological reinforcement learning, it is necessary to identify the intracellular events that trigger gene transcription alterations supporting long-lasting synaptic changes[5–7]. DARPP-32 (ref. 8) is a prominent mediator of dopamine signalling in the striatum[9] and is highly enriched in striatal medium-size spiny GABA neurons[10]. After activation of the dopamine D1 receptor (D1R), DARPP-32 is phosphorylated by cAMP-dependent protein kinase (PKA) at Thr 34 and converted into a potent inhibitor of the multifunctional serine/threonine protein phosphatase-1 (PP1)[11]. DARPP-32-mediated inhibition of PP1 increases the phosphorylation of neurotransmitter receptors and ion channels crucial for synaptic function and plasticity[9]. DARPP-32 also regulates nuclear events, as demonstrated by alterations of drug-induced gene expression in mice lacking DARPP-32 or bearing a point mutation of Thr 34 (refs 12, 13). Part of the control exerted by DARPP-32 on transcription is mediated by activation of the extracellular signal-regulated kinase pathway, which is dependent on the concomitant stimulation of D1R and glutamate *N*-methyl-D-aspartate receptors[13,14]. However, the precise mechanisms of information transfer from the cytoplasm to the nucleus of striatal neurons are still poorly understood.

Drugs and incentive learning increase nuclear DARPP-32

DARPP-32 has been extensively characterized as a cytoplasmic protein[9]. In contrast, electron microscopy has revealed immunoreactivity to DARPP-32 in the nucleus of some striatal medium-size spiny neurons[15], although the significance of this observation has not been explained. In mice treated with D-amphetamine, immunoreactivity to DARPP-32 phosphorylated on Thr 34 was strong in the nuclei of striatal neurons (Fig. 1a). D-Amphetamine also increased total DARPP-32 immunoreactivity in nuclei (Fig. 1a, b). Cocaine induced a rapid accumulation of DARPP-32 in the nucleus that lasted several hours (Fig. 1b and Supplementary Fig. 1a). D-Amphetamine and cocaine triggered the nuclear accumulation of DARPP-32 in both the dorsal striatum and the NAc (data not shown). In contrast, morphine increased nuclear DARPP-32 only in the NAc shell (Fig. 1b and Supplementary Fig. 1b), where it potently increases dopamine release[16]. A reinforcement learning model, in which mice nose-poked for food, also triggered the nuclear accumulation of DARPP-32 in the dorsal striatum and in the NAc shell and core (Fig. 1c). In mice receiving no food, or in yoked mice that received food when the 'active' animal nose-poked, no nuclear accumulation was observed (Fig. 1c).

In *drd1a*-EGFP mice, which specifically express enhanced green fluorescent protein (EGFP) in striatonigral neurons under the control of the D1R promoter (*drd1a*)[17], nuclear accumulation of DARPP-32 occurred in D1R-expressing neurons 8 min after cocaine injection (Supplementary Fig. 2a). Cocaine-induced rapid nuclear accumulation of DARPP-32 was not observed in D1R-null mice (Supplementary Fig. 2b), indicating the requirement for D1R. In striatal neurons in culture, DARPP-32 immunofluorescence was predominantly cytoplasmic; application of SKF81297, a D1R agonist, induced its nuclear accumulation, which was prevented by a D1R antagonist, SCH23390, or a cAMP antagonist, Rp-cAMPS (Supplementary Fig. 3a, b). Conversely, a cAMP analogue, Sp-5,6-DCl-cBIMPS, increased nuclear DARPP-32 immunoreactivity in a

[1]Inserm, UMR-S 839, 75005 Paris, France. [2]Université Pierre et Marie Curie (UPMC), 75005 Paris, France. [3]Institut du Fer à Moulin, 75005 Paris, France. [4]Department of Pharmacology, Kurume University School of Medicine, Kurume, Fukuoka 830-0011, Japan. [5]Laboratory of Molecular and Cellular Neuroscience, The Rockefeller University, New York, New York 10021, USA. [6]Inserm, U873, CEA, 38054 Grenoble, France. [7]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut 06508, USA. *These authors contributed equally to this work.

time-dependent manner (Supplementary Fig. 3c). Similar results were obtained with DARPP-32 fused to EGFP (D32–GFP), co-expressed with D1R in striatal neurons in culture (Supplementary Fig. 4a–c). Taken together, our results show that nuclear accumulation of DARPP-32 in response to physiological reward-controlled learning or to drugs of abuse was mediated by D1R and that cAMP was necessary and sufficient for this effect.

Ser 97 phosphorylation controls DARPP-32 localization

Leptomycin B (LMB), a specific inhibitor of CRM1-mediated nuclear export[18], induced a rapid nuclear accumulation of endogenous DARPP-32 (Fig. 2a) or transfected D32–GFP (Supplementary Fig. 5a and Supplementary Video) in virtually all neurons. At residues 103–111 of DARPP-32 we identified a putative nuclear export signal (NES; Fig. 2b) that was homologous to sequences known to bind CRM1 (ref. 19). Mutation of Leu 103 or Leu 109 to alanine induced a marked accumulation of D32–GFP in the nucleus (Fig. 2c), confirming the critical role of this sequence in nuclear export. Similar results were observed in Chinese hamster ovary cells (Supplementary Fig. 5b, c). We also found that the amino-terminal region of DARPP-32, which encompasses several putative nuclear localization signals, had a critical function in its nuclear import (data not shown).

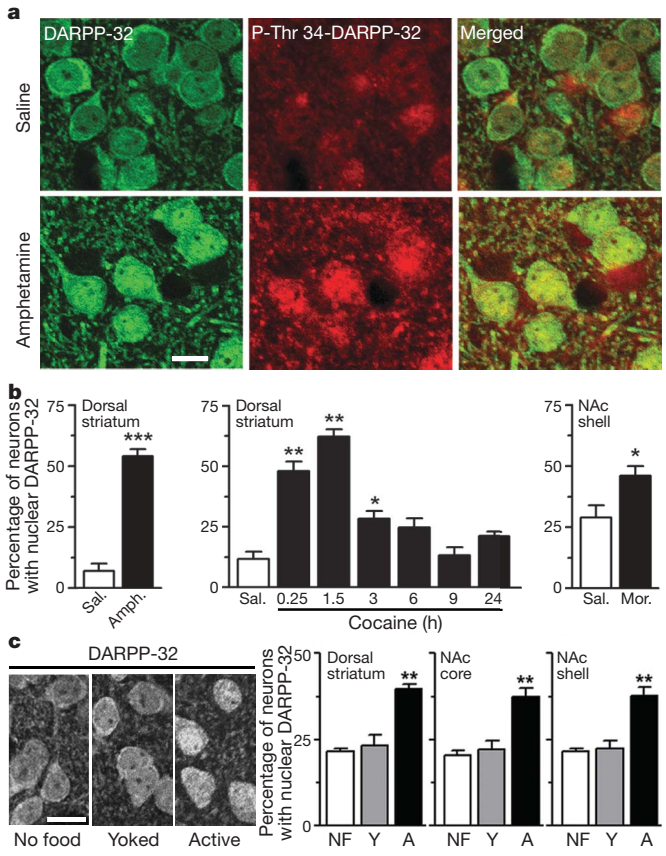


Figure 1 | Drugs of abuse and food self-administration induce nuclear accumulation of DARPP-32 in striatal neurons. **a**, DARPP-32 (green) and phospho-Thr 34-DARPP-32 (red) immunofluorescence in dorsal striatum of mice injected with saline or D-amphetamine (10 mg kg^{-1} intraperitoneally for 15 min). **b**, Effects of D-amphetamine (amph.), cocaine (20 mg kg^{-1}, time course) and morphine (mor., 5 mg kg^{-1} subcutaneously). Sal, saline. **c**, Effects of 1 h of reinforcement learning: mice in self-administration chamber without food (NF), learning to nose-poke for food pellets (active; A) or receiving pellets when the active animal nose-poked (yoked; Y). Error bars represent s.e.m., $n = 4$–8, *t*-test (amphetamine, morphine) or one-way analysis of variance (Supplementary Table 1) and Bonferroni test. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. Single confocal sections. Scale bars, 10 μm.

These results suggested that D1R stimulation acted by reducing the export of DARPP-32 out of the nucleus. We ruled out a general inhibitory influence of D1R stimulation on CRM1-dependent nuclear export, by showing that SKF81297 did not alter the localization of another protein with an active NES[20], namely Ca^{2+}/calmodulin-dependent kinase-Iα (CaMKIα–GFP; Fig. 2d). Introduction of an ectopic NES sequence derived from prohibitin[21] (NES^{proh}) at the carboxy terminus of DARPP-32 (D32–NES^{proh}–GFP) blocked the nuclear accumulation induced by SKF81297 (Fig. 2d).

To determine how the D1R/cAMP pathway regulated DARPP-32 localization we mutated its major phosphorylation sites (Thr 34, Thr 75, Ser 97 and Ser 130; Fig. 2b) to alanine. In striatal neurons in culture, basal and D1R-induced localization of T34A-D32–GFP, T75A-D32–GFP and S130A-D32–GFP mutants was similar to that of wild-type protein (Supplementary Fig. 6a). In support of these observations, in knock-in mutant mice bearing an alanine point mutation at Thr 34 or Thr 75, the basal localization of DARPP-32 and its cocaine-induced nuclear accumulation was similar to that of the wild type (Supplementary Fig. 6b, c). In contrast, when Ser 97, located close to the NES of DARPP-32, was mutated to alanine, the protein was nuclear in basal conditions and D1R stimulation did not alter this distribution (S97A-D32–GFP; Fig. 3a). Mutation of Ser 97 to the acidic residues glutamate (S97E) or aspartate (S97D), which mimics phosphorylation, induced a preferential cytoplasmic localization, which was not modified by SKF81297 (Fig. 3a and Supplementary Fig. 7a).

Ser 97 (Ser 102 in rat) is highly phosphorylated in basal conditions by casein kinase 2 (CK2)[22]. An inhibitor of CK2, 4,5,6,7-tetrabromobenzotriazole (TBB)[23], markedly increased nuclear D32–GFP and prevented the effects of treatment with D1R agonist (Fig. 3b and Supplementary Fig. 7b). The effects of TBB resulted from the prevention of Ser 97 phosphorylation, because they were

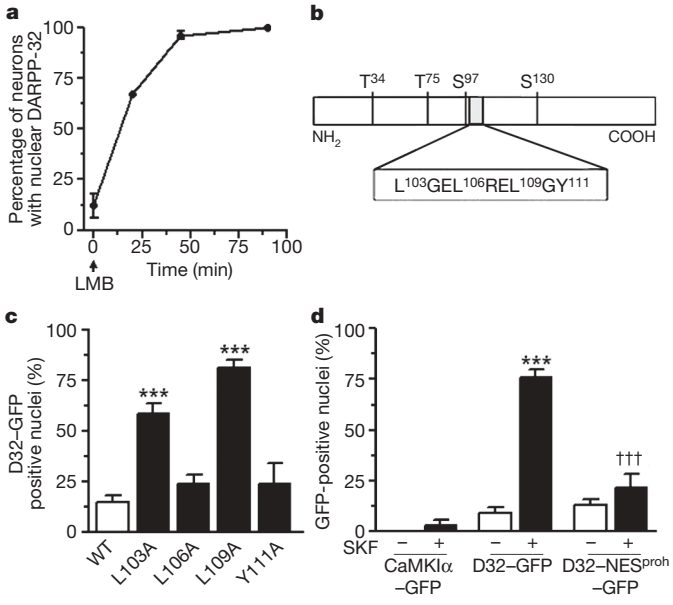


Figure 2 | DARPP-32 undergoes continuous cyto-nuclear shuttling. **a**, Leptomycin B (LMB; 10 ng ml^{-1}) induces nuclear accumulation of DARPP-32 in striatal neurons in culture. **b**, Putative NES in DARPP-32 and major phosphorylation sites. **c**, Mutagenesis of hydrophobic residues in NES increases nuclear localization of D32–GFP in transfected neurons. Three asterisks, $P < 0.001$ compared with wild type (WT). **d**, Stimulation of D1R (10 μM SKF81297 for 15 min) has no effect on the localization of an unrelated protein with a NES (CaMKIα–GFP) or when DARPP-32 contains an additional NES (D32–NES^{proh}–GFP). Three asterisks, $P < 0.001$ for SKF compared with vehicle; three daggers, $P < 0.001$ for mutant compared with WT. Error bars represent s.e.m., $n = 3$–6, one-way analysis of variance (Supplementary Table 1), Bonferroni test.

not observed with S97E-D32–GFP (Fig. 3c and Supplementary Fig. 7c). These results support the hypothesis that DARPP-32 is mainly nuclear when Ser 97 is dephosphorylated, whereas it is preferentially cytoplasmic when phosphorylated by CK2.

Protein phosphatase-2A dephosphorylates Ser 97

Because our results predicted that dephosphorylation of Ser 97 would enhance nuclear DARPP-32, we tested the role of protein phosphatase-2A (PP2A), which dephosphorylates Ser 97 *in vitro*[22]. Pretreatment of striatal neurons with okadaic acid, a potent inhibitor of PP2A, did not alter the basal cytoplasmic localization of D32–GFP, whereas it blocked its D1R-induced nuclear translocation (Fig. 3d and Supplementary Fig. 7d). We next examined the regulation of Ser 97 phosphorylation in striatal slices[24]. SKF81297 induced a marked increase in Thr 34 phosphorylation, as expected, and a dephosphorylation of Ser 97 (Fig. 3e). The precise time course of the two responses was different, with dephosphorylation of Ser 97 being slower and more persistent than Thr 34 phosphorylation. Forskolin, a potent activator of adenylyl cyclase, induced a pronounced and sustained increase in Thr 34 phosphorylation and a slightly delayed, but persistent, decrease in Ser 97 phosphorylation (Fig. 3f and Supplementary Fig. 8). Forskolin-stimulated dephosphorylation of Ser 97 was prevented by okadaic acid (Fig. 3g), at a concentration that blocks PP2A activity fully, and PP1 activity only partly, in striatal slices[25].

PP2A activity is controlled by various regulatory subunits, among which the B56δ subunit is highly expressed in the striatum (Supplementary Fig. 9a) and is activated by cAMP-dependent phosphorylation[26,27]. When B56δ was co-expressed with DARPP-32 in HEK293 cells, it promoted dephosphorylation of Ser 97 in response to forskolin (Fig. 3h and Supplementary Fig. 9b). In contrast, in cells transfected with the vector alone or with another isoform of the B subunit (Bα), forskolin slightly increased Ser 97 phosphorylation (Fig. 3h and Supplementary Fig. 9b). Thus, D1R stimulation and increased cAMP levels induce the dephosphorylation of Ser 97 selectively through stimulation of PP2A containing the B56δ subunit.

Altered behaviour in S97A-DARPP-32 mice

To examine the role of Ser 97 *in vivo*, we used a knock-in mouse line bearing a point mutation of this residue to alanine (S97A). DARPP-32 was nuclear in about half of medium-size spiny neurons in the various striatal regions of the mutant mice (Fig. 4a). Cocaine injection did not alter the nuclear localization of S97A-DARPP-32 (Fig. 4a). The spontaneous activity, habituation and acute locomotor response to cocaine were similar in S97A-DARPP-32 mice and wild-type mice (Supplementary Fig. 10a–c), whereas the acute locomotor response to morphine ($5\,mg\,kg^{-1}$) was decreased in the mutant (Supplementary Fig. 10d). In wild-type mice a second injection of cocaine or morphine, seven days after the first injection, had a much stronger effect on locomotion than the first, showing a robust sensitization (Supplementary Fig. 10e, f) as reported previously[13]. In S97A-DARPP-32 mice, locomotor sensitization to cocaine and morphine, although still present, was less pronounced than in wild-type mice (Fig. 4b and Supplementary Fig. 10e, f). Moreover, the rewarding effects of cocaine, evaluated by conditioned place preference, were not observed in S97A-DARPP-32 mice (Fig. 4c).

We next examined whether mutation of Ser 97 interfered with responses to physiological rewarding stimuli by using a model in

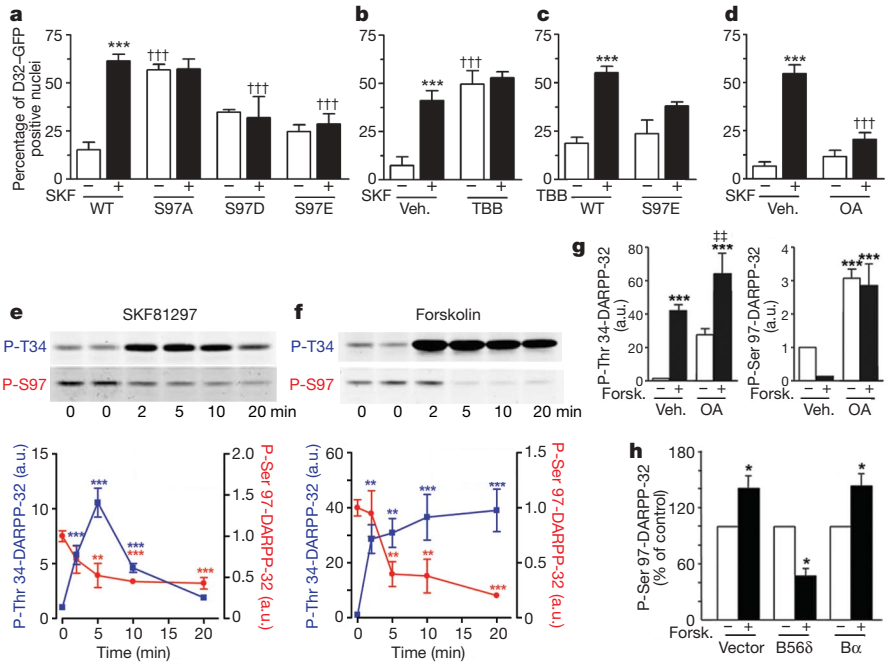


Figure 3 | Phosphorylation at Ser 97 controls intracellular localization of DARPP-32. **a**, Striatal neurons transfected with D1R and wild-type (WT), S97A-D32–GFP, S97D-D32–GFP or S97E-D32–GFP, treated with vehicle (Veh., −) or 10 μM SKF81297 (SKF, +, 15 min). Three asterisks, $P < 0.001$ for SKF compared with vehicle; two daggers, $P < 0.01$; three daggers, $P < 0.001$ for mutant compared with WT. **b**, CK2 inhibition (TBB, 50 μM, 45 min before SKF) increases nuclear DARPP-32. Three asterisks, $P < 0.001$ for SKF compared with vehicle; three daggers, $P < 0.001$ for TBB compared with vehicle. **c**, S97E mutation prevents TBB effects. Three asterisks, $P < 0.001$ for TBB compared with vehicle. **d**, Okadaic acid (OA; 500 nM, 45 min before SKF) prevents D1R-induced DARPP-32 nuclear accumulation. Three asterisks, $P < 0.001$ for SKF compared with vehicle; three daggers, $P < 0.001$ for OA compared with vehicle. **e**, SKF81297 (10 μM) induces Thr 34 phosphorylation (blue) and Ser 97 dephosphorylation (red) in mouse striatal slices (immunoblotting; $n = 5$–7). Two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. **f**, Forskolin (10 μM) also induces Thr 34 phosphorylation (blue) and Ser 97 dephosphorylation (red) ($n = 5$–7). Two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. **g**, Okadaic acid (1 μM) prevents this forskolin-induced Ser 97 dephosphorylation. Three asterisks, $P < 0.001$ compared with control; two double daggers, $P < 0.01$ compared with OA. **h**, B56δ allows Ser 97 dephosphorylation by forskolin (Forsk., 10 μM, 10 min): HEK293 cells transfected with DARPP-32 and vector, B56δ-PP2A or Bα-PP2A subunit (immunoblotting, $n = 3$, *t*-test). Error bars represent s.e.m., one-way ANOVA (Supplementary Table 1) and Bonferroni test (unless otherwise indicated). Asterisk, $P < 0.05$.

which mildly food-deprived mice learned to obtain food pellets by nose-poking. Wild-type and mutant mice learned equally well, even when the number of nose-pokes necessary to obtain a pellet was increased to a fixed ratio of five or when the rewarded hole was reversed (Supplementary Fig. 10g). However, when the required number of nose-pokes was progressively increased until the mice interrupted their behaviour, the breaking point was lower in S97A mutant mice than in wild-type mice (Fig. 4d). Taken together, these results show that mutation of Ser 97 alters long-lasting responses to drugs of abuse, and decreases motivation for food reward.

In S97A-DARPP-32 mice we found a decreased phosphorylation of DARPP-32 on Thr 34 and of several proteins regulated through DARPP-32 (data not shown). We previously observed, with purified bovine DARPP-32, that phosphorylation by CK2 facilitated the phosphorylation of Thr 34 by PKA[22]. However, there was no deficit in cAMP-induced Thr 34 phosphorylation of S97A mutant protein in transfected COS7 cells in which overexpressed DARPP-32 was cytoplasmic (Supplementary Fig. 11). We therefore conclude that the observed alterations in S97A mice probably resulted from the blockade of cyto-nuclear shuttling of DARPP-32, rather than from an intrinsic effect of Ser 97 phosphorylation on Thr 34 phosphorylation.

Nuclear DARPP-32 regulates histone H3 phosphorylation

Because D1R stimulation induces the accumulation of Thr 34-phospho-DARPP-32 (Fig. 1a), a potent inhibitor of PP1, we proposed that DARPP-32 translocation might regulate the phosphorylation of nuclear proteins. We therefore investigated the phosphorylation of histone H3 on Ser 10, a substrate for several kinases[28,29], which is known to be regulated by PP1 (refs 30, 31). The phosphorylation of H3 on Ser 10 is a key step in nucleosomal response[28,32] and is crucial for memory formation[33]. It is increased in the striatum after cocaine administration *in vivo*, and this response is functionally important[14,34]. Cocaine induced an intense phospho-Ser 10-H3 signal in the nuclei of several striatal neurons in wild-type mice, whereas this effect was absent in both T34A-DARPP-32 and S97A-DARPP-32 mice (Fig. 5a). Similar results were observed with an antibody that detected phospho-acetyl-H3 (phospho-Ser 10-acetyl-Lys 14-H3; data not shown).

In striatal neurons transfected with D1R and D32–GFP, the D1R agonist SKF81297 markedly increased the phosphorylation of H3 on Ser 10 (Fig. 5b, c, and Supplementary Fig. 12). SKF81297 did not alter the acetylation of H3 on Lys 14, but increased the immunoreactivity of phospho-Ser 10, acetyl-Lys 14 H3 (Fig. 5b and Supplementary Fig. 12). D1R-induced phosphorylation of histone H3 was lost when neurons were transfected with T34A-D32–GFP instead of the wild type (Fig. 5c), confirming that the ability of DARPP-32 to inhibit PP1 was essential for H3 phosphorylation. We then directly evaluated the role of DARPP-32 translocation to the nucleus by using mutant forms of D32–GFP that are constitutively excluded from the nucleus: S97E-D32–GFP (see Fig. 3a) and DARPP-32-NES^{proh}–GFP (see Fig. 2d). In neurons transfected with these 'cytoplasmic' variants of D32–GFP, the D1R agonist SKF81297 did not induce H3 phosphorylation (S97E and NES^{proh}; Fig. 5c). These results showed that DARPP-32 phosphorylation on Thr 34, and thus PP1 inhibition, was essential for D1R-induced phosphorylation of Ser 10-H3 and that this effect required the ability of DARPP-32 to accumulate in the nucleus in response to stimulation of D1R.

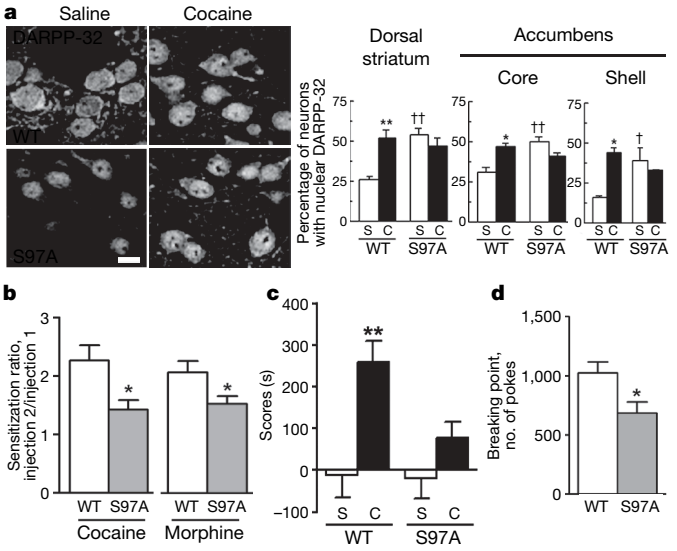


Figure 4 | Mutation of Ser 97 alters localization of DARPP-32, behavioural responses to drugs of abuse and motivation for food reward. a, DARPP-32 immunoreactivity (dorsal striatum) in wild-type (WT) or S97A-DARPP-32 mice injected with saline (S) or cocaine (C, 20 mg kg^{-1}, for 10 min). Single confocal sections. Scale bar, 10 μm; $n = 3$–8. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ for saline compared with cocaine; dagger, $P < 0.05$; two daggers, $P < 0.01$ for S97A compared with WT. **b**, Decreased locomotor sensitization to a second injection (day 7) of cocaine (20 mg kg^{-1}, for 1 h; $n = 16$) or morphine (5 mg kg^{-1}, for 3 h; $n = 8$) in S97A mice (*t*-test). Asterisk, $P < 0.05$. **c**, Conditioned place preference to cocaine was prevented in S97A mice (scores are calculated as time in cocaine-paired compartment after conditioning minus that before conditioning; $n = 7$–8). Two asterisks, $P < 0.01$. **d**, Decreased breaking point for food reward under progressive ratio schedule in S97A mice (*t*-test, $n = 7$–8). Error bars represent s.e.m., two-way analysis of variance (Supplementary Table 1) and Bonferroni test (unless otherwise indicated). Asterisk, $P < 0.05$.

Discussion

Our study identifies nuclear DARPP-32 as a key factor in the phosphorylation of histone H3, a component of the nucleosomal response that is essential for gene expression[28,32,33]. Psychostimulant drugs (D-amphetamine and cocaine) and morphine, which act by completely different mechanisms, share the ability to increase extracellular dopamine in the NAc and to trigger the nuclear accumulation of DARPP-32. A natural stimulus such as a simple food-reinforced learning model caused a similar effect. DARPP-32 undergoes a very active, continuous cyto-nuclear shuttling regulated by the phosphorylation of Ser 97, which is in the vicinity of its NES (summarized in Fig. 5d). Phosphorylation of Ser 97 by CK2 seems to be crucial for the nuclear export of DARPP-32. CK2 is present in nuclei of striatal neurons (Supplementary Fig. 13) and can phosphorylate nuclear DARPP-32 and promote its export. Stimulation of D1R triggers the rapid phosphorylation of Thr 34 responsible for cytoplasmic effects of DARPP-32, and the slower dephosphorylation of Ser 97, by activating PP2A through cAMP/PKA-mediated phosphorylation of its B56δ subunit[26,27], thus decreasing the nuclear export of DARPP-32. Although phosphorylation has been reported to regulate the cyto-nuclear shuttling of several proteins[35], this is a rare example of facilitation of CRM1-mediated nuclear export by phosphorylation[36,37]. cAMP-independent mechanisms also couple dopamine D2 receptor (D2R) to a delayed activation of PP2A (ref. 38). D2R does not seem to be important in the early nuclear accumulation of DARPP-32, but it could have a regulatory role on Ser 97 phosphorylation and DARPP-32 localization in other circumstances.

A large body of evidence demonstrates that long-term synaptic plasticity requires the control of gene expression in the striatum as in other brain regions[39]. We have identified a previously unrecognized mechanism by which dopamine controls chromatin through the regulated translocation of a PP1 inhibitor to the nucleus, which promotes histone H3 phosphorylation. Dopamine-controlled inhibition of nuclear PP1, a wide-spectrum protein phosphatase, is likely to have other targets, making it a general means for the control of nuclear function. This mechanism is undoubtedly important in the long-term effects of drugs of abuse, which activate gene transcription through the stimulation of D1R, and in physiological reward-controlled learning. Regulation of the nuclear accumulation of

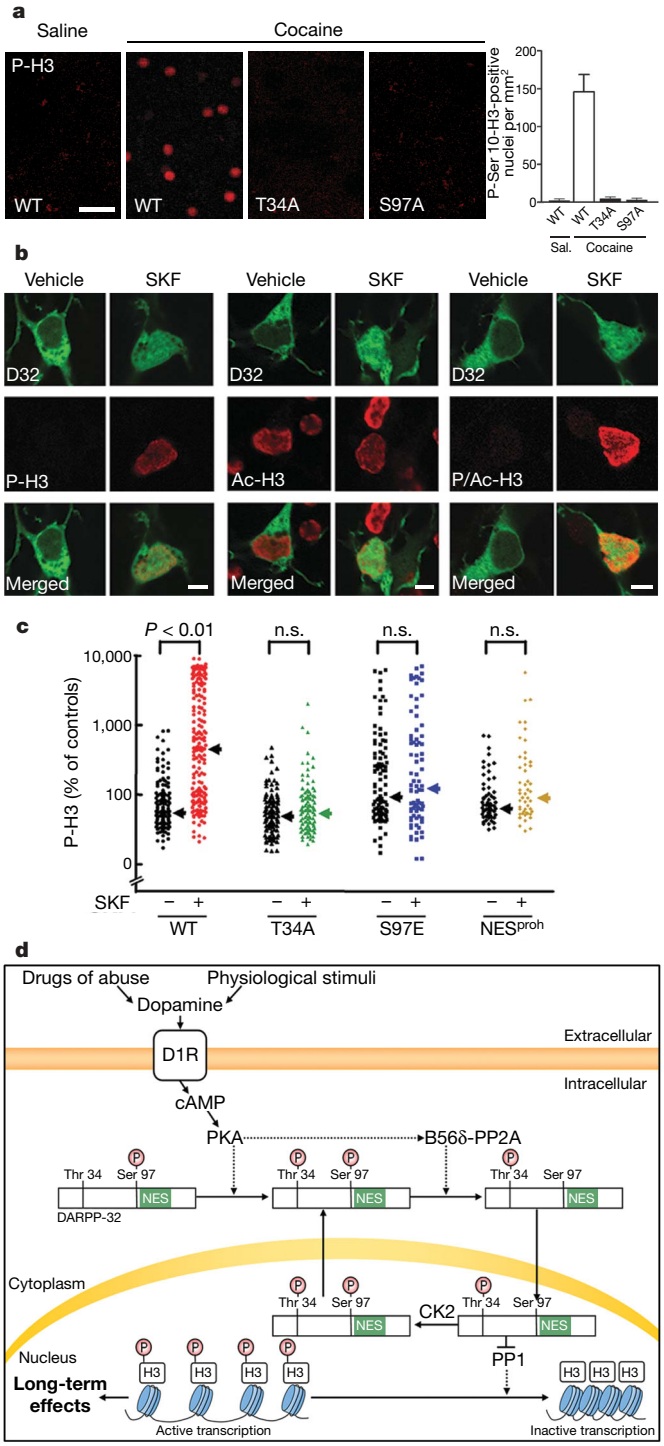


Figure 5 | Role of nuclear DARPP-32 in the phosphorylation of histone H3 in striatal neurons. **a**, Phospho-Ser 10-H3 (P-H3) immunofluorescence induced by cocaine (20 mg kg^{-1}, for 30 min) in wild-type (WT) dorsal striatum is abolished in T34A or S97A DARPP-32 mutant mice. Scale bar, 40 μm. **b**, Stimulation of D1R (SKF81297 (SKF), 10 μM, for 15 min) increases Ser 10-H3 phosphorylation in striatal neurons transfected with D32–GFP and D1R. Confocal section, GFP fluorescence, P-H3, acetyl-Lys 14-H3 (Ac-H3) and phospho-Ser 10-acetyl-Lys 14-H3 (P/Ac-H3) immunofluorescence. Scale bars, 5 μm. **c**, Phosphorylation of H3 requires DARPP-32 Thr 34 phosphorylation and nuclear accumulation. Striatal neurons transfected with WT, T34A-D32–GFP, S97E-D32–GFP or D32–NES^{proh}–GFP, treated with SKF81297. Arrows indicate medians. Kruskal–Wallis ($P < 0.0001$) and Dunn's tests. n.s., not significant. **d**, Working model of regulation of nucleosomal response in striatal neurons by DARPP-32.

DARPP-32 may also be important to the functions or dysfunctions of dopamine that involve transcription-dependent long-term plasticity, including neuroleptic- and L-DOPA-induced dyskinesia, and other diseases of the basal ganglia.

METHODS SUMMARY

Knock-in T34A-DARPP-32, T75A-DARPP-32 and S97A-DARPP-32 (refs 40, 41), D1R-deficient[42], *drd1a*-EGFP[17] and matched wild-type control mice were habituated to saline injections for three days before experiments. Drug-treated mice were anaesthetized with pentobarbital and received an intracardiac perfusion of 4% paraformaldehyde. Vibratome-cut brain sections were processed for immunofluorescence[13]. Primary striatal neurons, from 14-day embryonic Swiss mice, were grown in supplemented Neurobasal medium and transfected a week later with plasmids expressing D1R and DARPP-32 fused with C-terminal EGFP (D32–GFP). HEK293, CHO-K1 or COS-7 cells were transfected with DARPP-32 constructs associated or not with PP2A subunits, and studied 24–48 h later. On the basis of confocal single sections, the subcellular localization of DARPP-32 or GFP was classified into three categories: nuclear staining stronger than, equal to or weaker than cytoplasmic. For cells in culture, the first two categories were pooled and considered as cells with nuclear staining. Phosphorylated and/or acetylated H3 immunofluorescence was quantified with Image-J-1.34s (National Institutes of Health). Striatal coronal slices (350 μm) were incubated in oxygenated artificial cerebrospinal fluid[25], and phospho-DARPP-32 was analysed in slices or transfected cells by western blotting[25]. Locomotor activity, sensitization produced by a single injection of drug, and conditioned place preference were measured as described[13,14,43]. In the operant model for food self-administration (incentive learning), mildly food-deprived mice were placed in an operant chamber with two nose-poke holes. One or five nose-pokes (fixed ratios 1 and 5) in the rewarded hole, indicated by a cue light, triggered the release of a food pellet. Controls were 'no food' mice (mice in the same chamber but with no pellet delivered) and 'yoked' mice (paired to active mice, in exactly the same conditions, but rewarded when the active mouse nose-poked). For the reversal test, the previously inactive hole became active and vice versa. For analysis of the progressive ratio, the number of successive pokes required for pellet release was progressively increased until the mouse stopped poking (breaking point).

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

Received 20 January; accepted 4 April 2008.
Published online 21 May 2008.

1. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500 (2000).
2. Berke, J. D. & Hyman, S. E. Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**, 515–532 (2000).
3. Di Chiara, G. Drug addiction as dopamine-dependent associative learning disorder. *Eur. J. Pharmacol.* **375**, 13–30 (1999).
4. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neurosci.* **8**, 1481–1489 (2005).
5. Nicola, S. M., Surmeier, J. & Malenka, R. C. Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu. Rev. Neurosci.* **23**, 185–215 (2000).
6. Reynolds, J. N. & Wickens, J. R. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.* **15**, 507–521 (2002).
7. Hyman, S. E., Malenka, R. C. & Nestler, E. J. Neural mechanisms of addiction: the role of reward-related learning and memory. *Annu. Rev. Neurosci.* **29**, 565–598 (2006).
8. Walaas, S. I., Aswad, D. W. & Greengard, P. A dopamine- and cyclic AMP-regulated phosphoprotein enriched in dopamine-innervated brain regions. *Nature* **301**, 69–71 (1983).
9. Svenningsson, P. *et al.* DARPP-32: an integrator of neurotransmission. *Annu. Rev. Pharmacol. Toxicol.* **44**, 269–296 (2004).
10. Ouimet, C. C. *et al.* DARPP-32, a dopamine- and adenosine 3′:5′-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. III. Immunocytochemical localization. *J. Neurosci.* **4**, 111–124 (1984).
11. Hemmings, H. C. Jr, Greengard, P., Tung, H. Y. L. & Cohen, P. DARPP-32, a dopamine-regulated neuronal phosphoprotein, is a potent inhibitor of protein phosphatase-1. *Nature* **310**, 503–505 (1984).
12. Fienberg, A. A. *et al.* DARPP-32: Regulator of the efficacy of dopaminergic neurotransmission. *Science* **281**, 838–839 (1998).
13. Valjent, E. *et al.* Regulation of a protein phosphatase cascade allows convergent dopamine and glutamate signals to activate ERK in the striatum. *Proc. Natl Acad. Sci. USA* **102**, 491–496 (2005).
14. Brami-Cherrier, K. *et al.* Parsing molecular and behavioral effects of cocaine in mitogen- and stress-activated protein kinase-1-deficient mice. *J. Neurosci.* **25**, 11444–11454 (2005).

15. Ouimet, C. C. & Greengard, P. Distribution of DARPP-32 in the basal ganglia: An electron microscopic study. *J. Neurocytol.* **19,** 39–52 (1990).
16. Pontieri, F. E., Tanda, G. & Di Chiara, G. Intravenous cocaine, morphine, and amphetamine preferentially increase extracellular dopamine in the 'shell' as compared with the 'core' of the rat nucleus accumbens. *Proc. Natl Acad. Sci. USA* **92,** 12304–12308 (1995).
17. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425,** 917–925 (2003).
18. Nishi, K. *et al.* Leptomycin B targets a regulatory cascade of crm1, a fission yeast nuclear protein, involved in control of higher order chromosome structure and gene expression. *J. Biol. Chem.* **269,** 6320–6324 (1994).
19. Henderson, B. R. & Eleftheriou, A. A comparison of the activity, sequence specificity, and CRM1-dependence of different nuclear export signals. *Exp. Cell Res.* **256,** 213–224 (2000).
20. Stedman, D. R. *et al.* Cytoplasmic localization of calcium/calmodulin-dependent protein kinase I-α depends on a nuclear export signal in its regulatory domain. *FEBS Lett.* **566,** 275–280 (2004).
21. Rastogi, S., Joshi, B., Fusaro, G. & Chellappan, S. Camptothecin induces nuclear export of prohibitin preferentially in transformed cells through a CRM-1-dependent mechanism. *J. Biol. Chem.* **281,** 2951–2959 (2006).
22. Girault, J. A. *et al.* Phosphorylation of DARPP-32, a dopamine- and cAMP-regulated phosphoprotein, by casein kinase II. *J. Biol. Chem.* **264,** 21748–21759 (1989).
23. Sarno, S. *et al.* Selectivity of 4,5,6,7-tetrabromobenzotriazole, an ATP site-directed inhibitor of protein kinase CK2 ('casein kinase-2'). *FEBS Lett.* **496,** 44–48 (2001).
24. Nishi, A. *et al.* Amplification of dopaminergic signaling by a positive feedback loop. *Proc. Natl Acad. Sci. USA* **97,** 12840–12845 (2000).
25. Nishi, A., Snyder, G. L., Nairn, A. C. & Greengard, P. Role of calcineurin and protein phosphatase-2A in the regulation of DARPP-32 dephosphorylation in neostriatal neurons. *J. Neurochem.* **72,** 2015–2021 (1999).
26. Usui, H. *et al.* Activation of protein phosphatase 2A by cAMP-dependent protein kinase-catalyzed phosphorylation of the 74-kDa B″ (δ) regulatory subunit *in vitro* and identification of the phosphorylation sites. *FEBS Lett.* **430,** 312–316 (1998).
27. Ahn, J. H. *et al.* Protein kinase A activates protein phosphatase 2A by phosphorylation of the B56δ subunit. *Proc. Natl Acad. Sci. USA* **104,** 2979–2984 (2007).
28. Nowak, S. J. & Corces, V. G. Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet.* **20,** 214–220 (2004).
29. Salvador, L. M. *et al.* Follicle-stimulating hormone stimulates protein kinase A-mediated histone H3 phosphorylation and acetylation leading to select gene activation in ovarian granulosa cells. *J. Biol. Chem.* **276,** 40146–40155 (2001).
30. Murnion, M. E. *et al.* Chromatin-associated protein phosphatase 1 regulates aurora-B and histone H3 phosphorylation. *J. Biol. Chem.* **276,** 26656–26665 (2001).
31. Hsu, J. Y. *et al.* Mitotic phosphorylation of histone H3 is governed by Ipl1/aurora kinase and Glc7/PP1 phosphatase in budding yeast and nematodes. *Cell* **102,** 279–291 (2000).
32. Bode, A. M. & Dong, Z. G. Inducible covalent postranslational modification of histone H3. *Sci. STKE* **281,** 1–12 (2005).
33. Levenson, J. M. & Sweatt, J. D. Epigenetic mechanisms: a common theme in vertebrate and invertebrate memory formation. *Cell. Mol. Life Sci.* **63,** 1009–1016 (2006).
34. Kumar, A. *et al.* Chromatin remodeling is a key mechanism underlying cocaine-induced plasticity in striatum. *Neuron* **48,** 303–314 (2005).
35. Poon, I. K. & Jans, D. A. Regulation of nuclear transport: central role in development and transformation? *Traffic* **6,** 173–186 (2005).
36. Panasyuk, G. *et al.* Nuclear export of S6K1 II is regulated by protein kinase CK2 phosphorylation at Ser-17. *J. Biol. Chem.* **281,** 31188–31201 (2006).
37. Alt, J. R., Cleveland, J. L., Hannink, M. & Diehl, J. A. Phosphorylation-dependent regulation of cyclin D1 nuclear export and cyclin D1-dependent cellular transformation. *Genes Dev.* **14,** 3102–3114 (2000).
38. Beaulieu, J. M. *et al.* An Akt/β-arrestin 2/PP2A signaling complex mediates dopaminergic neurotransmission and behavior. *Cell* **122,** 261–273 (2005).
39. Nestler, E. J. Molecular basis of long-term plasticity underlying addiction. *Nature Rev. Neurosci.* **2,** 119–128 (2001).
40. Svenningsson, P. *et al.* Diverse psychotomimetics act through a common signaling pathway. *Science* **302,** 1412–1415 (2003).
41. Zhang, Y. *et al.* Cocaine self-administration in mice is inversely related to phosphorylation at Thr34 (protein kinase A site) and Ser130 (kinase CK1 site) of DARPP-32. *J. Neurosci.* **26,** 2645–2651 (2006).
42. Drago, J. *et al.* Altered striatal function in a mutant mouse lacking D1A dopamine receptors. *Proc. Natl Acad. Sci. USA* **91,** 12564–12568 (1994).
43. Vanderschuren, L. J. *et al.* A single exposure to amphetamine is sufficient to induce long-term behavioral, neuroendocrine, and neurochemical sensitization in rats. *J. Neurosci.* **19,** 9579–9586 (1999).
44. Valjent, E. *et al.* Involvement of the extracellular signal-regulated kinase cascade for cocaine-rewarding properties. *J. Neurosci.* **20,** 8701–8709 (2000).

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

Acknowledgements We thank M. Lambert for her help with time-lapse video; P. Ingrassia and P. Bernard for their help with mutant mice; and M. R. Picciotto, S. Cottecchia, J. P. Hornung, R. Luedtke, M. Takeda and Intracellular Therapies Inc. for reagents. This work was supported by Inserm, and by grants from Agence Nationale de la Recherche (05-NEUR-020-01), Fondation Bettencourt-Schueller (Coup d'élan) and Association pour la Recherche contre le Cancer (ARC-3118 and -7905) to J.A.G., from Fondation pour la Recherche Médicale (FRM) to D.H., a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science to A.N., and grants from the National Institute on Drug Abuse (DA10044), the National Institute of Mental Health (MH74866), the US Department of Defense (W81XWH-05-1-0146), the Picower Foundation, the Michael Stern Parkinson's Research Foundation and the US Army Medical Research Acquisition Activity (DAMD17-02-1-0705 and W81XWH-05-1-0146) to P.G. and A.C.N. A.S. was supported by Mission Interministérielle de Lutte contre la Drogue et la Toxicomanie and FRM, and J.B.G. by FRM.

Author Contributions A.S. and Mi.M. performed experiments *in vivo* and in transfected cultures, immunofluorescence and molecular biology. E.V. conducted *in vivo*, behavioural and immunohistofluorescence experiments. A.S., E.V. and Mi.M. prepared the figures. A.N. performed the slice experiments, J.H.A. the phosphatase experiments, and Ma.M. the incentive learning experiments. J.B.G. and A.G.C. contributed to *in vivo* and immunohistofluorescence experiments, and K.B.C. and H.E. to cell culture experiments. O.F. provided advice and reagents. A.S., A.C.N., P.G., D.H. and J.A.G. were involved in the study design and manuscript writing. J.A.G. coordinated the study. All authors analysed data they generated, discussed results and commented on the manuscript. A.S., E.V. and Mi.M. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J-A.G. (girault@fer-a-moulin.inserm.fr).

METHODS

Animals. Male eight-week-old C57BL/6J mice from Elevage Janvier or Japan-SLC were used for *in vivo* and slice experiments, respectively. Production of knock-in T34A-DARPP-32, T75A-DARPP-32 and S97A-DARPP-32 mutant[40,41], D1R-deficient[42] and *drd1a*-EGFP[17] mice were described previously. Mutant and wild-type mice used in experiments, or their parents, were littermates. Mice were kept for at least one week in a local animal house in stable conditions (22 °C, 12 h/12 h light/dark cycle, free access to food and water). During the three days preceding the experiments, mice were habituated to behavioural apparatus, when applicable, and to injections by daily intraperitoneal saline administration. Animal care was in accordance with ethical guidelines (Declaration of Helsinki and NIH, publication no. 85-23, revised 1985, European Community Guidelines, and French Agriculture and Forestry Ministry guidelines for handling animals, decree 87849, license A 75-05-22) and approved by the local ethical committees.

Drugs. (+)-D-Methylphenethylamine (D-amphetamine) sulphate salt, cocaine-HCl, SKF81297, SCH23390 (both in 0.1% v/v dimethylsulphoxide) and okadaic acid potassium salt were from Sigma-Aldrich or Tocris; morphine sulphate salt was from Francopia; 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole-3′,5′-cyclic-monophosphoro-thioate, Sp-isomer (Sp-5,6-DCl-cBIMPS) and adenosine-3′,5′-cyclic-mono-phosphorothioate, Rp-isomer (Rp-cAMPS) were from Biolog; leptomycin-B was from Amersham; 4,5,6,7-tetrabromobenzotriazole (50 μM in 0.1% dimethylsulphoxide) was from Calbiochem. For *in vivo* experiments, drugs were dissolved in 0.9% NaCl and injected intraperitoneally (i.p.) except morphine, which was injected subcutaneously (s.c.).

Plasmid constructs and DARPP-32 mutants. *Rattus norvegicus* DARPP-32 cDNA (mouse residue numbering is used throughout) was amplified by PCR with Pfu (Stratagene) and cloned in the *Xho*I/*Eco*RI sites of EGFP-N2 (Clontech), and subjected to site-directed mutagenesis (Quick-Change; Stratagene). DARPP-32-NES^{proh} was constructed by *Eco*RI/*Sal*I digestion and ligation of a sequence encoding AAEDIAYQLSRSRNITYLPAGQSVLLQLPQ (NES of prohibitin; ref. 21) to the C terminus of DARPP-32. Constructs were verified by DNA sequencing. CaMKIα–GFP and pRK5-D1R were gifts.

Cell culture and transfection. Striatal neurons (180,000 per well) from 14-day embryonic Swiss mice (Janvier) were transfected after seven days in culture[14] with D32–GFP (1 μg) and pRK5-D1R (1 μg) DNA, using Lipofectamine 2000 in OptiMEM serum-free medium. Treatments were done 24 h later, in fresh Neurobasal medium. CHO-K1 cells (American Type Culture Collection) (25,000 cells per 14-mm diameter coverslip) were cultured in Ham's F-12 medium with 10% foetal bovine serum (FBS). COS-7 and HEK293 cells were cultured in six-well plates (10^6 per well) in DMEM medium with 10% FBS. Culture media were from Invitrogen. CHO-K1 and COS-7 cells were transfected as neurons with D32–GFP DNA (1 μg). HEK293 cells were co-transfected with Myc-tagged DARPP-32 (0.5 μg) and Flag-tagged B-subunits of PP2A vector, Bα or B56δ (0.5 μg) DNA, using Fugene-6 (Roche). After 48 h, cells were treated with forskolin and lysed in 150 μl of 50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% Triton X-100, 0.1% SDS, protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail I and II (Calbiochem), sonicated for 5 s, and centrifuged for 10 min at 16,000*g*.

Immunostaining of striatal neurons in culture. Cells were fixed with PBS containing 2% paraformaldehyde (40 min, 20–23 °C) and incubated with methanol/acetone (1:1, 10 min, 4 °C). After three rinses in PBS, cells were treated with blocking buffer (3% BSA in PBS except for H3 immunofluorescence: 1% BSA and 1% FBS) for 45 min (2 h for P-Ser 10-H3) at room temperature. Coverslips were incubated overnight at 4 °C in PBS with 1% BSA and DARPP-32 (1:4,000 dilution) and D1R (1:500 dilution) monoclonal antibodies. P-Ser 10-H3, acetyl-Lys 14-H3 and P-Ser 10-acetyl-Lys 14-H3 antibodies (1:1,000 dilution; Upstate) were added overnight at 4 °C in PBS containing 1% BSA and 0.05% Tween 20. Anti-mouse Alexa488-conjugated antibody or anti-mouse Cy3-conjugated antibody (1:400 dilution; Molecular Probes) were added for 1 h at room temperature. Nuclei were counterstained with SYTOX (1:2,000 dilution; Molecular Probes). Cells rinsed three times were mounted under coverslips with Vectashield (Vector Laboratories).

Immunohistofluorescence of brain sections and subcellular distribution of DARPP-32. Brains from paraformaldehyde-perfused mice were postfixed and 30-μm sections were cut with a vibratome (Leica)[13,44]. Free-floating sections were incubated overnight with antibodies at 4 °C, and after rinses in TBS, for 2 h at room temperature with secondary antibody (1:400 dilution; Cy3-coupled anti-rabbit IgG or Alexa488-conjugated anti-mouse antibodies; Molecular Probes). Nuclei were counterstained with 1 μM TO-PRO-3 iodide (Molecular Probes). Sections were rinsed three times in TBS and mounted in Vectashield.

Image acquisition and statistical analysis. Image acquisition with sequential laser scanning confocal microscopy (SP2; Leica) was performed at the Institut du Fer à Moulin Imaging Facility. For brain immunofluorescence, two sections were analysed bilaterally and cells were counted in one field for dorsal striatum, NAc shell and core (data points for an animal correspond to four acquisitions per structure). For cultures, data points correspond to a minimum of 15 transfected cells (minimum of two different experiments with three independent transfections). Data were analysed with Prism (GraphPad) software, using parametric or non-parametric statistics as indicated, depending on the normality of the distribution of variables.

Behavioural analysis. Locomotor activity was measured in a circular corridor (Imetronic)[13]. Mice were placed in the corridor 30 min before injection for the acute experiments. Locomotor sensitization was induced by a single injection of cocaine or morphine[13,43]. Food self-administration experiments were conducted in operant chambers (Imetronic) with two holes, one being the active hole, indicated by a cue light. Nose-poking in the active hole resulted in delivery of a food pellet (Phymep). The light was switched off for 10 s after the consumption of the food pellet, during which no reward was provided. Mice were food-deprived (3.5 g/d, water *ad libitum*) starting five days before the experiment, to reach 85% of initial weight. Self-administration sessions (1 h per day) were conducted five days per week. Mice were trained under a fixed ratio 1 (FR1, 1 pellet per nose poke) schedule. The session was terminated after 100 pellets or 1 h. Acquisition criteria were: stable response with less than 20% deviation from the mean of the total number of reinforcers earned in three consecutive sessions (80% stability). Criteria were then changed to a fixed ratio 5 (FR5, 1 pellet per 5 nose pokes) schedule. When the 80% stability criterion had been reached, a reversal procedure was started (rFR5). During reversal the previously active hole became inactive, and vice versa. When the same criteria as above had been met, mice were subjected to a single 2-h progressive ratio (PR) session in which the requirement to earn a pellet escalated according to the following series: 1, 2, 3, 5, 12, 18, 27, 40, 60, 90, 135, 200, 300, 450, 675, 1,000, 1,500.

ARTICLES

Structural basis for the regulated protease and chaperone function of DegP

Tobias Krojer[1]*, Justyna Sawa[1]*, Eva Schäfer[2]*, Helen R. Saibil[2], Michael Ehrmann[3] & Tim Clausen[1]

All organisms have to monitor the folding state of cellular proteins precisely. The heat-shock protein DegP is a protein quality control factor in the bacterial envelope that is involved in eliminating misfolded proteins and in the biogenesis of outer-membrane proteins. Here we describe the molecular mechanisms underlying the regulated protease and chaperone function of DegP from *Escherichia coli*. We show that binding of misfolded proteins transforms hexameric DegP into large, catalytically active 12-meric and 24-meric multimers. A structural analysis of these particles revealed that DegP represents a protein packaging device whose central compartment is adaptable to the size and concentration of substrate. Moreover, the inner cavity serves antagonistic functions. Whereas the encapsulation of folded protomers of outer-membrane proteins is protective and might allow safe transit through the periplasm, misfolded proteins are eliminated in the molecular reaction chamber. Oligomer reassembly and concomitant activation on substrate binding may also be critical in regulating other HtrA proteases implicated in protein-folding diseases.

All living organisms employ dedicated chaperones and proteases to monitor and control the state of cellular proteins. Failure of this quality control can lead to protein aggregation, a malfunction that is correlated with fatal protein-folding diseases[1,2]. The protease–chaperone DegP is a unique model system for uncovering mechanisms that protect cells from misfolded or damaged proteins, because it combines digestive and remodelling activities on a single polypeptide and can switch between these dual functions in a tightly regulated manner[3–5]. DegP is a member of the widely conserved HtrA family of serine proteases that are crucial to maintain protein homeostasis in extracytoplasmic compartments[6]. The bacterial representatives DegP and DegS have key functions in the unfolded protein response of the cell envelope, whereas the four human HtrA proteins are implicated in many severe disorders including Parkinson's and Alzheimer's diseases[7,8]. HtrA proteins encompass a catalytic domain with a chymotrypsin-like fold and one or two carboxy-terminal PDZ domains, which are well-characterized protein–protein interaction modules[9]. The protease domains of three protomers interact closely to form a trimer that is the basic building block of HtrA oligomers. The projecting PDZ domains either participate in protein degradation by presenting substrates to the protease[4] or offer a binding site for an allosteric activator that stimulates protease function[10,11]. The reversible activation mechanism ensures that the digestive mode of HtrA proteins can be precisely switched on and off depending on the needs of the cell. The available crystal structures suggest that HtrA proteins differ in their molecular architecture, ranging from trimers with surface-accessible active sites to hexamers that belong to the class of self-compartmentalizing proteases[6]. For these proteases, trimer association positions a regulatory loop in the active site of a neighbouring molecule, thereby blocking substrate access and deforming the proteolytic site[12]. In addition to its housekeeping function, DegP is also involved in the biogenesis of outer-membrane proteins (OMPs)[13–16]. OMPs are translocated as unfolded polypeptide chains across the cytoplasmic membrane by means of the general SecYEG secretion complex[17]. In the periplasm, OMPs are targeted to a translocation machinery in the outer membrane composed of the integral OMP YaeT and the four lipoproteins NlpB, SmpA, YfgL and YfiO[18]. Because partly folded OMPs would be substrates for various periplasmic proteases, cells must ensure a safe transit of OMP precursors between inner and outer membrane. Moreover, unfolded OMPs could be prone to protein aggregation and would continuously stimulate the σE stress response[19]. So far, the three chaperones SurA, Skp and DegP have been implicated in guiding OMPs through the periplasm[20–23]. However, their exact contributions remain to be fully understood.

For a better understanding of how a single cellular factor selectively binds unstructured proteins and then decides whether a substrate will be degraded, repaired or transported to its ultimate cellular destination, we sought to characterize *in vivo* substrates of DegP and tested which proteins co-purify with the proteolytically inactive $DegP_{S210A}$ (Supplementary Fig. 1). Size-exclusion chromatography (SEC) led to the identification of three DegP oligomers, namely the 6-mer ($DegP_6$), the 12-mer ($DegP_{12}$) and the 24-mer ($DegP_{24}$), of which the two larger particles had additional proteins bound (Fig. 1a). Analysis of solubilized crystals of the $DegP_{24}$ complex revealed that the co-purified and co-crystallized proteins were the OMPs OmpA, OmpC, OmpF and LamB.

Crystal structure of $DegP_{24}$

The crystal structure of the $DegP_{24}$ complex was solved by the single-wavelength anomalous dispersion method and refined to an *R*-factor of 21.2% at 3.0 Å resolution (R_{free} 27.4%; Supplementary Table 1). In contrast to the previously solved hexameric structure of DegP, both the protease domain and the PDZ1 and PDZ2 domains are well defined by electron density and show good stereochemistry. Only one protease loop (residues 36–81) was too flexible to be traced in the electron density. The co-crystallized OMPs were also not defined by electron density, presumably as a result of conformational and chemical heterogeneity.

[1]Research Institute for Molecular Pathology – IMP, Dr Bohrgasse 7, A-1030 Vienna, Austria. [2]Crystallography Department and Institute of Structural Molecular Biology, Birkbeck College, Malet Street, London WC1E 7HX, UK. [3]Centre for Medical Biotechnology, FB Biology and Geography, University Duisburg-Essen, Universitätsstrasse, D-45117 Essen, Germany.
*These authors contributed equally to the work.

The 24-mer of DegP has a molecular mass of 1.13 MDa and forms a spherical shell with 432 symmetry (Fig. 1b). Its diameter of 195 Å is consistent with electron-microscopic images of negatively stained $DegP_{24}$–OMP particles, which were about 190 Å in diameter (Supplementary Fig. 2c, g). In the crystal structure of $DegP_{24}$, eight trimers are located at the vertices of an octahedron that assembles a protein shell of about 31 Å thickness enclosing a large internal cavity about 110 Å in diameter. The inside volume of the sphere is about 700,000 $Å^3$, which is roughly eightfold that of an open cavity of GroEL (Fig. 1c). Superposition of $DegP_{24}$ on $DegP_6$ illustrates that this remarkably large cavity could, in theory, accommodate a 300-kDa protein (Supplementary Fig. 3). The protein shell has wide pores allowing access to the inner cavity. The largest of these pores is 35 Å wide and runs along the particle's four-fold axes, whereas smaller channels that coincide with the two-fold axes are 14 Å in diameter. The 24 proteolytic sites are accessible only from the interior of the cavity. Thus, protein substrates would have to be encapsulated in the central compartment during oligomer assembly or would have to enter the particle through one of the six pores. The size of these pores is large enough to allow small folded proteins (less than 25 kDa) or unfolded polypeptides to diffuse in and out of the protein shell. The overall organization of the DegP trimer, in which three protease domains are encircled by six PDZ domains, dictates the assembly of $DegP_{24}$. The outward-extending PDZ domains, PDZ1 and PDZ2, are in close contact with the PDZ domains of two adjacent trimeric rings. Four DegP trimers are arranged by these interactions around the four-fold symmetry axis and form the large pores of the particle by constituting a ring of four PDZ1/PDZ2* pairs (the asterisk denotes a neighbouring molecule) (Supplementary Fig. 4).

Regulation of protease activity by oligomer reassembly

The crystal structure of the DegP hexamer revealed that regulation of protease activity depends on loop LA (ref. 12; for the nomenclature of protease loops see Fig. 2a). In the corresponding inactive conformation, loop LA protrudes into the active site of one subunit of the opposite trimeric ring, where it closely interacts with the active-site loops L1* and L2*. The resulting loop triad LA–L1*–L2* adopts an entirely twisted conformation that blocks the entrance to the active site and distorts adjustment of the catalytic triad, the oxyanion hole and the substrate-specificity pocket (Fig. 2a). Our structural data indicate that transformation of the hexamer into the larger oligomers extracts loop LA from the active site of the molecular neighbour and releases loops L1 and L2 to set up a functional proteolytic site. For example, the stretched conformation of loop L1 observed in the inactive DegP is remodelled into the typical turn structure that is essential in forming the oxyanion hole (Fig. 2b). Thus the conversion of $DegP_6$ into $DegP_{12}$ or $DegP_{24}$ is crucial in regulating protease activity.

To study determinants of oligomer reassembly, we incubated unfolded protein substrates with the hexameric form of $DegP_{S210A}$ and followed complex formation by SEC. Whereas the larger substrates bovine serum albumin and casein were generally captured in the $DegP_{24}$ complex (Supplementary Fig. 5a), lysozyme affected oligomerization in a concentration-dependent manner, triggering the formation of $DegP_{12}$ at lower concentrations and that of $DegP_{24}$ at elevated concentrations. Because the redistribution of oligomers did not depend on the amount of DegP (Fig. 2c), we presume that the higher-order particles reflect mainly the size and concentration of substrate.

When we tested the ability of proteolytically active DegP to form such complexes, we detected the transient formation of $DegP_{24}$ and $DegP_{12}$. A short incubation with substrates transformed $DegP_6$ into the larger oligomers, whereas after prolonged incubation, when degradation had been completed, DegP reverted to its hexameric state (Fig. 2d and Supplementary Fig. 5b). To probe the functionality of $DegP_6$, we assayed protease activity with a previously identified chromogenic peptide substrate[5]. Although no larger complexes were formed (data not shown), DegP slowly hydrolysed the model peptide, suggesting that $DegP_6$ is capable of hydrolysing oligopeptides. However, when we added denatured lysozyme and stimulated the formation of $DegP_{12}$ and $DegP_{24}$, degradation of the chromogenic substrate was accelerated 15-fold (Supplementary Fig. 5c). Similarly,

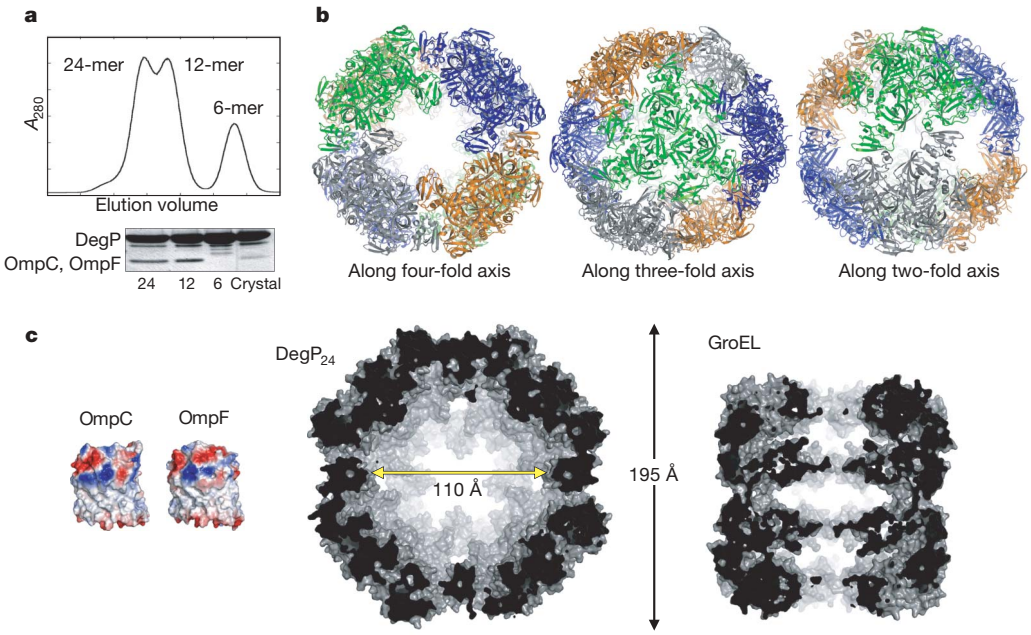


Figure 1 | The $DegP_{24}$ particle. **a**, Identification of $DegP_6$, $DegP_{12}$ and $DegP_{24}$ by SEC. Only the two larger oligomers had different OMPs bound. Mass spectrometry of dissolved $DegP_{24}$ crystals revealed that the visible additional band corresponds to OmpC and OmpF. **b**, Ribbon plot of $DegP_{24}$ illustrating its overall architecture with the trimeric units coloured differently. The particle is shown in three different orientations along the molecular four-fold, three-fold and two-fold axes. **c**, To illustrate the size of the inner cavity of $DegP_{24}$, the molecular surfaces of OmpF (PDB accession number 2omf) and OmpC (2j1n), with mapped electrostatic potential, are shown together with the surfaces of the sliced-open $DegP_{24}$ and GroEL (1kpo) chaperones.

at elevated temperatures, at which the protease activity of DegP is markedly upregulated[3], $DegP_6$ was destabilized and the oligomer equilibrium shifted to $DegP_3$ (Supplementary Fig. 5d). Taken together, these data indicate that DegP exists in a dynamic equilibrium of different multimers that have specific functions in protein quality control. Whereas $DegP_6$ seems to represent the resting state with reduced peptidase activity, $DegP_{12}$ and $DegP_{24}$ should function as protease–chaperone complexes acting on misfolded proteins. Because the high-molecular-mass particles are stabilized only as long as misfolded proteins are bound, the activity of DegP is linked directly to folding stress. Furthermore, the flexible encapsulation mechanism should guarantee quality control of a broad range of client proteins.

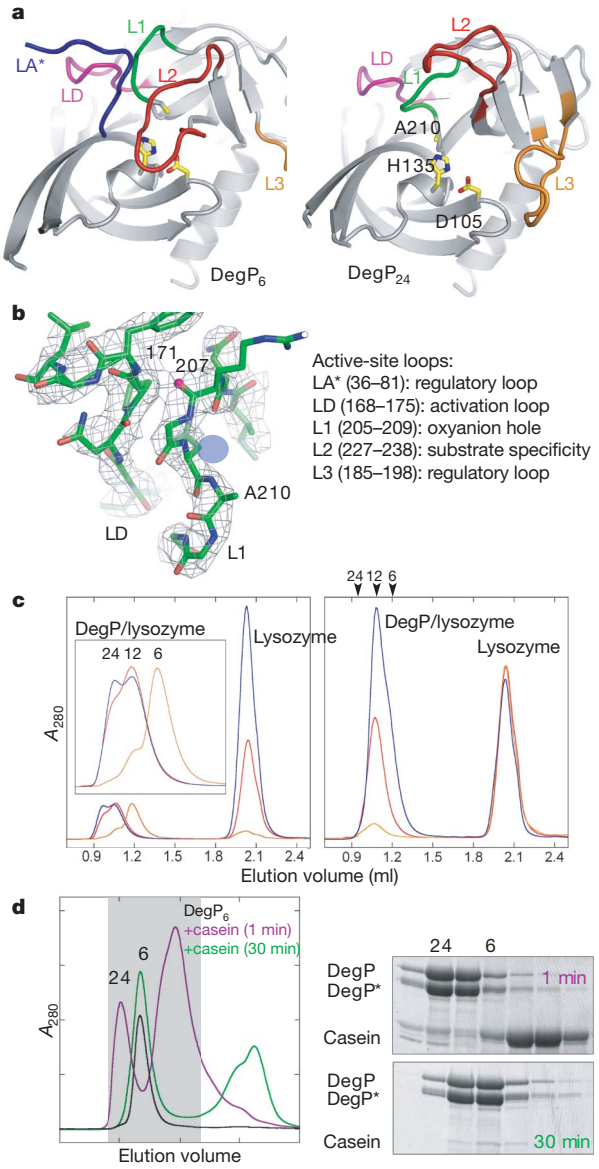


Figure 2 | Regulation of protease activity by oligomer reassembly. **a**, Ribbon plot of the protease domain of $DegP_6$ (1kj9) and $DegP_{24}$, highlighting the mechanistically important loops LA*, LD, L1, L2 and L3. Residues of the catalytic triad (Asp 105, His 135 and Ala 210) are shown in stick mode and the loop nomenclature used[12,41] is indicated. **b**, Electron density of the active-site loops L1 and LD. The $2F_o - F_c$ simulated annealing omit map was calculated at 3.0 Å resolution (contoured at 1.1σ) after omitting loops L1 and LD from the refined model. The oxyanion hole (blue sphere) and the main-chain carbonyl group of Arg 207 are highlighted. The position of the latter oxygen is a distinctive feature of proteolytically active HtrA proteases. **c**, Denatured lysozyme and $DegP_6$ were incubated in different ratios and the resulting complexes were analysed by SEC. Left: incubation of different amounts of lysozyme (orange, 30 μM; red, 300 μM; blue, 600 μM) with $DegP_6$ (15 μM). Right: incubation of different amounts of $DegP_6$ (orange, 3 μM; red, 15 μM; blue, 65 μM) with lysozyme (170 μM). **d**, Brief incubation of wild-type DegP with casein (1 min, magenta line) resulted in the formation of the $DegP_{24}$–casein complex (the pronounced low-molecular-mass peak represents unprocessed casein). After completion of degradation (30 min, green line), DegP recycled into its hexameric state. Composites of individual elution peaks are indicated on the SDS gel; the self-cleavage products of DegP are labelled DegP*.

DegP is a chaperone for folded OMPs

Identification of the co-purified and co-crystallized DegP–OMP complexes indicates that DegP has an active role in OMP biogenesis. To address the *in vivo* relevance of our findings, we analysed the OMP composition of wild-type and *degP*-null mutant strains (Fig. 3a). In the *degP* mutant, the levels of OmpA and OmpF in the outer membrane were decreased, whereas the level of OmpC was also lowered but to a smaller degree. It is known that the expression of OMPs is tightly regulated. For example, the σE stress response that is triggered by folding stress can decrease the synthesis of several OMPs by RNA-regulated transcriptional repression[24,25]. To test the consequence of deleting *degP* on OMP expression levels, we determined the amounts of OMPs in whole cell lysates and observed that the total amounts of expressed OMPs were similar in wild-type and *degP*-null strains (Fig. 3a). Thus the observed depletion of several OMPs in the outer membrane seems to be due to the lack of DegP activity in OMP biogenesis.

To characterize the observed complexes functionally, we examined the stability of OMPs bound to proteolytically active DegP. In

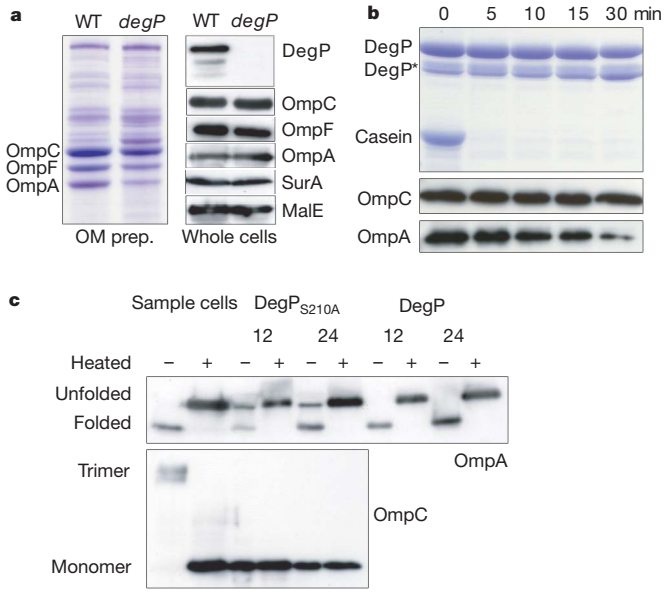


Figure 3 | Function of DegP in OMP biogenesis. **a**, Steady-state levels of OmpA, OmpC and OmpF in wild-type (WT) and *degP*-null mutant strains. Outer membranes were prepared from an equivalent number of wild-type and *degP*-null cells. Right: the corresponding steady-state levels of whole cell extracts. The constant SurA levels exclude stimulation of the σE stress response in the mutant, whereas MalE serves as a loading control. **b**, SDS–PAGE of the cleavage of different substrates by higher-order DegP complexes (40 μM). The cleavage reactions with casein (130 μM), OmpC and OmpA were performed at 37 °C and stopped at various time points. Self-cleavage products of DegP are labelled DegP*. **c**, Folding state of OmpA and OmpC in $DegP_{12}$ and $DegP_{24}$. Top: heated and unheated samples of whole cells, $DegP_{12}$ and $DegP_{24}$ were analysed by SDS–PAGE to distinguish between folded and unfolded OmpA. Unboiled samples of folded OmpA migrate at 30 kDa, whereas unboiled samples of partly folded or unfolded OmpA migrate at 35 kDa. Bottom: without heating, OmpC trimers do not dissociate on SDS gels. Therefore both DegP particles bind OmpC in its monomeric state.

contrast to misfolded model substrates, which were degraded within a few minutes, the co-purified OMPs were stable (Fig. 3b). Even in the presence of externally applied proteases, the bound OMPs were almost entirely resistant to proteolytic degradation (data not shown). They remained stably bound to DegP over a period of 30 min, a time frame that should be sufficient for targeted transport to the outer membrane.

Because it is known that DegP specifically degrades misfolded proteins, we examined whether the bound OMPs might contain tertiary structures that protect them from degradation. Thermodynamic stability studies of OmpA and other β-barrel membrane proteins indicated that the formation of tertiary structure can be conveniently followed by a shift in the apparent mass on SDS–PAGE gels[26]. The SDS gel-shift assay revealed that at least 50% of bound OmpA is present in a folded state in the higher-order particles of $DegP_{S210A}$ (Fig. 3c). Thus DegP seems to stabilize an assembly intermediate similar to the functionally related SurA chaperone, which favours the formation of a folded LamB protomer[21]. When we analysed the folding state of OmpA bound to the proteolytically active DegP, we observed that DegP degrades unfolded OmpA and stabilizes the folded protomers. Thus DegP functions as a genuine OMP chaperone. Moreover, OmpC trimers could not be detected in the large oligomeric complexes (Fig. 3c), suggesting that DegP selectively stabilizes folded OMP protomers but cannot support subsequent assembly (trimerization) steps, which are known to require additional folding factors such as lipopolysaccharide and YaeT[27].

Membrane attachment of $DegP_{24}$

To explore how DegP might interact with other molecules, we calculated the electrostatic potential of $DegP_{24}$ (Fig. 4a). We found that clusters of lysine and arginine residues originating from both PDZ domains render the electrostatic potential of the outer rim of the large pores strongly positive (Fig. 4a), thereby generating candidate sites for membrane attachment. Consistently, recent reports have emphasized the importance of PDZ domains in membrane localization[28–31]. To test the binding of $DegP_{24}$ to lipid membranes, we conducted a lipid sedimentation assay with liposomes prepared from bovine brain lipid extracts (Fig. 4b). Remarkably, $DegP_{24}$ bound to liposomes with a similar affinity to that of other membrane-associated PDZ proteins[32]. To monitor directly the influence of the PDZ domains on lipid binding, we generated two mutants in which the surface-exposed lysine residues 305, 379, 381 and 416 were replaced by either alanine ($DegP_{4A}$) or glutamate ($DegP_{3E}$). Dose–response experiments revealed that the lipid affinity of the $DegP_{4A}$ 24-mer is significantly decreased and that lipid binding of $DegP_{3E}$ is almost entirely impaired (Fig. 4b). These data indicate that DegP has exploited the PDZ domains to target cellular membranes. Alternatively, because the distance between the cytoplasmic and outer membranes is believed to be between 150 and 330 Å (refs 33, 34), the assembled $DegP_{24}$ could become wedged between the two membranes with the positively charged openings directly facing the phospholipid layers. Thus DegP could function as a periplasmic macropore, allowing the protected diffusion of OMP precursors from the inner membrane to the outer membrane.

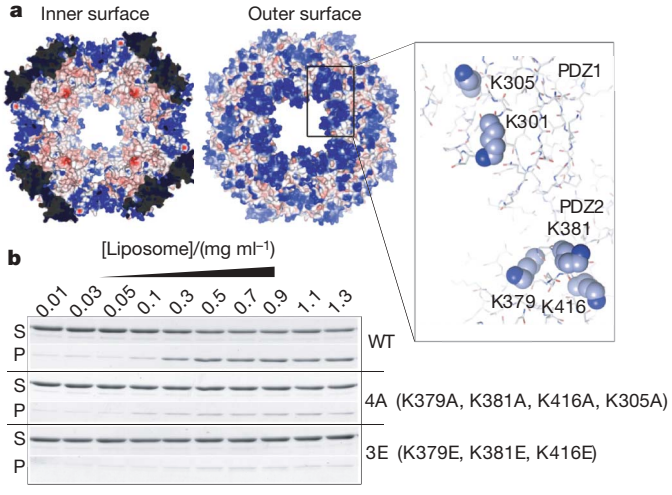


Figure 4 | Membrane attachment of $DegP_{24}$. **a**, The electrostatic potential of $DegP_{24}$ was calculated with PYMOL[42] and mapped on the molecular surface of the particle. Red indicates negatively charged regions, blue positively charged regions. Lysine residues of PDZ1 and PDZ2 that enclose the outer rim of the large pore and contribute to the positively charged patches are shown at the right. **b**, Sedimentation assay of $DegP_{24}$–OMP binding to bovine brain liposomes. Dose–response experiments were performed with a fixed amount of $DegP_{24}$–OMP (0.1 mg ml^{-1}) and increasing concentrations of liposomes. 'S' and 'P' refer to proteins present in the supernatant or pellet after centrifugation, and mutations of '4A' and '3E' are listed.

Electron-microscopic analysis reveals the encapsulated OMP

The $DegP_{12}$–OMP and $DegP_{24}$–OMP complexes were analysed by electron microscopy. Negative-stain electron-microscopic analysis of $DegP_{24}$–OMP yielded a map with octahedral symmetry that resembles the X-ray data filtered to an equivalent resolution (Supplementary Fig. 2g–k). However, the $DegP_{12}$–OMP complex was more homogeneous (compare Supplementary Fig. 2a and 2c). To define the subunit assembly and OMP density, we examined the $DegP_{12}$–OMP complex by cryo-electron microscopy.

The cryo-electron microscopy map of $DegP_{12}$–OMP shows a tetrahedral cage with a diameter of about 160 Å (Fig. 5a). Each face is made of a triangular density that fits well to a DegP trimer. In contrast to $DegP_{24}$, the inter-trimer contacts of $DegP_{12}$ are made by adjacent PDZ1 domains and do not seem to involve PDZ2 (Fig. 5b). Furthermore, the fitted cryo-electron microscopy map indicates that the catalytic sites open into the central cavity of the particle, which has a diameter of about 78 Å. The central cavity is occupied by a cylindrical density that fits remarkably well to the native β-barrel of OmpC (Fig. 5b). Because OmpA, OmpC, OmpF and LamB all form β-barrels with similar dimensions, the density observed in the central compartment could accommodate any of the potential OMP substrates. It is unlikely that unfolded proteins or the unstructured loop LA give rise to such a defined shape. Therefore, the extra density most probably represents an OMP monomer in an almost native conformation. Thus, the cryo-electron microscopy data provide further evidence that DegP sequesters OMP monomers in a substantially folded state and provides a view of a membrane protein precursor before its insertion into the membrane.

Conclusion

Large protein complexes with octahedral 432 or icosahedral 532 symmetry often form hollow protein shells that are used to store specific molecules. Classical storage devices for iron atoms and nucleic acids are ferritin and virus particles, respectively[35,36]. Our data show that DegP is another high-symmetry packaging device, whose central compartment is used to sequester unfolded proteins in the periplasm and to partition them between refolding and degradation pathways. In a first step, DegP has to sort out aberrant proteins with partly folded, aggregation-prone structures from properly folded proteins. $DegP_6$ seems to function as a substrate filter, because only unfolded proteins are capable of entering the cavity and assembling the functional protease–chaperone. Oligomer formation should not depend on the nature of the unfolded substrate, and thus both OMPs secreted into the periplasm and stress-damaged proteins should be equally well encapsulated by DegP. However, in contrast to other protease–chaperone systems, the inner cavity of $DegP_{12}$/$DegP_{24}$ combines the dual characteristics of a folding compartment and a proteolytic compartment. Because the proteolytic activity of DegP is

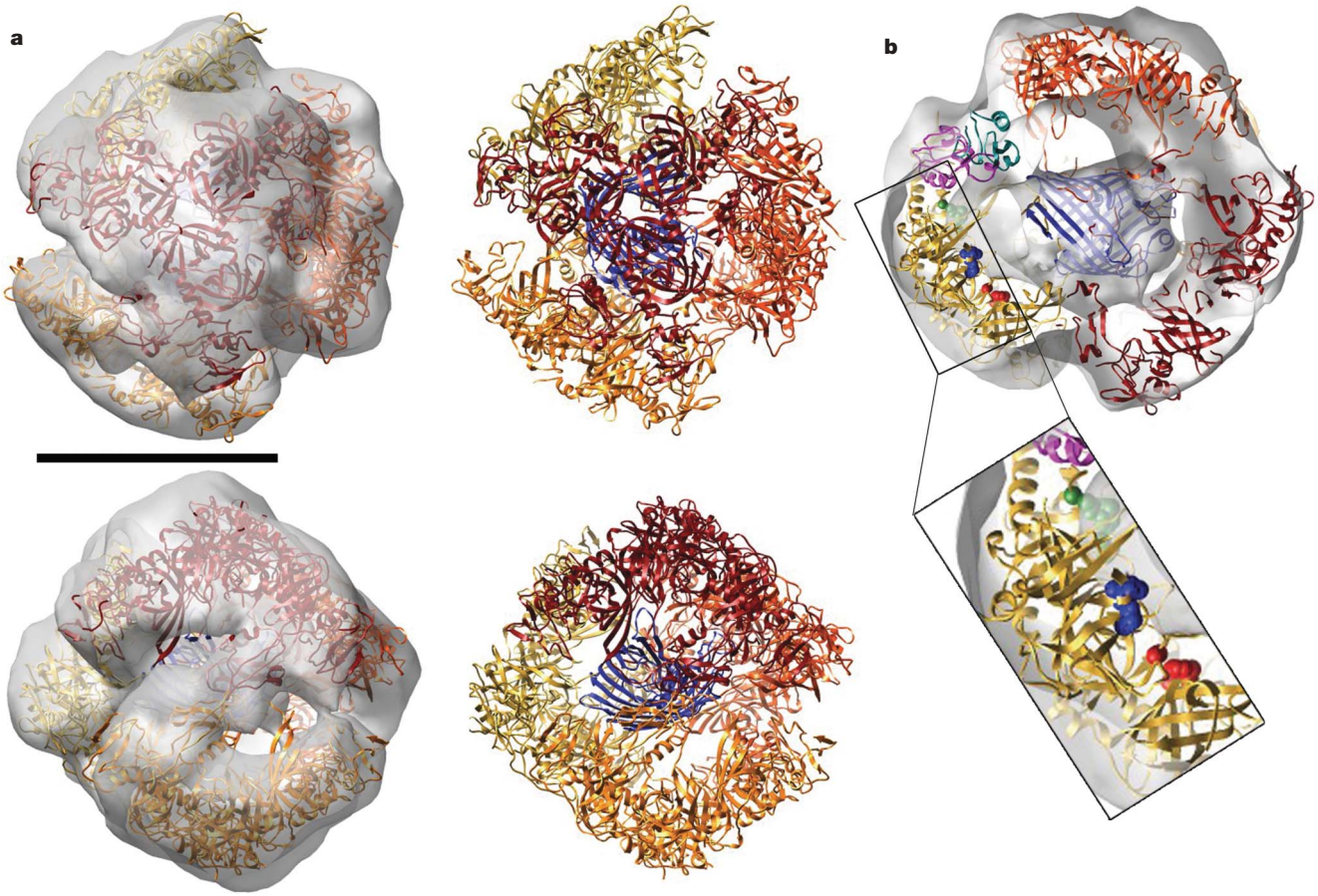


Figure 5 | Cryo-electron microscopy structure of the $DegP_{12}$–OMP complex. **a**, The asymmetric $DegP_{12}$–OMP complex viewed along the approximate three-fold (top) and two-fold (bottom) axes. In the left panels the ribbon model of the DegP dodecamer is overlaid with the semi-transparent three-dimensional map. **b**, Central section of the $DegP_{12}$–OMP electron microscopy map with an OmpC monomer (blue) modelled in the central density. The adjacent PDZ1 domains from neighbouring trimers are coloured in cyan and magenta. Three catalytic triads are coloured in red, green and blue and are shown magnified in the lower panel. Scale bar, 100 Å.

restricted to unfolded peptide structures, the fate of an encapsulated protein should depend mainly on its propensity to readily adopt its native, folded conformation and to escape the degradative machinery of DegP. Consistently, *in vitro* studies showed that several OMPs, including PhoE, LamB and OmpA, spontaneously fold into their β-barrel structures[37–40] and could therefore, in contrast to unfolded soluble proteins, remain protected inside the high-molecular-mass DegP particle.

METHODS SUMMARY

An improved purification procedure allowed separation of the $DegP_6$, $DegP_{12}$ and $DegP_{24}$ multimers. Mass spectrometry and western blot analysis revealed that specific OMPs were bound to $DegP_{12}$ and $DegP_{24}$. The $DegP_{24}$–OMP complex was crystallized and the structure was solved by the single anomalous dispersion method. In parallel, the structure of the $DegP_{12}$–OMP complex was determined by single-particle cryo-electron microscopy. *In vitro* complexes of DegP with several model substrates were analysed by SEC and SDS–PAGE, illustrating the reassembly of the resting $DegP_6$ into the proteolytically active $DegP_{12}$ and $DegP_{24}$ complexes. The transient nature of the higher-order complexes was shown by incubating wild-type DegP with an excess of substrate and immediate separation of the mixture by SEC. Degradation assays revealed the remarkable stability of OmpA and OmpC bound to proteolytically active DegP. The folding state of OMPs in corresponding complexes was analysed by mobility-shift assays, taking advantage of the different migration behaviours of folded and unfolded OmpA and of monomeric and trimeric OmpC. To explore the relevance of our findings *in vivo*, we isolated the outer-membrane fraction from *Escherichia coli* wild-type and *degP*-null mutant strains and analysed OMP levels on SDS–PAGE in the presence of 6 M urea to resolve OmpC and OmpF bands. As a control we analysed the corresponding OMP composition of whole-cell extracts. Mutational analysis uncovered PDZ residues that are crucial for membrane attachment of $DegP_{24}$, as shown by liposome binding assays.

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

Received 18 December 2007; accepted 15 April 2008.
Published online 21 May 2008.

1. Selkoe, D. J. Folding proteins in fatal ways. *Nature* **426,** 900–904 (2003).
2. Macario, A. J. & Conway de Macario, E. Sick chaperones, cellular stress, and disease. *N. Engl. J. Med.* **353,** 1489–1501 (2005).
3. Spiess, C., Beil, A. & Ehrmann, M. A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* **97,** 339–347 (1999).
4. Iwanczyk, J. *et al.* Role of the PDZ domains in *Escherichia coli* DegP protein. *J. Bacteriol.* **189,** 3176–3186 (2007).
5. Meltzer, M. *et al.* Allosteric activation of HtrA protease DegP by stress signals during bacterial protein quality control. *Angew. Chem. Int. Edn Engl.* **47,** 1332–1334 (2008).
6. Clausen, T., Southan, C. & Ehrmann, M. The HtrA family of proteases: implications for protein composition and cell fate. *Mol. Cell* **10,** 443–455 (2002).
7. Grau, S. *et al.* Implications of the serine protease HtrA1 in amyloid precursor protein processing. *Proc. Natl Acad. Sci. USA* **102,** 6021–6026 (2005).
8. Plun-Favreau, H. *et al.* The mitochondrial protease HtrA2 is regulated by Parkinson's disease-associated kinase PINK1. *Nature Cell Biol.* **9,** 1243–1252 (2007).
9. Harris, B. Z. & Lim, W. A. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **114,** 3219–3231 (2001).
10. Wilken, C., Kitzing, K., Kurzbauer, R., Ehrmann, M. & Clausen, T. Crystal structure of the DegS stress sensor: How a PDZ domain recognizes misfolded protein and activates a protease. *Cell* **117,** 483–494 (2004).
11. Hasselblatt, H. *et al.* Regulation of the σE stress response by DegS: how the PDZ domain keeps the protease inactive in the resting state and allows integration of

different OMP-derived stress signals upon folding stress. *Genes Dev.* **21,** 2659–2670 (2007).
12. Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M. & Clausen, T. Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature* **416,** 455–459 (2002).
13. Misra, R., Peterson, A., Ferenci, T. & Silhavy, T. J. A genetic approach for analyzing the pathway of LamB assembly into the outer membrane of *Escherichia coli*. *J. Biol. Chem.* **266,** 13592–13597 (1991).
14. Misra, R., CastilloKeller, M. & Deng, M. Overexpression of protease-deficient DegP(S210A) rescues the lethal phenotype of *Escherichia coli* OmpF assembly mutants in a degP background. *J. Bacteriol.* **182,** 4882–4888 (2000).
15. CastilloKeller, M. & Misra, R. Protease-deficient DegP suppresses lethal effects of a mutant OmpC protein by its capture. *J. Bacteriol.* **185,** 148–154 (2003).
16. Purdy, G. E., Fisher, C. R. & Payne, S. M. IcsA surface presentation in *Shigella flexneri* requires the periplasmic chaperones DegP, Skp, and SurA. *J. Bacteriol.* **189,** 5566–5573 (2007).
17. Ruiz, N., Kahne, D. & Silhavy, T. J. Advances in understanding bacterial outer-membrane biogenesis. *Nat. Rev. Microbiol.* **4,** 57–66 (2006).
18. Wu, T. *et al.* Identification of a multicomponent complex required for outer membrane biogenesis in *Escherichia coli*. *Cell* **121,** 235–245 (2005).
19. Alba, B. M. & Gross, C. A. Regulation of the *Escherichia coli* sigma-dependent envelope stress response. *Mol. Microbiol.* **52,** 613–619 (2004).
20. Chen, R. & Henning, U. A periplasmic protein (Skp) of *Escherichia coli* selectively binds a class of outer membrane proteins. *Mol. Microbiol.* **19,** 1287–1294 (1996).
21. Rouviere, P. E. & Gross, C. A. SurA, a periplasmic protein with peptidyl-prolyl isomerase activity, participates in the assembly of outer membrane porins. *Genes Dev.* **10,** 3170–3182 (1996).
22. Rizzitello, A. E., Harper, J. R. & Silhavy, T. J. Genetic evidence for parallel pathways of chaperone activity in the periplasm of *Escherichia coli*. *J. Bacteriol.* **183,** 6794–6800 (2001).
23. Sklar, J. G., Wu, T., Kahne, D. & Silhavy, T. J. Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in *Escherichia coli*. *Genes Dev.* **21,** 2473–2484 (2007).
24. Rhodius, V. A., Suh, W. C., Nonaka, G., West, J. & Gross, C. A. Conserved and variable functions of the σE stress response in related genomes. *PLoS Biol.* **4,** e2 (2006).
25. Guisbert, E., Rhodius, V. A., Ahuja, N., Witkin, E. & Gross, C. A. Hfq modulates the σE-mediated envelope stress response and the σ32-mediated cytoplasmic stress response in *Escherichia coli*. *J. Bacteriol.* **189,** 1963–1973 (2007).
26. Schweizer, M., Hindennach, I., Garten, W. & Henning, U. Major proteins of the *Escherichia coli* outer cell envelope membrane. Interaction of protein II with lipopolysaccharide. *Eur. J. Biochem.* **82,** 211–217 (1978).
27. Sen, K. & Nikaido, H. Lipopolysaccharide structure required for *in vitro* trimerization of *Escherichia coli* OmpF porin. *J. Bacteriol.* **173,** 926–928 (1991).
28. Zimmermann, P. *et al.* PIP(2)-PDZ domain binding controls the association of syntenin with the plasma membrane. *Mol. Cell* **9,** 1215–1225 (2002).
29. Yan, J. *et al.* Structure of the split PH domain and distinct lipid-binding properties of the PH-PDZ supramodule of α-syntrophin. *EMBO J.* **24,** 3985–3995 (2005).
30. Mortier, E. *et al.* Nuclear speckles and nucleoli targeting by PIP2–PDZ domain interactions. *EMBO J.* **24,** 2556–2565 (2005).
31. Pan, L. *et al.* Clustering and synaptic targeting of PICK1 requires direct interaction between the PDZ domain and lipid membranes. *EMBO J.* **26,** 4576–4587 (2007).
32. Wu, H. *et al.* PDZ domains of Par-3 as potential phosphoinositide signaling integrators. *Mol. Cell* **28,** 886–898 (2007).
33. Dubochet, J., McDowall, A. W., Menge, B., Schmid, E. N. & Lickfeld, K. G. Electron microscopy of frozen-hydrated bacteria. *J. Bacteriol.* **155,** 381–390 (1983).
34. Leduc, M., Frehel, C. & van Heijenoort, J. Correlation between degradation and ultrastructure of peptidoglycan during autolysis of *Escherichia coli*. *J. Bacteriol.* **161,** 627–635 (1985).
35. Winkler, F. K., Schutt, C. E., Harrison, S. C. & Bricogne, G. Tomato bushy stunt virus at 5.5-Å resolution. *Nature* **265,** 509–513 (1977).
36. Banyard, S. H., Stammers, D. K. & Harrison, P. M. Electron density map of apoferritin at 2.8-Å resolution. *Nature* **271,** 282–284 (1978).
37. Park, S., Liu, G., Topping, T. B., Cover, W. H. & Randall, L. L. Modulation of folding pathways of exported proteins by the leader sequence. *Science* **239,** 1033–1035 (1988).
38. de Cock, H., Hendriks, R., de Vrije, T. & Tommassen, J. Assembly of an *in vitro* synthesized *Escherichia coli* outer membrane porin into its stable trimeric configuration. *J. Biol. Chem.* **265,** 4646–4651 (1990).
39. Klose, M. *et al.* Membrane assembly of the outer membrane protein OmpA of *Escherichia coli*. *J. Biol. Chem.* **268,** 25664–25670 (1993).
40. Rouviere, P. E. & Gross, C. A. SurA, a periplasmic protein with peptidyl-prolyl isomerase activity, participates in the assembly of outer membrane porins. *Genes Dev.* **10,** 3170–3182 (1996).
41. Perona, J. J. & Craik, C. S. Structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4,** 337–360 (1995).
42. DeLano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific, Palo Alto, CA, 2002).

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

Acknowledgements We thank K. Mechtler and I. Steinmacher for assistance with mass spectrometry; the staff at the European Synchrotron Radiation Facility and the Swiss Light Source for assistance with collecting synchrotron data; N. Boisset and R. Trujillo for providing image-processing scripts; L. Wang for support with electron microscopy; D. Houldershaw for computer support; and E. Orlova for discussion. The Research Institute of Molecular Pathology (IMP) is funded by Boehringer Ingelheim. E.S. and H.R.S. were supported by the UK Biotechnology and Biological Sciences Research Council, M.E. by the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie, T.C. by the EMBO Young Investigator Program, and T.K. and J.S. by the Austrian Science Fund.

Author Information The $DegP_{24}$ structure is deposited in the Protein Data Bank under accession number 3cs0. The fitted model of the electron microscopic three-dimensional map of $DegP_{12}$–OMP is deposited in the Protein Data Bank under accession number 2zle. The electron-microscopic three-dimensional maps are deposited at the Electron Microscopy Data Bank (http://www.ebi.ac.uk/msd-srv/emsearch/index.html) under accession codes EMD-1504 and EMD-1505 for $DegP_{24}$–OMP and $DegP_{12}$–OMP, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.C. (clausen@imp.univie.ac.at).

METHODS

Isolation of $DegP_6$, $DegP_{12}$ and $DegP_{24}$. C-terminally His-tagged DegP, $DegP_{S210A}$, $DegP_{4A}$ and $DegP_{3E}$ were expressed as full-length proteins containing the periplasmic signal sequence in a *degP*-null strain (CLC198, *degP*::Tn*10*) (ref. 3). Site-directed mutagenesis was performed with the QuikChange multi site-directed mutagenesis kit (Stratagene) and all DegP variants were isolated at 4 °C by means of the following purification procedure. Cells were harvested by centrifugation, resuspended in 200 mM NaCl, 100 mM HEPES-NaOH pH 8.0 (buffer A) and disrupted by sonication. The cleared lysate was purified with a Ni^{2+}-nitrilotriacetate resin (Qiagen) with standard procedures. Samples containing DegP were applied to a hydroxyapatite column (Bio-Rad) and eluted with a linear gradient of 0–500 mM potassium phosphate in buffer A (Supplementary Fig. 1a). Two distinct DegP fractions could be discerned, one of which represented $DegP_6$ that was crystallized previously[12], whereas the second fraction contained DegP together with prominent protein bands of 35 kDa. The latter sample was concentrated with VIVASPIN concentrators (cutoff 50 kDa) and applied to a Superdex 200 column (preparative grade; GE Healthcare) equilibrated with 300 mM NaCl, 50 mM HEPES-NaOH pH 8.0. During the SEC run, three DegP oligomers were separated, representing $DegP_6$, $DegP_{12}$ and $DegP_{24}$. The overall sizes of the particles were determined by dynamic light scattering (DynaPro-801; Protein-Solutions Inc.) at 20 °C, with protein concentrations from 0.5 to 2 mg ml^{-1} and 10-s acquisition times.

Crystallization and structure solution of the $DegP_{24}$–OMP complex. The SeMet-containing $DegP_{24}$(S210A)–OMP complex was crystallized at 19 °C by the vapour diffusion method. For crystallization, 1 µl of 10 mg ml^{-1} $DegP_{24}$–OMP was mixed with 0.3 µl of FOS-choline-10 and 0.5 µl of a reservoir solution containing 23% (v/v) PEG550MME, 0.1 M Tris-HCl pH 8.5 and 0.1 M NaCl. Before flash-freezing the crystals in liquid nitrogen, the drop was left open for 20 min at 19 °C. The protein crystallized in the cubic space group *F*432 with unit cell parameters of $a = b = c = 253.9$ Å and one DegP protomer in the crystallographic asymmetric unit. A single-wavelength anomalous dispersion data set to 3.0 Å resolution was collected at beamline ID23-1 at the European Synchrotron Radiation Facility. Diffraction data were processed with programs from the XDS package[43], and 11 out of 14 selenium sites of DegP were instantly located using Shake and Bake[44]. Subsequent phasing was performed with Sharp[45]. The model was built with O[46] and refined with CNS[47]. Data collection, phasing and refinement statistics are summarized in Supplementary Table 1.

Identification of substrate proteins. To identify co-purified protein substrates, we examined crystals of the $DegP_{24}$ complex for potential binding partners. After extensive washing, the $DegP_{24}$ crystals were solubilized and the proteins were separated by SDS–PAGE. In addition to the strong band representing DegP, we detected several faint bands, which were analysed by mass spectrometry. Protein spots were excised from the SDS–PAGE gel, washed, reduced, *S*-alkylated, and digested with trypsin. Resulting peptide fragments were analysed on a hybrid linear ion trap – Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (LTQ-FT Ultra; ThermoFisher). For peptide identification a database search was performed with Mascot (Matrix Science). Ultimately, hits were confirmed by western blot analysis.

Electron microscopy, image processing and atomic structure fitting. Samples of the $DegP_{12}$–OMP and $DegP_{24}$–OMP complexes at final concentrations of 0.0064 and 0.01 mg ml^{-1}, respectively, were negatively stained with 2% (w/v) uranyl acetate on glow-discharged, carbon-coated grids (Agar Scientific). For cryo-electron microscopic measurements, samples containing the $DegP_{12}$(S210A)–OMP complex (0.16 mg ml^{-1}) were embedded in vitreous ice by using C-flat holey carbon grids (CF-2/2-4C-100; Protochip) and a Vitrobot (FEI) at 20 °C and 100% relative humidity. Low-dose charge-coupled device (CCD) images were recorded on a 4,096 × 4,096 Gatan CCD camera with a Tecnai F20 (FEI) at 200 kV and a defocus of about 2 µm, at a magnification of ×68,100 corresponding to a pixel size of 2.22 Å. Subsequently, adjacent pixels were 2 × 2 averaged to yield a pixel size of 4.44 Å. A total of 9,890 particles of the $DegP_{12}$(S210A)–OMP complex were selected from 64 CCD images and windowed into 100-pixel × 100-pixel boxes using the EMAN/BOXER software package[48], corrected for the effects of the contrast transfer function by phase flipping and the contrast was inverted. Images were processed with SPIDER, version 11.12 (ref. 49) and IMAGIC-5 (ref. 50). Images were centred against a circular mask. Initial reference-free alignment[51] was refined by multi-reference alignment in SPIDER[49]. The first three-dimensional reconstruction was calculated by angular reconstitution[52] in IMAGIC. On the basis of the observation of two-fold and three-fold views and the subunit number of 12, tetrahedral symmetry was applied. Other possible symmetries such as *C*3, *D*2 and *D*3 were not consistent with the data. After initial projection matching with tetrahedral symmetry, the structure was subsequently refined without symmetry. To validate the substrate density, it was masked out of the map; however, it fully reappeared in subsequent refinement. The final three-dimensional map was calculated from 6,285 particles (Supplementary Fig. 2c) and had a resolution of 28 Å as determined by Fourier shell correlation at 0.5 correlation (Supplementary Fig. 6). The three-dimensional map was contoured at threshold of 3σ, giving a volume of 600 kDa, corresponding to the expected molecular mass of $DegP_{12}$–OMP. The X-ray structures of the DegP trimer[12] and OmpC[53] were fitted manually and then refined in Chimera[54]. For the $DegP_{24}$–OMP complex, 3,828 particles from 40 negative-stain CCD images were windowed into 130-pixel × 130-pixel boxes and processed as for the $DegP_{12}$–OMP complex. The starting map was obtained by angular reconstitution and refined by projection matching with octahedral symmetry. The $DegP_{24}$ X-ray structure was filtered to 25 Å resolution for comparison with the electron-microscopic images (Supplementary Fig. 2j, k).

Biochemical characterization of DegP–substrate complexes. To follow the degradation of the model substrate, casein, we incubated 40 µM DegP with 130 µM casein in 25 mM Hepes–NaOH pH 7.5, 150 mM NaCl and 5 mM $MgCl_2$ at 37 °C. In parallel we followed the degradation of OmpA and OmpC by incubating the DegP–OMP complexes in the same degradation buffer at 37 °C. At certain time points the reaction was stopped by adding SDS loading buffer supplemented with 8 M urea to take aliquots. Subsequently, the aliquots were incubated for 15 min at 95 °C and analysed by SDS–PAGE. DegP and casein were detected by Coomassie stain, and degradation of OmpA and OmpC was monitored by western blot analysis (see below).

To follow the complex formation of hexameric $DegP_{S210A}$ with casein, we incubated 20 µl of 80 µM $DegP_{S210A}$ with 20 µl of 170 µM casein in 50 mM Hepes-NaOH pH 7.5, 150 mM NaCl, 10 mM dithiothreitol (DTT). To survey the formation of complexes with lysozyme (or bovine serum albumin), the protein substrate was initially denatured by preparing a 100 (50) mg ml^{-1} protein solution in 4 (8) M urea and 10 mM DTT. Subsequently, 2.5 (1) µl of substrate was added to 50 µl of degradation buffer containing $DegP_{S210A}$. Both assays were incubated for 10 min at 37 °C before samples were injected on a Superdex 200 gel-filtration column (PC 3.2/30; GE Healthcare). Hexameric $DegP_{S210A}$ was used as a control.

To determine the dynamics of complex formation, we incubated wild-type DegP with casein for different durations and explored the size of the resulting complexes. First, we mixed 15 µl of 320 µM DegP with 50 µl of 2 mM casein and analysed the reaction mixture immediately by SEC. For a second gel-filtration run, we mixed wild-type DegP (320 µM) with a lower concentration of casein (425 µM) and incubated the sample for 30 min at 37 °C. Analogously, we incubated wild-type $DegP_{12}$–OMP for 3 h at 37 °C and used SEC and SDS–PAGE to follow the conversion of $DegP_{12}$ to $DegP_6$ during OMP degradation.

Determination of OMP levels in outer membranes and whole cells. Outer membranes of *E. coli* MC4100 wild-type and *degP*-null mutant strain were prepared as described previously[55]. To evaluate protein levels in whole cells, both strains were grown in Luria–Bertani medium until they reached the stationary phase. After measuring the optical density, we took standardized aliquots to obtain pellets of equal cell number. Cell pellets were dissolved in equivalent volumes of SDS loading buffer and boiled for 15 min at 95 °C. Each sample (10 µl) was resolved by SDS–PAGE, transferred to a poly(vinylidene difluoride) membrane (Immobilon-P; Millipore) and probed with antibodies against DegP (1:10,000 dilution), OmpA (1:20,000), OmpC (1:20,000), OmpF (1:20,000) and SurA (1:20,000). After incubation with the secondary antibody, protein bands were revealed with an ECL Plus Western Blotting detection System (GE Healthcare) and Hyperfilm ECL (Amersham Biosciences).

Mobility-shift assay of folded and unfolded OMPs. Cells of wild-type strain MC4100 were harvested in the stationary phase and lysed by incubation with 50 mM Tris–HCl pH 7.5, 500 mM NaCl, 5 mM EDTA, 1% SDS and 1% 2-mercaptoethanol. The protein concentration of the cleared lysate was determined and samples with equal protein amounts were prepared. To ensure the complete unfolding of OMPs, one sample was boiled at 95 °C for 15 min in SDS loading buffer supplemented with 4 M urea. To maintain the folded states of OMPs, the second sample was incubated at 25 °C with a loading buffer lacking SDS[56]. Analogously, samples with or without SDS and with or without heating were prepared from the isolated $DegP_{12}$–OMP and $DegP_{24}$–OMP complexes. Finally, all samples were characterized by SDS–PAGE and western blot analysis.

Lipid-binding assays. Lipid binding of $DegP_{wt}$, $DegP_{4A}$ and $DegP_{3E}$ 24-mers was performed as described[29]. In brief, brain lipid extracts (Folch fraction I; Sigma) were suspended in 140 mM NaCl, 20 mM Hepes-NaOH pH 7.4. The proteins were incubated for 15 min at 37 °C at a concentration of 0.1 mg ml^{-1} with various lipid concentrations. Subsequently, samples were centrifuged for 30 min at 100,000*g* and 4 °C. After the supernatant was removed and the pellet had been resuspended with an equivalent volume, samples were analysed by SDS–PAGE.

43. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
44. Weeks, C. M. & Miller, R. The design and implementation of SnB version 2.0. *J. Appl. Cryst.* **32,** 120–124 (1999).
45. de la Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
46. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
47. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
48. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
49. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116,** 190–199 (1996).
50. van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116,** 17–24 (1996).
51. Penczek, P., Radermacher, M. & Frank, J. Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* **40,** 33–53 (1992).
52. van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21,** 111–123 (1987).
53. Basle, A., Rummel, G., Storici, P., Rosenbusch, J. P. & Schirmer, T. Crystal structure of osmoporin OmpC from *E. coli* at 2.0 Å. *J. Mol. Biol.* **362,** 933–942 (2006).
54. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
55. Matsuyama, S., Inokuchi, K. & Mizushima, S. Promoter exchange between Ompf and Ompc, genes for osmoregulated major outer-membrane proteins of *Escherichia coli* K-12. *J. Bacteriol.* **158,** 1041–1047 (1984).
56. Nakae, T., Ishii, J. & Tokunaga, M. Subunit structure of functional porin oligomers that form permeability channels in the outer membrane of *Escherichia coli*. *J. Biol. Chem.* **254,** 1457–1461 (1979).

LETTERS

Direct observation of Anderson localization of matter waves in a controlled disorder

Juliette Billy[1], Vincent Josse[1], Zhanchun Zuo[1], Alain Bernard[1], Ben Hambrecht[1], Pierre Lugan[1], David Clément[1], Laurent Sanchez-Palencia[1], Philippe Bouyer[1] & Alain Aspect[1]

In 1958, Anderson predicted the localization[1] of electronic wavefunctions in disordered crystals and the resulting absence of diffusion. It is now recognized that Anderson localization is ubiquitous in wave physics[2] because it originates from the interference between multiple scattering paths. Experimentally, localization has been reported for light waves[3–7], microwaves[8,9], sound waves[10] and electron gases[11]. However, there has been no direct observation of exponential spatial localization of matter waves of any type. Here we observe exponential localization of a Bose–Einstein condensate released into a one-dimensional waveguide in the presence of a controlled disorder created by laser speckle[12]. We operate in a regime of pure Anderson localization, that is, with weak disorder—such that localization results from many quantum reflections of low amplitude—and an atomic density low enough to render interactions negligible. We directly image the atomic density profiles as a function of time, and find that weak disorder can stop the expansion and lead to the formation of a stationary, exponentially localized wavefunction—a direct signature of Anderson localization. We extract the localization length by fitting the exponential wings of the profiles, and compare it to theoretical calculations. The power spectrum of the one-dimensional speckle potentials has a high spatial frequency cutoff, causing exponential localization to occur only when the de Broglie wavelengths of the atoms in the expanding condensate are greater than an effective mobility edge corresponding to that cutoff. In the opposite case, we find that the density profiles decay algebraically, as predicted in ref. 13. The method presented here can be extended to localization of atomic quantum gases in higher dimensions, and with controlled interactions.

The transport of quantum particles in non-ideal material media (for example the conduction of electrons in an imperfect crystal) is strongly affected by scattering from the impurities in the medium. Even for weak disorder, semiclassical theories, such as those based on the Boltzmann equation for matter waves scattering from the impurities, often fail to describe transport properties[2], and fully quantum approaches are necessary. For example, Anderson localization[1], which predicts metal–insulator transitions, is based on interference between multiple scattering paths, leading to localized wavefunctions with exponentially decaying profiles. While Anderson's theory applies to non-interacting particles in static (quenched) disordered potentials[1], both thermal phonons and repulsive interparticle interactions significantly affect Anderson localization[14,15]. To our knowledge, no direct observation of exponentially localized wavefunctions in space has been reported in condensed matter.

Degenerate atomic quantum gases can be used to study a number of basic models in condensed matter theory experimentally, with unprecedented control and measurement possibilities (see refs 16, 17 and references therein). In investigating the behaviour of matter waves in disordered potentials[18], key advantages of atomic quantum gases are the possibility to implement systems in any dimension; the control of the interatomic interactions, either by density control or by Feshbach resonances; the possibility to design perfectly controlled and phonon-free disordered potentials; and the opportunity to measure *in situ* atomic density profiles via direct imaging. The quest for evidence of Anderson localization of Bose–Einstein condensates (BECs) in optical disordered potentials has thus attracted considerable attention in recent years[19–22]. Experiments using ultracold atoms have shown evidence of dynamical localization associated with a kicked rotor[23,24], which can be considered as a mapping onto momentum space of the Anderson localization phenomenon. Suppression of one-dimensional transport of BECs has been observed[19,20], but this occurred in a regime of strong disorder and strong interactions where localization is due to classical reflections from large peaks of the disordered potential. Here we report direct observation in real space of one-dimensional localization of a BEC in the regime of Anderson localization, that is, with weak disorder and negligible interatomic interactions.

Our experiment (sketched in Fig. 1a, b), starts with a small, elongated BEC (1.7×10^4 atoms of rubidium-87, with transverse and longitudinal radii of 3 μm and 35 μm, respectively, corresponding to the trapping frequencies given below, and a chemical potential of $\mu_{in}/h = 219$ Hz, where h is Planck's constant). The BEC is produced in an anisotropic opto-magnetic hybrid trap. A far-off-resonance laser beam (wavelength 1.06 μm, to be compared with the resonant wavelength of rubidium, 0.78 μm) creates an optical waveguide along the horizontal z axis[25], with a transverse harmonic confinement of frequency $\omega_\perp/2\pi = 70$ Hz. A shallow magnetic trap confines the BEC in the longitudinal direction ($\omega_z/2\pi = 5.4$ Hz).

The longitudinal confinement is switched off at $t = 0$, and the BEC starts to expand in the guide along the z direction under the effect of the initial repulsive interaction energy. A weakly anti-trapping magnetic field compensates the residual longitudinal trapping of the optical waveguide, so that the atoms can freely expand along the z direction over several millimetres. The expanding BEC can be imaged at any chosen time t after release by switching off the optical guide and irradiating the atoms with a resonant probe of duration 50 μs. An ultrasensitive electron-multiplying charge-coupled-device camera allows us to make an image of the fluorescing atoms with a resolution of 15 μm and a one-dimensional atomic density sensitivity of close to one atom per micrometre.

A disordered potential is applied to the expanding BEC using an optical speckle field produced by passing a laser beam (wavelength 0.514 μm) through a diffusing plate[22]. The detuning from the atomic frequency is great enough, and the intensity low enough, that spontaneous photon scattering on the atoms is negligible during the expansion, and we have a purely conservative disordered potential,

[1]Laboratoire Charles Fabry de l'Institut d'Optique, CNRS and Univ. Paris-Sud, Campus Polytechnique, RD 128, F-91127 Palaiseau cedex, France.

which extends 4 mm along the z direction. The three-dimensional autocorrelation of the disordered potential—that is, of the light intensity—is determined by diffraction from the diffusive plate onto the atoms' location[22].

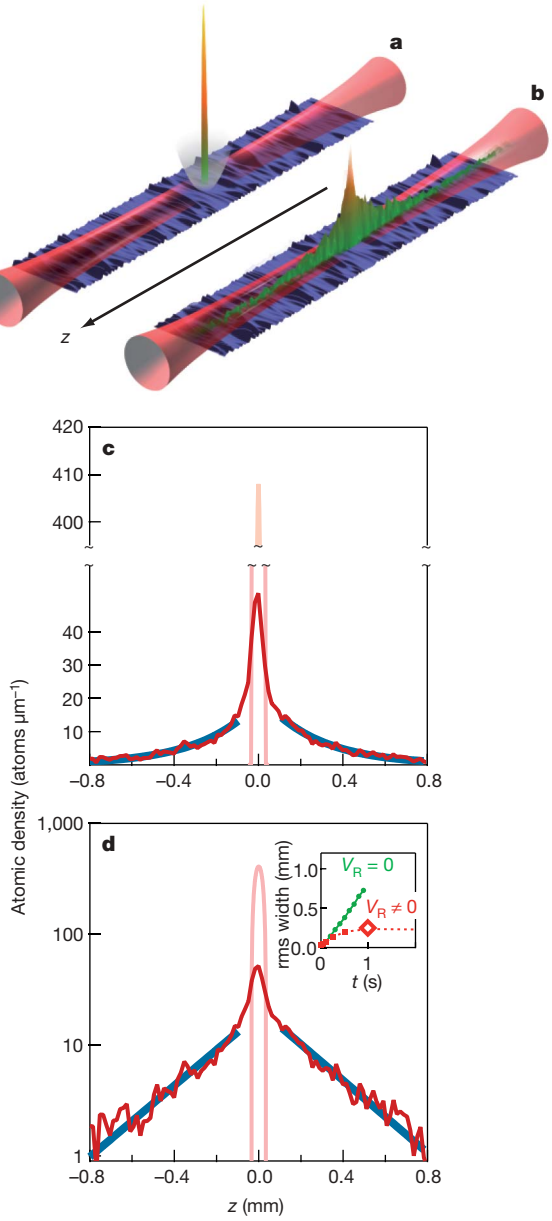


Figure 1 | Observation of exponential localization. a, A small BEC (1.7×10^4 atoms) is formed in a hybrid trap that is the combination of a horizontal optical waveguide, ensuring a strong transverse confinement, and a loose magnetic longitudinal trap. A weak disordered optical potential, transversely invariant over the atomic cloud, is superimposed (disorder amplitude V_R low in comparison with the chemical potential μ_{in} of the initial BEC). **b**, When the longitudinal trap is switched off, the BEC starts expanding and then localizes, as observed by direct imaging of the fluorescence of the atoms irradiated by a resonant probe. In **a** and **b**, false-colour images and sketched profiles are for illustration purposes; they are not exactly to scale. **c**, **d**, Density profiles (red) of the localized BEC one second after release, in linear (**c**) and semi-log (**d**) coordinates. In the inset in **d** we display the root-mean-square (rms) width of the profile versus time t, with ($V_R \neq 0$) and without ($V_R = 0$) disordered potential. This shows that the stationary regime is reached after 0.5 s. The diamond at $t = 1$ s corresponds to the data shown in **c** and the main panel of **d**. Blue lines in **c** are exponential fits to the wings, and correspond to the straight blue lines in **d**. The narrow central profiles (pink) represent the trapped condensate before release ($t = 0$ s).

Transversely, the correlation function (an ellipse with semi-axis lengths of 97 μm and 10 μm) is much wider than the atomic matter wave, and we can therefore consider the disorder as being one-dimensional for the BEC expanding along the z direction. Along this direction, the correlation function of the disordered potential is $V_R^2(\sin(z/\sigma_R)/(z/\sigma_R))^2$, where the correlation length $\sigma_R = 0.26 \pm 0.03$ μm (±1 s.e.m.) is calculated knowing the numerical aperture of the optics, and V_R is the amplitude of the disorder. The corresponding speckle grain size is $\pi\sigma_R = 0.82$ μm. The power spectrum of this speckle potential is non-zero only for **k**-vectors lower in magnitude than a cutoff of $2/\sigma_R$. The amplitude of the disorder is directly proportional to the laser intensity[22]. The calibration factor is calculated knowing the geometry of the optical system and the properties of the rubidium-87 atoms.

When we switch off the longitudinal trapping in the presence of weak disorder, the BEC starts expanding, but the expansion rapidly stops, in stark contrast with the free expansion case (Fig. 1d inset, showing the evolution of the root-mean-square width of the observed profiles). Plots of the density profile in linear (Fig. 1c) and semi-log (Fig. 1d) coordinates then show clear exponential wings, a signature of Anderson localization. Our observations are made in a regime allowing Anderson localization, unlike in the experiments in refs 19 and 20. First, the disorder is weak enough ($V_R/\mu_{in} = 0.12$) that the initial interaction energy per atom is rapidly converted into a kinetic energy of the order of μ_{in} for atoms in the wings. This value is much greater than the amplitude of the disordered potential, so there is no possibility of a classical reflection from a potential barrier. Second, the atomic density in the wings is low enough (two orders of magnitude less than in the initial BEC) that the interaction energy is negligible in comparison with the atom kinetic energy. Last, we fulfil the criterion, emphasized in ref. 13, that the atomic matter wave **k**-vector distribution be bounded, with a maximum magnitude k_{max} of less than half the cutoff in the power spectrum of the speckle disordered potential used here, that is, $k_{max}\sigma_R < 1$. The value of k_{max} is measured directly by observing the free expansion of the BEC in the waveguide in the absence of disorder (see Methods). For the runs corresponding to Figs 1c, 1d, 2, and 3, we have $k_{max}\sigma_R = 0.65 \pm 0.09$ (±2 s.e.m.).

An exponential fit to the wings of the density profiles yields the localization length L_{loc}, which we can compare to the theoretical value[13]

$$L_{loc} = \frac{2\hbar^4 k_{max}^2}{\pi m^2 V_R^2 \sigma_R (1 - k_{max}\sigma_R)} \quad (1)$$

valid only for $k_{max}\sigma_R < 1$ (m is the atomic mass). To ensure that the comparison is meaningful, we first check that we have reached a stationary situation, in which the fitted value of L_{loc} no longer evolves, as shown in Fig. 2. In Fig. 3, we plot the variation of L_{loc} with the disorder amplitude V_R, for the same number of atoms, that is, the same k_{max}. The dash–dot line is a plot of equation (1) for the values of k_{max} and σ_R determined as explained above. It shows quite a good agreement between our measurements and the theoretical predictions: with no adjustable parameters we obtain the correct magnitude and general shape. The shaded area reflects the envelope of the dash–dot line when we take into account the uncertainties in σ_R and k_{max}. The uncertainty in the calibration of V_R does not appear in Fig. 3. We estimate it to be no greater than 30%, which does not affect the agreement between theory and experiment.

An intriguing result of ref. 13 is the prediction of density profiles with algebraic wings when $k_{max}\sigma_R > 1$, that is, when the initial interaction energy is great enough that a fraction of the atoms have a **k**-vector greater in magnitude than $1/\sigma_R$, which plays the role of an effective mobility edge. We investigate this regime by repeating the experiment with a BEC containing more atoms (1.7×10^5 atoms, $\mu_{in}/h = 519$ Hz), for $V_R/\mu_{in} = 0.15$. Figure 4a shows the observed density profile in such a situation ($k_{max}\sigma_R = 1.16 \pm 0.14$ (±2 s.e.m.)), and a log–log plot suggests a power-law decrease in the wings, with an

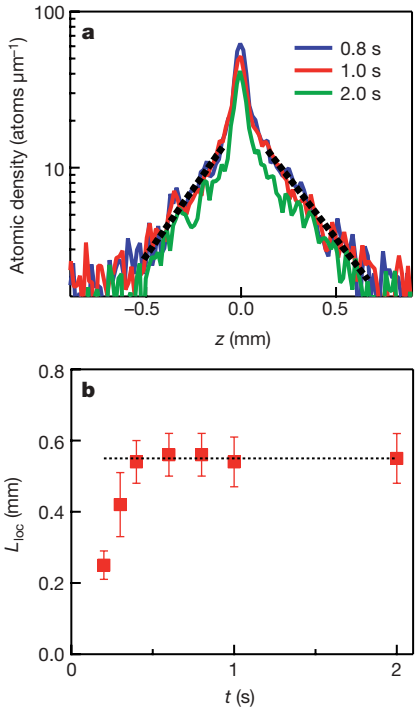


Figure 2 | Stationarity of the localized profile. **a**, Three successive density profiles, from which the localization length L_{loc} is extracted by fitting an exponential, $\exp(-2|z|/L_{loc})$ (dotted black lines), to the atomic density in the wings. **b**, Localization length L_{loc} versus expansion time t. Error bars, 95% confidence intervals for the fitted values (± 2 s.e.m.).

exponent of 1.95 ± 0.10 (± 2 s.e.m.), in agreement with the theoretical prediction that density decreases like $1/z^2$ in the wings. The semi-log plot (inset) confirms that an exponential would not work as well. For comparison, we present in Fig. 4b a log–log plot and a semi-log plot (inset) for the case with $k_{max}\sigma_R = 0.65$ and $V_R/\mu_{in} = 0.15$, where we conclude in favour of exponential rather than algebraic tails. These data support the existence of a crossover from an exponential to an algebraic regime in our speckle potential.

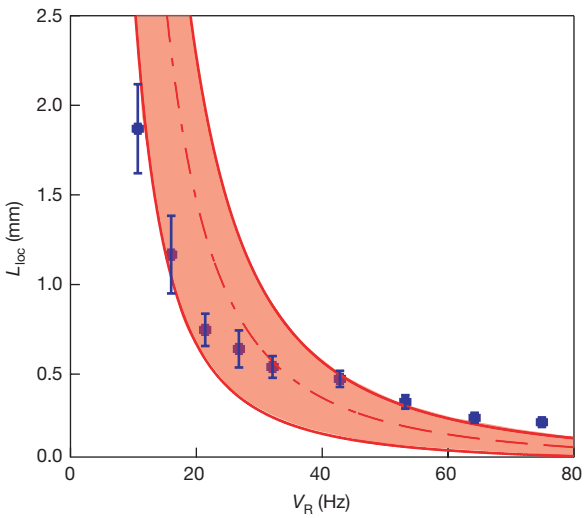


Figure 3 | Localization length versus amplitude of the disordered potential. L_{loc} is obtained by an exponential fit to the wings of the stationary localized density profiles, as shown in Fig. 2. Error bars, 95% confidence intervals for the fitted values (± 2 s.e.m.); 1.7×10^4 atoms ; $\mu_{in} = 219$ Hz. The dash–dot line is plotted using equation (1), where k_{max} is determined from the observed free expansion of the condensate (see Methods). The shaded area represents uncertainty associated with the evaluations of k_{max} and σ_R. We note that the limited extension of the disordered potential (4 mm) allows us to measure values of L_{loc} up to about 2 mm.

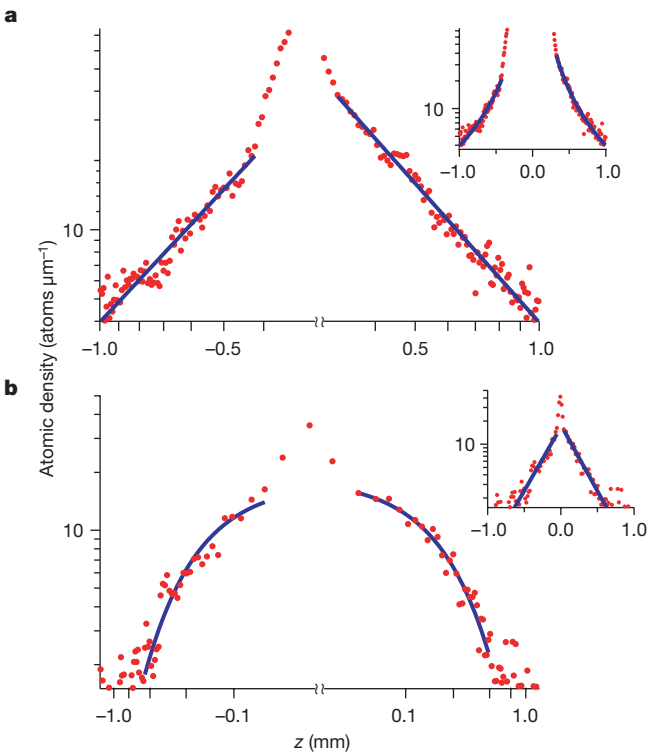


Figure 4 | Algebraic and exponential regimes in a one-dimensional speckle potential. Log–log and semi-log plots of the stationary atomic density profiles, showing the difference between the algebraic ($k_{max}\sigma_R > 1$) and exponential ($k_{max}\sigma_R < 1$) regimes. **a**, Density profile for $V_R/\mu_{in} = 0.15$ and $k_{max}\sigma_R = 1.16 \pm 0.14$ (± 2 s.e.m.). The momentum distribution of the released BEC has components beyond the effective mobility edge $1/\sigma_R$. The fit to the wings with a power-law decay $1/|z|^\beta$ yields $\beta = 1.92 \pm 0.06$ (± 2 s.e.m.) for the left-hand wing and $\beta = 2.01 \pm 0.03$ (± 2 s.e.m.) for the right-hand wing. The inset shows the same data in a semi-log plot, and confirms the non-exponential decay. **b**, For comparison, a similar set of plots (log–log and semi-log) in the exponential regime for the same $V_R/\mu_{in} = 0.15$ and $k_{max}\sigma_R = 0.65 \pm 0.09$ (± 2 s.e.m.).

Direct imaging of atomic quantum gases in controlled, optical disordered potentials is a promising technique to investigate a variety of open questions on disordered quantum systems. First, as in other problems of condensed matter simulated using ultracold atoms, direct imaging of atomic matter waves offers unprecedented possibilities to measure important properties, such as localization lengths. Second, our experiment can be extended to quantum gases with controlled interactions where localization of quasi-particles[26,27], Bose glass[14,15,28] and Lifshits glass[29] are expected, as well as to Fermi gases and to Bose–Fermi mixtures where rich phase diagrams have been predicted[30]. The reasonable quantitative agreement between our measurements and the theory of one-dimensional Anderson localization in a speckle potential demonstrates the high degree of control in our set-up. We thus anticipate that it can be used as a quantum simulator for investigating Anderson localization in higher dimensions[31,32], first to look for the mobility edge of the Anderson transition, and then to measure important features at the Anderson transition that are still under theoretical investigation, such as critical exponents. It will also become possible to investigate the effect of controlled interactions on Anderson localization.

METHODS SUMMARY

Momentum distribution of the expanding BEC. To compare measured localization lengths with those calculated from equation (1), we need to know k_{max}, the maximum amplitude of the **k**-vector distribution of the atoms, at the beginning of the expansion in the disordered potential. We measure k_{max} by releasing a BEC with the same number of atoms in the waveguide without disorder, and observing the density profiles at various times t. Density profiles are readily

converted into **k**-vector distributions ($|\mathbf{k}| = \hbar^{-1} m \mathrm{d}z/\mathrm{d}t$). The key step in obtaining $k_{\max}$ is accurately determining the position $z_{\max}$ of the front edge of the profile. To do this, we fit the whole profile to an inverted parabola, which is the expected shape for the one-dimensional expansion of a BEC in the fundamental transverse mode of the waveguide. Actually, the BEC has an initial transverse profile that is slightly enlarged owing to interactions between atoms, but its density rapidly decreases during the expansion, and a numerical calculation using our experimental parameters shows that for expansion times greater than $t = 0.2$ s, an inverted parabola correctly approximates the atomic density profile and allows accurate determination of the front edge position. Using this procedure, we measure $z_{\max}$ every 0.1 s from $t = 0$ to $t = 1$ s, and find it to be proportional to t for $t > 0.2$ s. We estimate the uncertainty in $k_{\max}$ to about 6% and 9% for 1.7×10^5 atoms and 1.7×10^4 atoms, respectively.

Received 14 March; accepted 9 April 2008.

1. Anderson, P. W. Absence of diffusion in certain random lattices. *Phys. Rev.* **109,** 1492–1505 (1958).
2. Van Tiggelen, B. in *Wave Diffusion in Complex Media 1998* (ed. Fouque, J. P.) 1–60 (Kluwer, Dordrecht, 1999).
3. Wiersma, D. S., Bartolini, P., Lagendijk, A. & Righini, R. Localization of light in a disordered medium. *Nature* **390,** 671–673 (1997).
4. Scheffold, F., Lenke, R., Tweer, R. & Maret, G. Localization or classical diffusion of light? *Nature* **398,** 206–270 (1999).
5. Störzer, M., Gross, P., Aegerter, C. M. & Maret, G. Observation of the critical regime near Anderson localization of light. *Phys. Rev. Lett.* **96,** 063904 (2006).
6. Schwartz, T., Bartal, G., Fishman, S. & Segev, M. Transport and Anderson localization in disordered two-dimensional photonic lattices. *Nature* **446,** 52–55 (2007).
7. Lahini, Y. *et al.* Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. *Phys. Rev. Lett.* **100,** 013906 (2008).
8. Dalichaouch, R., Armstrong, J. P., Schultz, S., Platzman, P. M. & McCall, S. L. Microwave localization by two-dimensional random scattering. *Nature* **354,** 53–55 (1991).
9. Chabanov, A. A., Stoytchev, M. & Genack, A. Z. Statistical signatures of photon localization. *Nature* **404,** 850–853 (2000).
10. Weaver, R. L. Anderson localization of ultrasound. *Wave Motion* **12,** 129–142 (1990).
11. Akkermans, E. & Montambaux, G. *Mesoscopic Physics of Electrons and Photons* (Cambridge Univ. Press, Cambridge, UK, 2006).
12. Goodman, J. W. *Speckle Phenomena in Optic*s (Roberts, Greenwood Village, Colorado, 2007).
13. Sanchez-Palencia, L. *et al.* Anderson localization of expanding Bose-Einstein condensates in random potentials. *Phys. Rev. Lett.* **98,** 210401 (2007).
14. Giamarchi, T. & Schulz, H. J. Anderson localization and interactions in one-dimensional metals. *Phys. Rev. B* **37,** 325–340 (1988).
15. Fisher, M. P. A., Weichman, P. B., Grinstein, G. & Fisher, D. S. Boson localization and the superfluid-insulator transition. *Phys. Rev. B* **40,** 546–570 (1989).
16. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold atoms. *Rev. Mod. Phys.* (in the press); preprint at ⟨http://arxiv.org/abs/0704.3011⟩ (2007).
17. Lewenstein, M. *et al.* Ultracold atomic gases in optical lattices: mimicking condensed matter physics and beyond. *Adv. Phys.* **56,** 243–379 (2007).
18. Damski, B., Zakrzewski, J., Santos, L., Zoller, P. & Lewenstein, M. Atomic Bose and Anderson glasses in optical lattices. *Phys. Rev. Lett.* **91,** 080403 (2003).
19. Clément, D. *et al.* Suppression of transport of an interacting elongated Bose–Einstein condensate in a random potential. *Phys. Rev. Lett.* **95,** 170409 (2005).
20. Fort, C. *et al.* Effect of optical disorder and single defects on the expansion of a Bose–Einstein condensate in a one-dimensional waveguide. *Phys. Rev. Lett.* **95,** 170410 (2005).
21. Schulte, T. *et al.* Routes towards Anderson-like localization of Bose–Einstein condensates in disordered optical lattices. *Phys. Rev. Lett.* **95,** 170411 (2005).
22. Clément, D. *et al.* Experimental study of the transport of coherent interacting matter-waves in a 1D random potential induced by laser speckle. *N. J. Phys.* **8,** 165, doi:10.1088/1367-2630/8/8/165 (2006).
23. Moore, J. L., Robinson, J. C., Bharucha, C., Williams, P. E. & Raizen, M. G. Observation of dynamical localization in atomic momentum transfer: A new testing ground for quantum chaos. *Phys. Rev. Lett.* **73,** 2974–2977 (1994).
24. Chabé, J. *et al.* Experimental observation of the Anderson transition with atomic matter waves. Preprint available at ⟨http://arxiv.org/abs/0709.4320⟩ (2007).
25. Guerin, W. *et al.* Guided quasicontinuous atom laser. *Phys. Rev. Lett.* **97,** 200402 (2006).
26. Bilas, N. & Pavloff, N. Anderson localization of elementary excitations in a one-dimensional Bose–Einstein condensate. *Eur. Phys. J. D* **40,** 387–397 (2006).
27. Lugan, P., Clément, D., Bouyer, P., Aspect, A. & Sanchez-Palencia, L. Anderson localization of Bogolyubov quasiparticles in interacting Bose–Einstein condensates. *Phys. Rev. Lett.* **99,** 180402 (2007).
28. Fallani, L., Lye, J. E., Guarrera, V., Fort, C. & Inguscio, M. Ultracold atoms in a disordered crystal of light: towards a Bose glass. *Phys. Rev. Lett.* **98,** 130404 (2007).
29. Lugan, P. *et al.* Ultracold Bose gases in 1D-disorder: from Lifshits glass to Bose–Einstein condensate. *Phys. Rev. Lett.* **98,** 170403 (2007).
30. Sanpera, A., Kantian, A., Sanchez-Palencia, L., Zakrzewski, J. & Lewenstein, M. Atomic Fermi–Bose mixtures in inhomogeneous and random optical lattices: from Fermi glass to quantum spin glass and quantum percolation. *Phys. Rev. Lett.* **93,** 040401 (2004).
31. Kuhn, R. C., Miniatura, C., Delande, D., Sigwarth, O. & Müller, C. A. Localization of matter waves in two-dimensional disordered optical potentials. *Phys. Rev. Lett.* **95,** 250403 (2005).
32. Skipetrov, S. E., Minguzzi, A., van Tiggelen, B. A. & Shapiro, B. Anderson localization of a Bose-Einstein condensate in a 3D random potential. *Phys. Rev. Lett.* **100,** 165301 (2008).

Acknowledgements: The authors are indebted to P. Chavel, T. Giamarchi, M. Lewenstein and G. Shlyapnikov for many discussions, to P. Georges and G. Roger for assistance with the laser, and to F. Moron, A. Villing and G. Colas for technical assistance on the experimental apparatus. This research was supported by the Centre National de la Recherche Scientifique (CNRS), the Délégation Générale de l'Armement (DGA), the Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT), the Agence Nationale de la Recherche (ANR), the Institut Francilien de Recherche sur les Atomes Froids (IFRAF) and IXSEA; by the STREP programme FINAQS of the European Union; and by the programme QUDEDIS of the European Science Foundation (ESF).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.B. (philippe.bouyer@institutoptique.fr).

LETTERS

Anderson localization of a non-interacting Bose–Einstein condensate

Giacomo Roati[1,2], Chiara D'Errico[1,2], Leonardo Fallani[1,2], Marco Fattori[1,2,3], Chiara Fort[1,2], Matteo Zaccanti[1,2], Giovanni Modugno[1,2], Michele Modugno[1,4,5] & Massimo Inguscio[1,2]

Anderson localization of waves in disordered media was originally predicted[1] fifty years ago, in the context of transport of electrons in crystals[2]. The phenomenon is much more general[3] and has been observed in a variety of systems, including light waves[4,5]. However, Anderson localization has not been observed directly for matter waves. Owing to the high degree of control over most of the system parameters (in particular the interaction strength), ultracold atoms offer opportunities for the study of disorder-induced localization[6]. Here we use a non-interacting Bose–Einstein condensate to study Anderson localization. The experiment is performed with a one-dimensional quasi-periodic lattice—a system that features a crossover between extended and exponentially localized states, as in the case of purely random disorder in higher dimensions. Localization is clearly demonstrated through investigations of the transport properties and spatial and momentum distributions. We characterize the crossover, finding that the critical disorder strength scales with the tunnelling energy of the atoms in the lattice. This controllable system may be used to investigate the interplay of disorder and interaction (ref. 7 and references therein), and to explore exotic quantum phases[8,9].

The transition between extended and localized states originally studied by Anderson for non-interacting electrons has not been directly observed in crystals, owing to the high electron–electron and electron–phonon interactions[2]. Researchers have therefore turned their attention to systems where interactions or nonlinearities are almost absent. Evidence of the Anderson localization for light waves in disordered media has been provided by an observed modification of the classical diffusive regime, featuring a conductor–insulator transition[4,5]. However, a clear understanding of the interplay between disorder and nonlinearity is considered crucial in contemporary condensed matter physics. First effects of weak nonlinearities have been recently shown in experiments with light waves in photonic lattices[10,11]. The combination of ultracold atoms and optical potentials offers a novel platform for the study of disorder-related phenomena where most of the relevant physical parameters, including those governing interactions, can be controlled[6,8]. The introduction of laser speckles[12] and quasi-periodic optical lattices[9] has made possible the investigation of the physics of disorder. The investigations reported so far have explored either quantum phases induced by interaction[9] or regimes of weak interaction where the observation of Anderson localization was precluded either by the size of the disorder or by delocalizing effects of nonlinearity[12–16].

In this work, we use a Bose–Einstein condensate in which the interactions can be tuned independently of the other parameters[17], to study localization due purely to disorder. We study localization in a one-dimensional lattice perturbed by a second, weak incommensurate lattice, which constitutes an experimental realization of the non-interacting Harper[18] or Aubry–André model[19]. This quasi-periodic system displays a transition from extended to localized states analogous to the Anderson transition, already in one dimension[20,21], whereas in the case of pure random disorder, more than two dimensions would be needed[22]. We clearly observe this transition by studying transport and both spatial and momentum distributions, and we verify the scaling behaviour of the critical disorder strength. Our system is described by the Aubry–André hamiltonian

$$H = J\sum_m (|w_m\rangle\langle w_{m+1}| + |w_{m+1}\rangle\langle w_m|) + \Delta\sum_m \cos(2\pi\beta m + \phi)|w_m\rangle\langle w_m| \quad (1)$$

where $|w_m\rangle$ is the Wannier state localized at the lattice site m, J is the site-to-site tunnelling energy, Δ is the strength of the disorder, $\beta = k_2/k_1$ is the ratio of the two lattice wave numbers, and ϕ is an arbitrary phase. In the experiment, the two relevant energies J and Δ (see Fig. 1a) can be controlled independently by changing the heights of the primary and secondary lattice potentials, respectively. For a maximally incommensurate ratio $\beta = (\sqrt{5}-1)/2$, the model exhibits a sharp transition from extended to localized states at $\Delta/J = 2$ (refs 18, 19, 21). For the actual experimental parameters, $\beta = 1.1972\ldots$ and the transition is broadened and shifted towards larger values of Δ/J (see Fig. 1b). Owing to the quasi-periodic nature of the potential, these localized states appear approximately every five sites (2.6 μm).

The non-interacting Bose–Einstein condensate is prepared by sympathetically cooling a cloud of interacting ^{39}K atoms in an optical trap, and then tuning the *s*-wave scattering length almost to zero by means of a Feshbach resonance[17,23] (see Methods). The spatial size of the condensate can be controlled by changing the harmonic confinement provided by the trap. For most of the measurements the size along the direction of the lattice is $\sigma \approx 5\,\mu$m. The quasi-periodic potential is imposed by using two lasers in a standing-wave configuration[16]. The gaussian shape of the laser beams forming the primary lattice also provides radial confinement of the condensate in the absence of the harmonic trap.

In our first experiment we investigated transport, by abruptly switching off the main harmonic confinement and letting the atoms expand along the one-dimensional bichromatic lattice. We detect the spatial distribution of the atoms at increasing evolution times using absorption imaging (see Fig. 2a). In a regular lattice ($\Delta = 0$) the eigenstates of the potential are extended Bloch states, and the system expands ballistically. In the limit of large disorder ($\Delta/J > 7$) we observe no diffusion, because in this regime the condensate can be described as the superposition of several localized eigenstates whose individual extensions are less than the initial size of the condensate. In the crossover between these two regimes we observe a ballistic expansion with reduced speed. This crossover is summarized in Fig. 2b, which, for a fixed evolution time of 750 ms, shows the width of the atomic distribution versus the rescaled disorder strength Δ/J,

[1]LENS and Dipartimento di Fisica, Università di Firenze, 50019 Sesto Fiorentino, Italy. [2]INFM-CNR, 50019 Sesto Fiorentino, Italy. [3]Museo Storico della Fisica e Centro Studi e Ricerche 'E. Fermi', 00184 Roma, Italy. [4]Dipartimento di Matematica Applicata, Università di Firenze, 50139 Firenze, Italy. [5]BEC-INFM Center, Università di Trento, 38050 Povo, Italy.

for three different values of J. In all three cases, the system enters the localized regime at the same disorder strength, providing compelling evidence of the scaling behaviour intrinsic to the model described in equation (1).

In this regime, the eigenstates of the hamiltonian in equation (1) are exponentially localized, and the tails of diffusing wave packets are expected to behave like stretched exponentials[24]. We therefore analysed the tails of the spatial distributions with an exponential function of the form $f_\alpha(x) = A\exp(-|(x-x_0)/l|^\alpha)$, the exponent α being a fitting parameter. Two examples of this analysis, one for weak disorder and one for strong disorder, are shown in Fig. 3a, b. The exponent α exhibits a smooth crossover from a value of two to a value of one as Δ/J increases (Fig. 3c), signalling the onset of an exponential localization. The value $\alpha = 2$ that we obtain for small Δ/J corresponds to the expected ballistic evolution of the initial gaussian momentum distribution of the non-interacting condensate. We note that in the radial direction, where the system is only harmonically trapped, the spatial distribution is always well fitted by a gaussian function ($\alpha = 2$).

Information on the eigenstates of the system can also be extracted from the analysis of the momentum distribution of the stationary atomic states in the presence of the harmonic confinement. The width of the axial momentum distribution $P(k)$ is inversely proportional to the spatial extent of the condensate in the lattice. We measure it by releasing the atoms from the lattice and imaging them after a ballistic expansion.

In Fig. 4, we show examples of the experimental momentum distributions that are in agreement with the model predictions for the low-lying eigenstates. Without disorder, we observe the typical grating interference pattern with three peaks at $k = 0, \pm 2k_1$, reflecting the periodicity of the primary lattice. The very small width of the peak at $k = 0$ indicates that the wavefunction is spread over many lattice sites[25]. For weak disorder, the eigenstates of the hamiltonian in equation (1) are still extended, and additional momentum peaks appear at momentum space distances $\pm 2(k_1 - k_2)$ from the main peaks, corresponding to the beating of the two lattices. As we further increase Δ/J, $P(k)$ broadens and its width eventually becomes comparable with that of the Brillouin zone, k_1, indicating that the extension of the localized states becomes comparable with the lattice spacing. From the theoretical analysis of the Aubry–André model, we have a clear indication that in this regime the eigenstates are exponentially localized on individual lattice sites.

We note that the side peaks in the two bottom profiles of Fig. 4a, b indicate that the localization is non-trivial, that is, the tails of the eigenstates extend over several lattice sites even for large disorder. The small modulation on top of the profiles is due to the interference between the several localized states over which the condensate is distributed. In Fig. 4c, we present the root-mean-squared width of the central peak of $P(k)$ as a function of Δ/J, for three different values of J. The three data sets lie on the same line, confirming the scaling behaviour of the system. A visibility of the interference pattern,

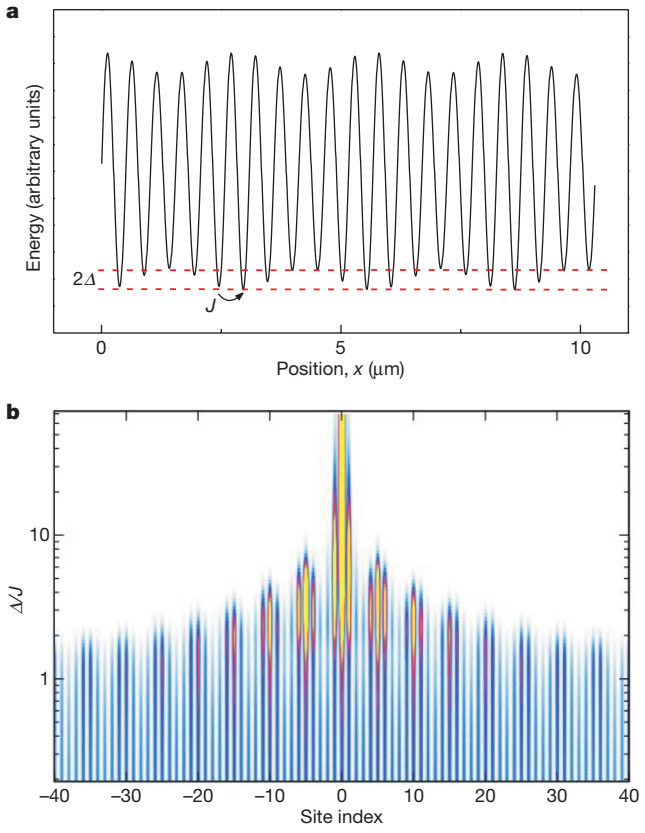


Figure 1 | The quasi-periodic optical lattice. a, The quasi-periodic potential realized in the experiment. The hopping energy J describes the tunnelling between different sites of the primary lattice and 2Δ is the maximum shift of the on-site energy induced by the secondary lattice. The lattice constant is 516 nm. **b**, Typical calculated density plot of a low-lying eigenstate of the bichromatic potential, as a function of Δ/J (vertical axis). For small values of Δ/J the state is delocalized over many lattice sites. For $\Delta/J \geq 7$ the state becomes exponentially localized on lengths smaller than the lattice constant.

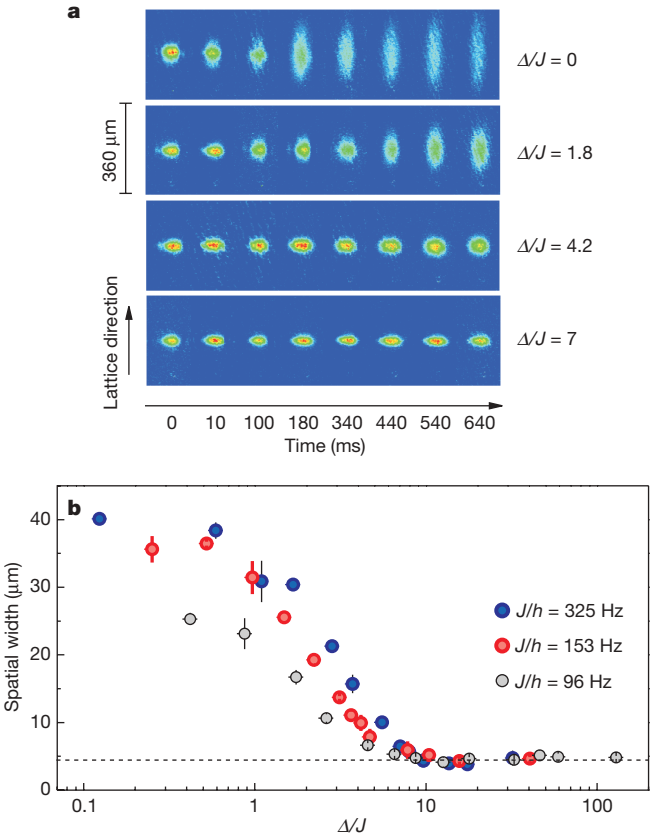


Figure 2 | Probing the localization with transport. a, *In situ* absorption images of the Bose–Einstein condensate diffusing along the quasi-periodic lattice for different values of Δ and $J/h = 153$ Hz (where h denotes Planck's constant). For $\Delta/J > 7$ the size of the condensate remains at its original value, reflecting the onset of localization. **b**, Root-mean-squared size of the condensate for three different values of J, at a fixed evolution time of 750 ms, versus the rescaled disorder strength Δ/J. The dashed line indicates the initial size of the condensate. The onset of localization appears in the same range of values of Δ/J in all three cases. Vertical error bars, 95% confidence level (± 2 s.e.m.); horizontal error bars, 10% uncertainty due to the nonlinearity of the modulators' response.

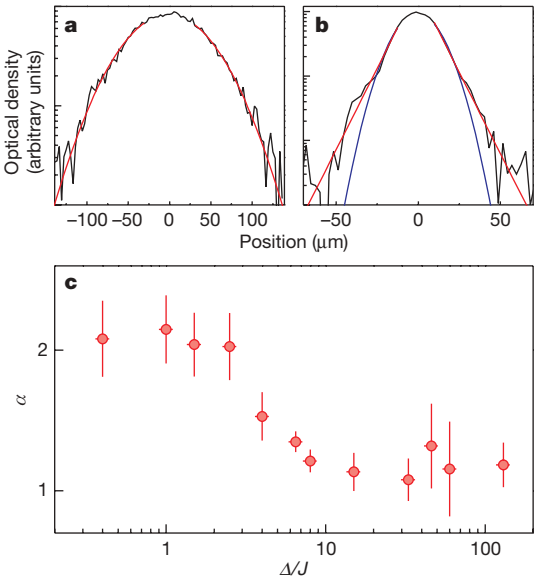
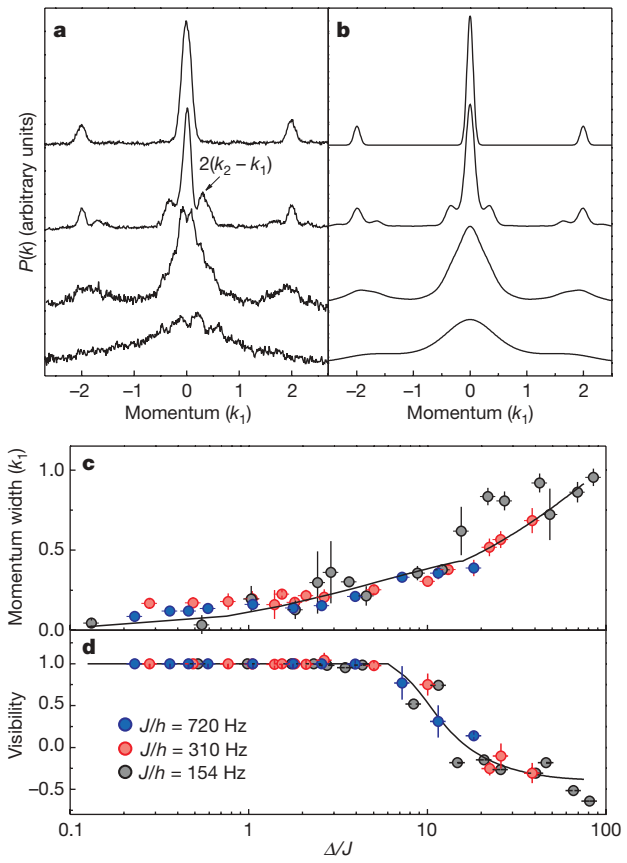


Figure 3 | Observing the nature of the localized states. **a**, **b**, Experimental profiles and fitting function $f_\alpha(x)$ (red) for $\Delta/J \approx 1$ (**a**) and $\Delta/J \approx 15$ (**b**). Note the vertical log scale. The blue line in **b** represents a gaussian fit, $\alpha = 2$. **c**, Dependence of the fitting parameter α on Δ/J, indicating a transition from a gaussian to an exponential distribution. Vertical error bars, 95% confidence level (± 2 s.e.m.); horizontal error bars, 10% uncertainty due to the nonlinearity of the modulators' response.

$V = (P(2k_1) - P(k_1))/(P(2k_1) + P(k_1))$, can be defined to highlight the appearance of a finite population in the momentum states $\pm k_1$ and, therefore, the onset of exponential localization with an extension comparable with the lattice spacing. In Fig. 4d, we show the visibility extracted from the same data as Fig. 4c. Experiment and theory are again in good agreement, and feature a sudden decrease in the visibility for $\Delta/J \approx 6$.



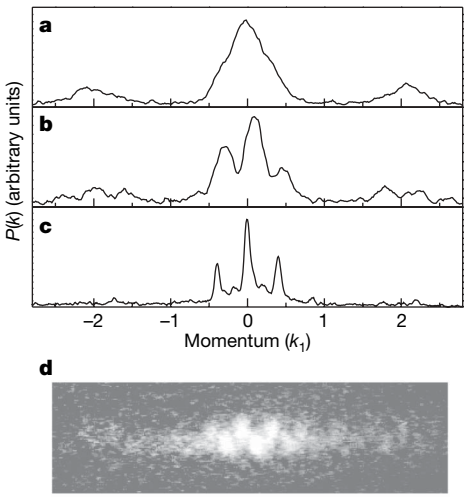


Figure 5 | Interference of localized states. Momentum distribution of the condensate prepared in a disordered lattice with $\Delta/J \approx 10$, for different values of the harmonic confinement. **a**, Profile of a single localized state (initial spatial size of the condensate, $\sigma = 1.2$ μm); **b**, interference of two localized states ($\sigma = 1.2$ μm); **c**, interference of three localized states ($\sigma = 2.1$ μm). Momentum is measured along the horizontal axes in units of k_1. **d**, Dislocated interference pattern resulting from the interference of two localized states, suggesting the presence of a thermally activated vortex.

Further information on the localized states can be extracted from the interference of a small number of them. This regime can be reached in the experiment by simply reducing the spatial extent of the condensate through an increase of the harmonic confinement. Typical profiles of $P(k)$ are displayed in Fig. 5a–c. Depending on the degree of confinement, we observe one, two or three states, featuring a smooth distribution or a clear multiple-slit interference pattern. The spacing of the fringes yields a spatial separation between the localized states of about five sites, as expected. The independent localized states have a quasi-two-dimensional geometry, because their axial extents are much smaller than their radial extents. This feature makes our system an excellent testing ground in which to study the physics of quasi-two-dimensional systems[26], which were recently investigated using widely spaced optical lattices[27]. We also observe (Fig. 5d) interference patterns which present a dislocation, possibly produced by thermal activation of a vortex in one of the two localized states[27], but in our case for non-interacting atoms.

In this work we have observed Anderson localization of coherent non-interacting matter waves. Future studies might reveal how a weak, controllable interaction affects the observed localization transition. More generally, the high theoretical and experimental control possible in this system makes it a novel platform for the study of exotic quantum phases arising from the interplay between interaction and disorder[6,8,28] (see also ref. 29 and references therein).

Figure 4 | Momentum distribution. **a**, **b**, Experimental and theoretical momentum distributions $P(k)$ for increasing Δ/J values (0, 1.1, 7.2 and 25, from top to bottom). The interference pattern of a regular lattice observed at $\Delta = 0$ is at first modified by the appearance of peaks at the beating between the two lattices, and then increasingly broadened. Momentum is measured along the horizontal axes in units of k_1. **c**, Root-mean-squared size of the central peak of $P(k)$ versus Δ/J, for three different values of J. The experimental data follow a unique scaling behaviour, as predicted by theory (continuous line). The width of the peak is measured in units of k_1. **d**, Visibility V of the interference pattern versus Δ/J. In both experiment and theory (continuous line), V decreases abruptly for $\Delta/J \approx 6$, indicating localization on distances comparable to the lattice period. Vertical error bars, 95% confidence level (± 2 s.e.m.); horizontal error bars, 10% uncertainty due to the nonlinearity of the modulators' response.

METHODS SUMMARY

Non-interacting Bose–Einstein condensate. The condensate of ^{39}K atoms was produced at a large positive value of the *s*-wave scattering length ($a \approx 180a_0$, $a_0 = 0.529 \times 10^{-10}$ m) at a magnetically tuneable Feshbach resonance[17,30]. The scattering length was then adiabatically reduced to values of the order of $0.1a_0$ (ref. 23), corresponding to an atom–atom interaction energy of $U \approx 10^{-5}J$ (ref. 6).

Quasi-periodic optical lattice. The quasi-periodic potential was created by superimposing two standing waves created with laser beams of wavelengths $\lambda_1 \approx 1{,}032$ nm (primary lattice) and $\lambda_2 \approx 862$ nm (secondary lattice), with waists of 150 μm. The two lattice depths were adjusted by varying the intensity of the beams, and calibrated by Bragg diffraction. The heights s_i of the lattices ($i = 1, 2$), in units of the recoil energies $E_{R,i} = h^2/(2M\lambda_i^2)$, are $s_1 \leq 10$ and $s_2 \leq 3$ (where M is the atomic mass). The estimated relative uncertainty at a 95% confidence level on Δ/J is 15%. An additional uncertainty of 10% due to the nonlinearity of the modulators' response is shown in Figs 2–4 as a horizontal error bar.

Analysing the atomic distributions. The atomic samples were imaged on a charge-coupled-device camera with a spatial resolution of about 5 μm. The images analysed in Figs 2 and 3 were recorded just after the release from the trap. The analysis of the profiles of Fig. 3 was carried out by integrating the atomic density distributions along the radial direction. The top 20% of the signal was dropped and only the remaining tails were fitted. This takes into account the possible gaussian broadening effects of the transfer function of the imaging system and the fact that we populate few localized states. The profiles in Figs 4 and 5 were obtained after a long ballistic expansion over 25 ms to reduce the contribution of the initial spatial distribution. The analysis of typically four or five different images was used to extract individual data points in Figs 2–4. The vertical error bars in Figs 2–4 correspond to a 95% confidence level (± 2 s.e.m.).

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

Received 31 March; accepted 6 May 2008.

1. Anderson, P. W. Absence of diffusion in certain random lattices. *Phys. Rev.* **109,** 1492–1505 (1958).
2. Lee, P. A. & Ramakrishnan, T. V. Disordered electronic systems. *Rev. Mod. Phys.* **57,** 287–337 (1985).
3. Kramer, B. & MacKinnon, A. Localization: theory and experiment. *Rep. Prog. Phys.* **56,** 1469–1564 (1993).
4. Van Albada, M. P. & Lagendijk, A. Observation of weak localization of light in a random medium. *Phys. Rev. Lett.* **55,** 2692–2695 (1985).
5. Wiersma, D. S., Bartolini, P., Lagendijk, A. & Righini, R. Localization of light in a disordered medium. *Nature* **390,** 671–673 (1997).
6. Damski, B., Zakrzewski, J., Santos, L., Zoller, P. & Lewenstein, M. Atomic Bose and Anderson glasses in optical lattices. *Phys. Rev. Lett.* **91,** 080403 (2003).
7. Dubi, Y., Meir, Y. & Avishai, Y. Nature of the superconductor-insulator transition in disordered superconductor. *Nature* **449,** 876–880 (2007).
8. Lewenstein, M. *et al.* Ultracold atomic gases in optical lattices: mimicking condensed matter physics and beyond. *Adv. Phys.* **56,** 243–379 (2007).
9. Fallani, L., Lye, J. E., Guarrera, V., Fort, C. & Inguscio, M. Ultracold atoms in a disordered crystal of light: towards a Bose glass. *Phys. Rev. Lett.* **98,** 130404 (2007).
10. Schwartz, T., Bartal, G., Fishman, S. & Segev, M. Transport and Anderson localization in disordered two-dimensional photonic lattices. *Nature* **446,** 52–55 (2007).
11. Lahini, Y. *et al.* Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. *Phys. Rev. Lett.* **100,** 013906 (2008).
12. Lye, J. E. *et al.* Bose-Einstein condensate in a random potential. *Phys. Rev. Lett.* **95,** 070401 (2005).
13. Clément, D. *et al.* Suppression of transport of an interacting elongated Bose-Einstein condensate in a random potential. *Phys. Rev. Lett.* **95,** 170409 (2005).
14. Fort, C. *et al.* Effect of optical disorder and single defects on the expansion of a Bose-Einstein condensate in a one-dimensional waveguide. *Phys. Rev. Lett.* **95,** 170410 (2005).
15. Schulte, T. *et al.* Routes towards Anderson-like localization of Bose-Einstein condensates in disordered optical lattices. *Phys. Rev. Lett.* **95,** 170411 (2005).
16. Lye, J. E. *et al.* Effect of interactions on the localization of a Bose-Einstein condensate in a quasi-periodic lattice. *Phys. Rev. A* **75,** 061603 (2007).
17. Roati, G. *et al.* ^{39}K Bose-Einstein condensate with tunable interactions. *Phys. Rev. Lett.* **99,** 010403 (2007).
18. Harper, P. G. Single band motion of conduction electrons in a uniform magnetic field. *Proc. Phys. Soc. A* **68,** 874–878 (1955).
19. Aubry, S. & André, G. Analyticity breaking and Anderson localization in incommensurate lattices. *Ann. Israel Phys. Soc.* **3,** 133–140 (1980).
20. Grempel, D. R., Fishman, S. & Prange, R. E. Localization in an incommensurate potential: an exactly solvable model. *Phys. Rev. Lett.* **49,** 833–836 (1982).
21. Aulbach, C., Wobst, A., Ingold, G.-L., Hänggi, P. & Varga, I. Phase-space visualization of a metal–insulator transition. *New J. Phys.* **6,** doi:10.1088/1367-2630/6/1/070 (2004).
22. Abrahams, E., Anderson, P. W., Licciardello, D. C. & Ramakrishnan, T. V. Scaling theory of localization: absence of quantum diffusion in two dimensions. *Phys. Rev. Lett.* **42,** 673–676 (1979).
23. Fattori, M. *et al.* Atom interferometry with a weakly interacting Bose-Einstein condensate. *Phys. Rev. Lett.* **100,** 080405 (2008).
24. Zhong, J. *et al.* Shape of the quantum front. *Phys. Rev. Lett.* **86,** 2485–2489 (2001).
25. Pedri, P. *et al.* Expansion of a coherent array of Bose-Einstein condensates. *Phys. Rev. Lett.* **87,** 220401 (2001).
26. Burger, S. *et al.* Quasi-2D Bose-Einstein condensation in an optical lattice. *Europhys. Lett.* **57,** 1–6 (2002).
27. Hadzibabic, Z., Krüger, P., Cheneau, M., Battelier, B. & Dalibard, J. Berezinskii-Kosterlitz-Thouless crossover in a trapped atomic gas. *Nature* **441,** 1118–1121 (2006).
28. Lugan, P. *et al.* Ultracold Bose gases in 1D disorder: from Lifshits glass to Bose-Einstein condensate. *Phys. Rev. Lett.* **98,** 170403 (2007).
29. Roux, G. *et al.* The quasi-periodic Bose-Hubbard model and localization in 1-dimensional cold atomic gases. Preprint at ⟨http://arxiv.org/abs/0802.3774⟩ (2008).
30. D'Errico, C. *et al.* Feshbach resonances in ultracold ^{39}K. *New J. Phys.* **9,** doi:10.1088/1367-2630/9/7/223 (2007).

Acknowledgements We thank J. Dalibard for discussions, S. Machluf for contributions, and all the colleagues of the Quantum Gases group at LENS. This work has been supported by MIUR, EU (IP SCALA), ESF (DQS–EuroQUAM), INFN and Ente CRF.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.I. (inguscio@lens.unifi.it).

METHODS

Non-interacting Bose–Einstein condensate. The condensate was prepared in a homogeneous magnetic field of about 396 G, where a broad Feshbach resonance raised the value of the *s*-wave scattering length from the background value of $-29a_0$ to about $180a_0$ (refs 17, 30). This allowed the efficient formation of a stable condensate. The condensate was trapped in a crossed dipole trap with an average harmonic frequency of 100 Hz, and contained about 10^5 atoms. The scattering length was then reduced by shifting the magnetic field to about 350 G, at the zero-crossing position. This magnetic field was adiabatically changed with a combined linear and exponential ramp lasting 110 ms, to avoid shape excitations of the cloud. We estimate a residual scattering length of the order of $0.1a_0$, limited by magnetic field instability (100 mG) and by the contribution to the scattering of higher-order partial waves.

Quasi-periodic lattice. The quasi-periodic lattice was created by superimposing two standing waves of incommensurate wavelengths. In particular, the primary lattice was generated by a single-mode Yb:YAG laser of wavelength $\lambda_1 = 1{,}032$ nm, whose linewidth and intensity were actively stabilized, and the secondary lattice was obtained by a single-mode Ti:Sapphire laser of wavelength $\lambda_2 = 862$ nm. The lattice beams were focused on the condensate with a beam waist of about 150 μm. In the absence of the harmonic trap, in the radial plane, the atoms felt a harmonic confinement of about 40 Hz due to the gaussian intensity profile of the laser beams of the primary lattice. The optical power in each of the two standing waves was independently adjusted by means of two acousto-optic modulators, and the lattice heights were calibrated by means of Bragg diffraction[31]. In the experiment, the primary lattice depth ranged up to 10 $E_{R,1}$, and that of the secondary lattice to 3 $E_{R,2}$, where $E_{R,i} = h^2/(2M\lambda_i^2)$ are the recoil energies ($i = 1, 2$). Before loading the condensate into the quasi-periodic lattice, we adiabatically reduced the harmonic trap frequency to 5 Hz by decreasing the intensity of the dipole trap beams. The lattices were then raised by means of 's'-shaped ramps, on a timescale of 100 ms.

Theory. Since our three-dimensional condensate is non-interacting, the problem is separable. Along the direction of the bichromatic lattice, the system is described by the single-particle hamiltonian $H_{1D} = -(h^2/(8\pi^2 M))\partial^2/\partial x^2 + s_1 E_{R,1}\cos^2(k_1 x) + s_2 E_{R,2}\cos^2(k_2 x)$, where $k_i = 2\pi/\lambda_i$ ($i = 1, 2$) are the lattice wavenumbers and s_i are the heights of the two lattices in units of the recoil energies $E_{R,i}$. In the tight-binding limit, the above hamiltonian can be mapped to the Aubry–André hamiltonian in equation (1), with $\Delta = s_2 E_{R,2}/(2E_{R,1})$ and $J \approx 1.43 {s_1}^{0.98} \exp(-2.07\sqrt{s_1})$ (ref. 32). Fig. 1b represents a typical density plot of a lowest lying eigenstate of the bichromatic potential, obtained by direct diagonalization of the full hamiltonian H_{1D}. The plots in Fig. 4b also include the effect of the harmonic trapping potential $V_{ho}(x) = M\omega^2 x^2/2$, with $\omega = 2\pi \times 5$ Hz.

Data analysis. The root-mean-squared widths presented in Fig. 2b were obtained by integrating the two-dimensional distributions of Fig. 2a in the radial direction, and by fitting the axial distribution with gaussian profiles. The analysis shown in Fig. 3 consisted of fitting only the tails of the axial distributions with an exponential decay of the form $f_z(x) = A\exp(-|(x - x_0)/l|^\alpha)$, leaving the exponent α as a fitting parameter. We exclude from the fit the central part of the distributions, within 0.6 times the root-mean-squared width from the centre. For our data, this choice represents the best compromise between dropping an expected gaussian component at the centre and having a sufficiently large dynamical range to fit the tails. The analysis of the data in Fig. 3 was based on the average of 4–5 different experimental images. All experimental momentum distribution profiles in Figs 4 and 5 were obtained by radial integration of the two-dimensional distributions. For the analysis in Fig. 4c, we directly measured the root-mean-squared width of the central peak when this could be resolved from the side peaks, or we fitted the overall distribution with three gaussian profiles to extract the width of the central peak. This procedure was applied to both experiment and theory.

31. Ovchinnikov, YuB *et al.* Diffraction of a released Bose-Einstein condensate by a pulsed standing light wave. *Phys. Rev. Lett.* **83**, 284–287 (1999).
32. Gerbier, F. *et al.* Interference pattern and visibility of a Mott insulator. *Phys. Rev. A* **72**, 053606 (2005).

LETTERS

Magnetic order close to superconductivity in the iron-based layered $LaO_{1-x}F_xFeAs$ systems

Clarina de la Cruz[1,2], Q. Huang[3], J. W. Lynn[3], Jiying Li[3,4], W. Ratcliff II[3], J. L. Zarestky[5], H. A. Mook[2], G. F. Chen[6], J. L. Luo[6], N. L. Wang[6] & Pengcheng Dai[1,2]

Following the discovery of long-range antiferromagnetic order in the parent compounds of high-transition-temperature (high-T_c) copper oxides[1,2], there have been efforts to understand the role of magnetism in the superconductivity that occurs when mobile 'electrons' or 'holes' are doped into the antiferromagnetic parent compounds. Superconductivity in the newly discovered rare-earth iron-based oxide systems ROFeAs (R, rare-earth metal) also arises from either electron[3–7] or hole[8] doping of their non-superconducting parent compounds. The parent material LaOFeAs is metallic but shows anomalies near 150 K in both resistivity and d.c. magnetic susceptibility[3]. Although optical conductivity and theoretical calculations suggest that LaOFeAs exhibits a spin-density-wave (SDW) instability that is suppressed by doping with electrons to induce superconductivity[9], there has been no direct evidence of SDW order. Here we report neutron-scattering experiments that demonstrate that LaOFeAs undergoes an abrupt structural distortion below 155 K, changing the symmetry from tetragonal (space group *P*4/*nmm*) to monoclinic (space group *P*112/*n*) at low temperatures, and then, at ~137 K, develops long-range SDW-type antiferromagnetic order with a small moment but simple magnetic structure[9]. Doping the system with fluorine suppresses both the magnetic order and the structural distortion in favour of superconductivity. Therefore, like high-T_c copper oxides, the superconducting regime in these iron-based materials occurs in close proximity to a long-range-ordered antiferromagnetic ground state.

The recent discovery of superconductivity in $RO_{1-x}F_xFeAs$ (refs 3–7) has generated enormous interest because these materials are the first non-copper oxide superconductors with T_cs exceeding 50 K. Because these superconductors are derived by doping their non-superconducting parent compounds, it is natural to wonder what the ground states of the parent compounds are. It has been argued theoretically that non-superconducting LaOFeAs is either a non-magnetic metal near a magnetic (antiferromagnetic and/or ferromagnetic) instability[10–12] or an antiferromagnetic semimetal[9,13–15]. As a function of temperature, the resistivity of LaOFeAs shows a clear drop around 150 K before increasing again below 50 K (refs 3, 9). The d.c. magnetic susceptibility also has a small anomaly near 150 K. From optical measurements and theoretical calculations[9], it has been argued that LaOFeAs has an antiferromagnetic SDW instability below 150 K and that superconductivity in these materials arises from the suppression of this SDW order.

We used neutron diffraction to study the structural and magnetic order in polycrystalline, non-superconducting LaOFeAs and superconducting $LaO_{1-x}F_xFeAs$ with $x = 0.08$ ($T_c = 26$ K). Our experiments were carried out on the BT-1 powder diffractometer and the BT-7 thermal triple-axis spectrometer at the NIST Center for Neutron Research, and on the HB-1A triple-axis spectrometer at the High Flux Isotope Reactor, Oak Ridge National Laboratory. Figure 1a shows the high-resolution neutron powder diffraction data obtained using the BT-1 diffractometer and our refinements for non-superconducting LaOFeAs at 170 K. Consistent with earlier results[3–9], we find that the crystal structure belongs to the tetragonal space group *P*4/*nmm*, with atomic positions given in Table 1. On cooling the sample to 4 K, the (2, 2, 0) reflection that has a single peak at 170 K (Fig. 1a, inset) is split into two peaks (Fig. 1b, inset). This suggests that a structural phase transition has occurred. For comparison, we note that the (2, 2, 0) peak remains a single peak even at 10 K (Fig. 1c inset) for superconducting $LaO_{0.92}F_{0.08}FeAs$. To understand the low-temperature structural distortion in LaOFeAs, we carried out refinements using the neutron data and found that the structure in fact becomes monoclinic and belongs to space group *P*112/*n* (see Supplementary Information). Table 2 summarizes the low-temperature lattice parameters and atomic positions for LaOFeAs. Table 3 summarizes the lattice parameters and atomic positions for superconducting $LaO_{0.92}F_{0.08}FeAs$ at 10 K, 35 K and 170 K. Figure 1d shows the LaOFeAs structure.

To see if the newly observed structural transition is related to the ~150 K resistivity anomaly, we carried out detailed temperature-dependent measurements of the (2, 2, 0) reflection and found that there is an abrupt splitting of the (2, 2, 0) peak at 155 K (Fig. 2). The peak intensity also shows a clear kink at 155 K. These results indicate that the non-superconducting system has a structural phase transition, and this phase transition is associated with the observed resistivity[3] and specific heat anomalies[9]. Because a similar splitting of the (2, 2, 0) peak is absent in superconducting $LaO_{0.92}F_{0.08}FeAs$ (Fig. 1c), we can safely assume that this transition is suppressed with the

Table 1 | Properties of LaOFeAs at 175 K

a, Refined structure parameters

Atom	Site	*x*	*y*	*z*	*B* (Å^2)
La	2*c*	¼	¼	0.1418(3)	0.65(7)
Fe	2*b*	¾	¼	½	0.39(5)
As	2*c*	¼	¼	0.6507(4)	0.23(8)
O	2*a*	¾	¼	0	0.69(7)

b, Selected interatomic distances and angles

La–As (×4)	3.378(1) Å	Fe–As (×4)	2.407(2) Å
La–O (×4)	2.365(2) Å	As–Fe–As	107.41(7)°
Fe–Fe	2.84969(7) Å	As–Fe–As	113.7(1)°

Space group, *P*4/*nmm*; $a = 4.03007(9)$ Å; $c = 8.7368(2)$ Å; $V = 141.898(9)$ Å^3. $R_p = 5.24\%$, weighted $R_p = 6.62\%$, $\chi^2 = 0.9821$.

[1]Department of Physics and Astronomy, The University of Tennessee, Knoxville, Tennessee 37996-1200, USA. [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [3]NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-6102, USA. [4]Department of Materials Science and Engineering, University of Maryland, College Park, Maryland 20742-6393, USA. [5]Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA. [6]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100080, China.

appearance of superconductivity in $LaO_{1-x}F_xFeAs$ through doping with fluorine.

It would be interesting to see whether this phase transition in non-superconducting LaOFeAs is indeed associated with SDW order. Figure 3a shows our raw data for LaOFeAs collected on the BT-7 spectrometer at 8 K and 170 K. Inspection of the figure immediately reveals that there are extra peaks in the low-temperature spectrum at wavevector magnitudes $Q = 1.15$, 1.53 and 2.5 Å^{-1}, which are not

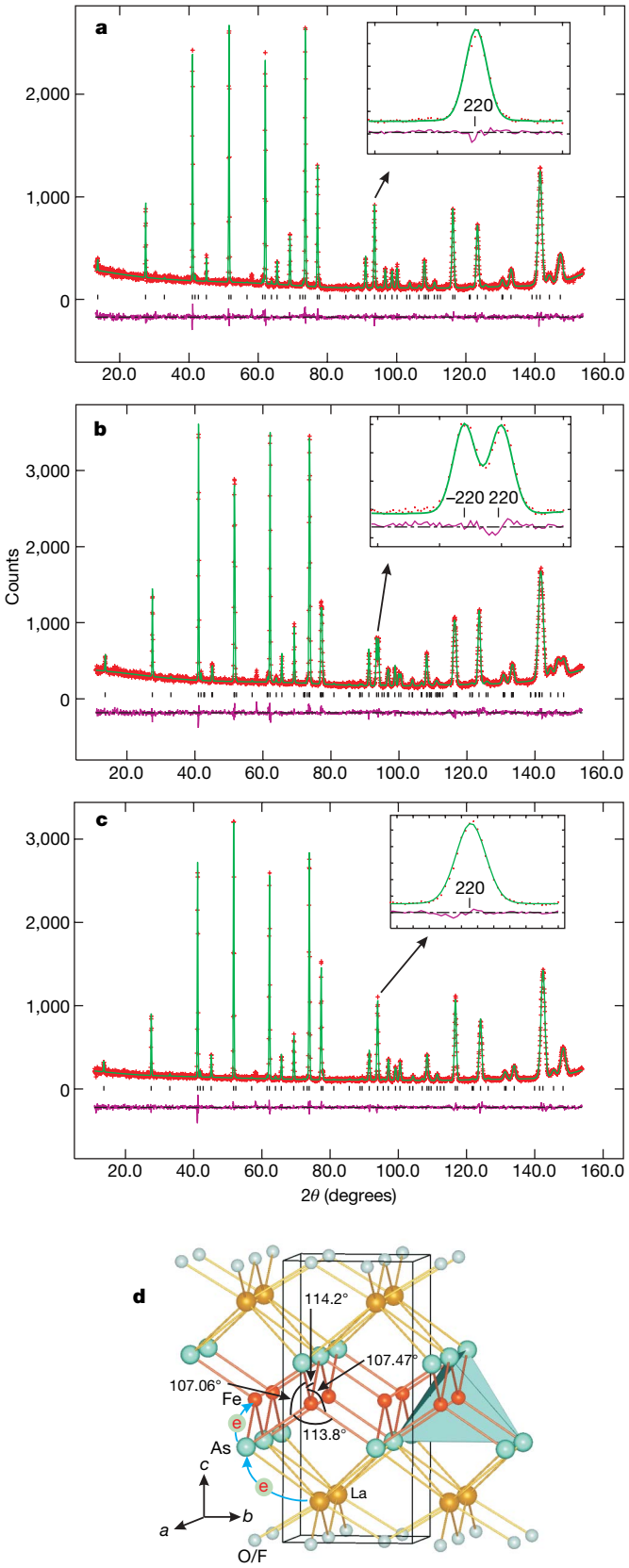


Figure 1 | Temperature dependence of the nuclear structures for LaOFeAs and $LaO_{0.92}F_{0.08}FeAs$. We prepared ~2 g each of LaOFeAs and $LaO_{0.92}F_{0.08}FeAs$ using the method described in ref. 9. The BT-1 diffractometer has a Ge(3, 1, 1) monochromator and an incident beam wavelength of $\lambda = 2.0785$ Å. **a**, Observed (red crosses) and calculated (green solid line) neutron powder diffraction intensities of LaOFeAs at 175 K using space group *P*4/*nmm*. The inset shows a single peak of the (2, 2, 0) reflection. Short black vertical lines show the Bragg peak positions. The purple trace indicates the intensity difference between the observed and calculated structures. θ, diffraction angle. **b**, The same scan at 4 K, where the (2, 2, 0) reflection is split owing to the monoclinic distortion. The fit was made using space group *P*112/*n*. **c**, The 10 K scan of superconducting $LaO_{0.92}F_{0.08}FeAs$; the space group *P*4/*nmm* can describe data at all measured temperatures. **d**, Crystal structure of $LaO_{1-x}F_xFeAs$. For $x = 0$, the compound has the charge-balance configuration $La^{3+}O^{2-}Fe^{2+}As^{3-}$. Electron doping can be achieved by replacing O with F.

Table 2 | Properties of LaOFeAs at 4 K

a, Refined structure parameters

Atom	Site	*x*	*y*	*z*	*B* (Å^2)
La	2*e*	¼	¼	0.1426(3)	0.54(6)
Fe	2*f*	¾	¼	0.5006(12)	0.16(4)
As	2*e*	¼	¼	0.6499(4)	0.23(7)
O	2*f*	¾	¼	−0.0057(17)	0.69(7)

b, Selected interatomic distances and angles

La–As (×2)	3.369(1) Å	Fe–As (×2)	2.398(6) Å
La–As (×2)	3.380(1) Å	Fe–As (×2)	2.405(6) Å
La–O (×2)	2.394(8) Å	As–Fe–As	114.2(5)°
La–O (×2)	2.342(7) Å	As–Fe–As	107.47(6)°
Fe–Fe	2.8409(2) Å	As–Fe–As	107.06(6)°
Fe–Fe	2.8548(2) Å	As–Fe–As	113.8(4)°

Space group, *P*112/*n*; $a = 4.0275(2)$ Å; $b = 4.0275(2)$ Å; $c = 8.7262(5)$ Å; $\gamma = 90.279(3)°$; $V = 141.54(2)$ Å^3. $R_p = 4.31\%$, weighted $R_p = 5.74\%$, $\chi^2 = 1.100$. Lattice constants *a* and *b* were constrained to be equal in the final refinement. For *z*(Fe) = 0.5 and *z*(O) = 0, the structure symmetry can also be described as an orthorhombic structure with *Cmma* space group (see Supplementary Information).

Table 3 | Properties of $LaO_{0.92}F_{0.08}FeAs$ at 10 K (first line), 35 K (second line) and 175 K (third line)

a, Refined structure parameters

Atom	Site	*x*	*y*	*z*	*B* (Å^2)
La	2*c*	¼	¼	0.1448(3)	0.40(5)
		¼	¼	0.1458(3)	0.50(5)
		¼	¼	0.1446(3)	0.73(5)
Fe	2*b*	¾	¼	½	0.32(4)
		¾	¼	½	0.41(4)
		¾	¼	½	0.65(4)
As	2*c*	¼	¼	0.6521(4)	0.41(7)
		¼	¼	0.6515(4)	0.40(6)
		¼	¼	0.6527(4)	0.69(7)
O/F	2*a*	¾	¼	0	0.53(6)
		¾	¼	0	0.62(6)
		¾	¼	0	0.71(6)

b, Selected interatomic distances and angles

La–As (×4)	3.347(1) Å	Fe–As (×4)	2.407(2) Å
	3.345(1) Å		2.404(2) Å
	3.349(1) Å		2.412(2) Å
La–O/F (×4)	2.373(2) Å	As–Fe–As	107.61(6)°
	2.377(1) Å		107.52(6)°
	2.373(1) Å		107.72(6)°
Fe–Fe	2.8427(1) Å	As–Fe–As	113.3(1)°
	2.8423(1) Å		113.5(1)°
	2.8446(1) Å		113.0(1)°

Space group, *P*4/*nmm*. At 10 K, $a = 4.0202(1)$ Å, $c = 8.7034(2)$ Å, $V = 140.66(1)$ Å^3, $R_p = 5.34\%$, weighted $R_p = 6.95\%$, $\chi^2 = 1.028$. At 35 K, $a = 4.0196(1)$ Å, $c = 8.7027(2)$ Å, $V = 140.61(1)$ Å^3, $R_p = 5.38\%$, weighted $R_p = 6.96\%$, $\chi^2 = 1.050$. At 175 K, $a = 4.0229(1)$ Å, $c = 8.7142(2)$ Å, $V = 141.03(1)$ Å^3, $R_p = 5.30\%$, weighted $R_p = 6.93\%$, $\chi^2 = 0.9882$. The small lattice parameter and cell volume differences between 10 K and 35 K data are within the uncertainties of our measurements.

present at 170 K. Indexing these peaks (Fig. 3a) indicates that these reflections are indeed directly related to the nuclear structure and are due to magnetic scattering arising from a simple stripe-type antiferromagnetic structure of iron moments with a magnetic cell $\sqrt{2}a_N \times \sqrt{2}b_N \times 2c_N$ (Fig. 4, top-right inset), where a_N, b_N, and c_N are the nuclear lattice parameters of the unit cell (see Table 1). Figure 3c shows our refinements considering both the magnetic and structural unit cell. Normalizing the magnetic intensity to the nuclear scattering, we find an ordered iron moment of $0.36(5)\mu_B$ at 8 K (in our results, figures in parentheses indicate uncertainty in the final decimal place; μ_B denotes the Bohr magneton). For comparison, Fig. 3b shows the HB-1A spectrometer data for LaOFeAs and $LaO_{0.92}F_{0.08}FeAs$. For non-superconducting LaOFeAs, the temperature difference spectrum (8–170 K) shows a clear peak at $Q = 1.53$ Å^{-1}, which corresponds to the magnetic (1, 0, 3) Bragg peak. This peak is absent from the scan of superconducting $LaO_{0.92}F_{0.08}FeAs$.

To see if the observed magnetic scattering at low temperature in LaOFeAs is indeed associated with the 150 K phase transition, we carried out order parameter measurements on the strongest (1, 0, 3) magnetic peak using both the BT-7 and the HB-1A spectrometers. Figure 4 shows the temperature dependence of the square of the ordered magnetic moment (normalized at low temperature), which vanishes at ~137 K, about ~18 K lower than the temperature at which the structural phase transition occurs (Fig. 2). Surprisingly, the magnetic order is established at lower temperatures than the structural distortion, much as spin ordering is established after the charge ordering in the static stripe-ordered copper oxide material $La_{1.6-x}Nd_{0.4}Sr_xCuO_4$ with $x = 0.12$ (ref. 16). The presence of the lattice distortion above the Néel temperature is established conclusively in the bottom-left inset of Fig. 4, where a clear lattice distortion is apparent at 138 K. Therefore, the resistivity anomaly at 150 K is caused by the structural distortion, not SDW ordering as originally suggested[9]. The top-right inset in Fig. 4 shows the magnetic structure we determined for the system.

To summarize, we have discovered that the parent compound of the iron-based superconductors is a long-range-ordered antiferromagnet with a simple stripe-type antiferromagnetic structure within the plane that is doubled along the c axis (Fig. 4, top-right inset). There is a structural phase transition before the antiferromagnetic phase transition that changes the structure from space group $P4/nmm$ to space group $P112/n$ at low temperature. The magnetic structure is consistent with the theoretical prediction[9], but the moment of $0.36(5)\mu_B$ per iron atom that we observe at 8 K is much smaller than the predicted value of $\sim 2.3\mu_B$ per iron atom (refs 13, 14). The presence of magnetic frustration might induce the reduced ordered moment[17]. The disappearance of the static antiferromagnetic order and lattice distortion in the doped superconducting materials suggests that the underlying physical properties of this class of superconductors may have important similarities to the high-T_c copper oxides. In any case, we are confident that this new class of materials will open new avenues of research regardless of the origin of the electron pairing and superconductivity.

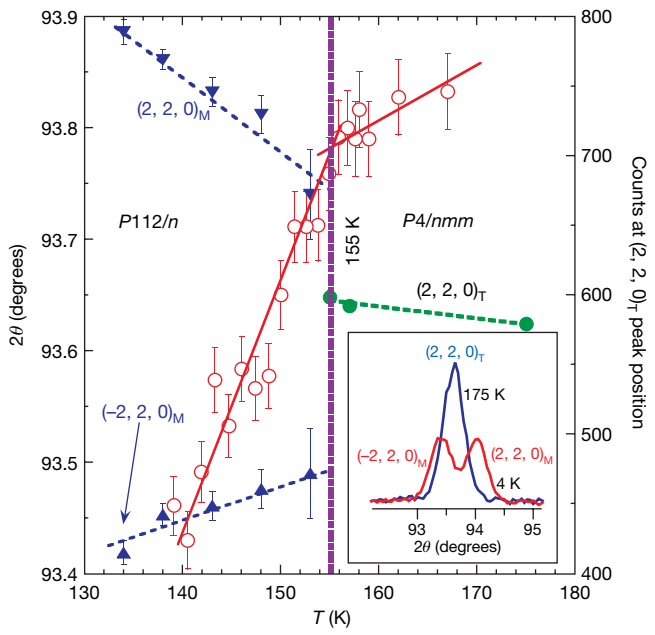


Figure 2 | Temperature dependence of the (2, 2, 0) nuclear reflection indicative of a structural phase transition at ~155 K in LaOFeAs. Peak intensities at the $(2, 2, 0)_T$ (tetragonal) reflection (open symbols, right-hand scale) and positions of the $(2, 2, 0)_T$, $(-2, 2, 0)_M$ (monoclinic) and $(2, 2, 0)_M$ peaks (solid symbols, left-hand scale) as a function of temperature on cooling. A structural transition from tetragonal symmetry $P4/nmm$ to monoclinic symmetry $P112/n$ occurs at ~155 K. Error bars, 1 s.d. Inset shows the $(2, 2, 0)_T$ reflection at 175 K and the $(-2, 2, 0)_M$ and $(2, 2, 0)_M$ reflections at 4 K.

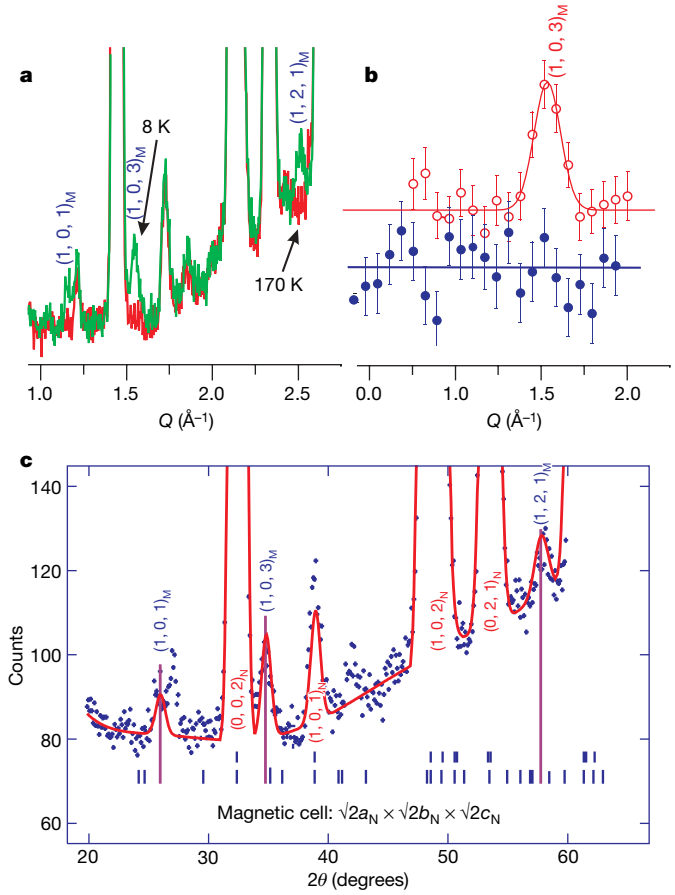


Figure 3 | Temperature dependence of the magnetic scattering for LaOFeAs and $LaO_{0.92}F_{0.08}FeAs$. **a**, LaOFeAs data clearly showing (marked) magnetic peaks at 8 K that disappear at 170 K, counting 1 min per point. **b**, The temperature difference spectra (8–170 K) measured using the HB-1A spectrometer for LaOFeAs (red) and $LaO_{0.92}F_{0.08}FeAs$ (blue), counting 4 min per point. The magnetic (1, 0, 3) peak is missing from the $LaO_{0.92}F_{0.08}FeAs$ scan. Error bars, 1 s.d. **c**, LaOFeAs data again showing both magnetic and nuclear Bragg peaks (data, crosses; Bragg peak positions, short vertical lines) together with the model fit (solid line), at 8 K. Data in **a** and **c** were collected using the BT-7 spectrometer with an incident beam wavelength of $\lambda = 2.44$ Å, a PG(0, 0, 2) monochromator and a PG (pyrolytic graphite) filter. Data in **b** were collected using the HB-1A spectrometer with $\lambda = 2.36$ Å and a PG filter.

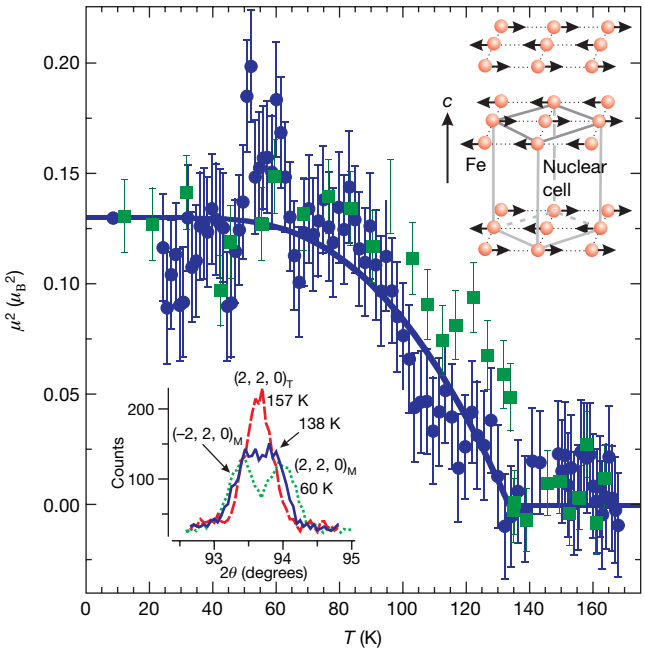


Figure 4 | Temperature dependence of the order parameter at $Q = 1.53$ Å^{-1}, obtained using the magnetic structure we determine for LaOFeAs. Blue circles, BT-7 spectrometer data; green squares, HB-1A spectrometer data. The solid line is a simple fit to mean field theory that gives a Néel temperature $T_N = 137(3)$ K. The bottom-left inset shows the temperature dependence of the nuclear (2, 2, 0) peak obtained using the BT-1 diffractometer. It is clear that the lattice is distorted at 138 K, before the long-range static antiferromagnetic order sets in at ~137 K. The top-right inset shows the antiferromagnetic structure of the system, giving a $\sqrt{2}a_N \times \sqrt{2}b_N \times 2c_N$ unit cell with moment directions parallel to the planes of iron atoms. To determine the magnetic structure, we note that three magnetic peaks with $h + k + l = 2n$, where h, k and l are Miller indices and $n = 0, 1, 2, \ldots$, suggest that the spin configuration has a body-centred symmetry. Our refinements including the spin direction with a c-axis component revealed that the c component converged to 0. Because of the very small difference (0.002 Å) in the a and b lattice constants in the orthorhombic magnetic unit cell (see Supplementary Information), it was not possible to determine the spin direction in the a–b plane. Error bars, 1 s.d.

Received 1 April; accepted 2 May 2008.
Published online 28 May 2008.

1. Vaknin, D. *et al.* Antiferromagnetism in La_2CuO_{4-y}. *Phys. Rev. Lett.* **58,** 2802–2805 (1987).
2. Tranquada, J. M. *et al.* Neutron-diffraction determination of antiferromagnetic structure of Cu ions in $YBa_2Cu_3O_{6+x}$ with x = 0.0 and 0.15. *Phys. Rev. Lett.* **60,** 156–159 (1988).
3. Kamihara, Y., Watanabe, T., Hirano, M. & Hosono, H. Iron-based layered superconductor La[$O_{1-x}F_x$]FeAs ($x = 0.05$–0.12) with $T_c = 26$ K. *J. Am. Chem. Soc.* **130,** 3296–3297 (2008).
4. Chen, G. F. *et al.* Superconducting properties of Fe-based layered superconductor $LaO_{0.9}Fe_{0.1-\delta}FeAs$. Preprint at ⟨http://arxiv.org/abs/0803.0128v1⟩ (2008).
5. Chen, X. H. *et al.* Superconductivity at 43 K in samarium-arsenic oxides SmFeAs$O_{1-x}F_x$. *Nature* advance online publication, doi:10.1038/nature07045 (25 May 2008).
6. Chen, G. F. *et al.* Superconductivity at 41 K and its competition with spin-density-wave instability in layered $CeO_{1-x}Fe_xFeAs$. *Phys. Rev. Lett.* (in the press); preprint at ⟨http://arxiv.org/abs/0803.3790v2⟩ (2008).
7. Ren, Z.-A. *et al.* Superconductivity at 52 K in iron-based F-doped layered quaternary compound Pr[$O_{1-x}F_x$]FeAs. *Mater. Res. Innov.* (in the press); preprint at ⟨http://arxiv.org/abs/0803.4283v1⟩ (2008).
8. Wen, H. H., Mu, G., Fang, L., Yang, H. & Zhu, X. Y. E. Superconductivity at 25 K in hole-doped ($La_{1-x}Sr_x$)OFeAs. *Europhys. Lett.* **82,** 17009 (2008).
9. Dong, J. *et al.* competing orders and spin-density-wave instability in La($O_{1-x}Fe_x$)FeAs. Preprint at ⟨http://arxiv.org/abs/0803.3426v1⟩ (2008).
10. Singh, D. J. & Du, M. H. $LaFeAsO_{1-x}F_x$: A low carrier density superconductor near itinerant magnetism. Preprint at ⟨http://arxiv.org/abs/0803.0429v1⟩ (2008).
11. Xu, G. *et al.* Doping-dependent phase diagram of LaOMAs (M = V–Cu) and electron-type superconductivity near ferromagnetic instability. *Europhys. Lett.* (in the press); preprint at ⟨http://arxiv.org/abs/0803.1282v2⟩ (2008).
12. Haule, K., Shim, J. H. & Kotliar, G. Correlated electronic structure of $LaO_{1-x}F_xFeAs$. *Phys. Rev. Lett.* (in the press); preprint at ⟨http://arxiv.org/abs/0803.1279v1⟩ (2008).
13. Cao, C., Hirschfeld, P. J. & Cheng, H. P. Coexistence of antiferromagnetism with superconductivity in $LaO_{1-x}F_xFeAs$: effective Hamiltonian from ab initio studies. *Phys. Rev. B* (in the press); preprint at ⟨http://arxiv.org/abs/0803.3236v1⟩ (2008).
14. Ma, F. J. & Lu, Z. Y. Iron-based layered superconductor $LaO_{1-x}F_xFeAs$: an antiferromagnetic semimetal. Preprint at ⟨http://arxiv.org/abs/0803.3286v1⟩ (2008).
15. Han, Q., Chen, Y. & Wang, Z. D. A generic two-band model for unconventional superconductivity and spin-density-wave order in electron- and hole-doped iron-based superconductors. *Europhys. Lett.* **82,** 37007 (2008).
16. Tranquada, J. M., Sternlieb, B. J., Axe, J. D., Nakamura, Y. & Uchida, S. Evidence for stripe correlations of spins and holes in copper oxide superconductors. *Nature* **375,** 561–563 (1995).
17. Si, Q. & Abrahams, A. Strong correlations and magnetic frustration in the high T_c iron pnictides. Preprint at ⟨http://arxiv.org/abs/0804.2480v1⟩ (2008).

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

Acknowledgements We thank J. A. Fernandez-Baca, H. P. Cheng, T. Yildirim and C. Brown for discussions. This work is supported by the US Department of Energy, Division of Materials Science and Division of Scientific User Facilities, Basic Energy Sciences. This work is also supported by the US Department of Energy through UT/Battelle LLC. The work at the Institute of Physics, Chinese Academy of Sciences, is supported by the Natural Science Foundation of China, the Chinese Academy of Sciences and the Ministry of Science and Technology of China.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.D. (daip@ornl.gov).

LETTERS

Two-band superconductivity in $LaFeAsO_{0.89}F_{0.11}$ at very high magnetic fields

F. Hunte[1], J. Jaroszynski[1], A. Gurevich[1], D. C. Larbalestier[1], R. Jin[2], A. S. Sefat[2], M. A. McGuire[2], B. C. Sales[2], D. K. Christen[2] & D. Mandrus[2]

The recent synthesis of the superconductor $LaFeAsO_{0.89}F_{0.11}$ with transition temperature $T_c \approx 26$ K (refs 1–4) has been quickly followed by reports of even higher transition temperatures in related compounds: 41 K in $CeFeAsO_{0.84}F_{0.16}$ (ref. 5), 43 K in $SmFeAsO_{0.9}F_{0.1}$ (ref. 6), and 52 K in $NdFeAsO_{0.89}F_{0.11}$ and $PrFeAsO_{0.89}F_{0.11}$ (refs 7, 8). These discoveries have generated much interest[9,10] in the mechanisms and manifestations of unconventional superconductivity in the family of doped quaternary layered oxypnictides LnOTMPn (Ln: La, Pr, Ce, Sm; TM: Mn, Fe, Co, Ni; Pn: P, As), because many features of these materials set them apart from other known superconductors. Here we report resistance measurements of $LaFeAsO_{0.89}F_{0.11}$ at high magnetic fields, up to 45 T, that show a remarkable enhancement of the upper critical field B_{c2} compared to values expected from the slopes $dB_{c2}/dT \approx 2\,T\,K^{-1}$ near T_c, particularly at low temperatures where the deduced $B_{c2}(0) \approx 63$–65 T exceeds the paramagnetic limit. We argue that oxypnictides represent a new class of high-field superconductors with B_{c2} values surpassing those of Nb_3Sn, MgB_2 and the Chevrel phases, and perhaps exceeding the 100 T magnetic field benchmark of the high-T_c copper oxides.

The continuing search for new superconductors has recently yielded a new family of oxypnictides composed of alternating $LaO_{1-x}F_x$ and FeAs layers[1–4] with T_c values of 25–28 K, which can be raised to 40–43 K by replacing La with Ce (ref. 5) or Sm (ref. 6) or to 52 K by replacing La with Nd and Pr (refs 7, 8). Several experiments and band structure calculations suggest unconventional superconductivity in the paramagnetic Fe layer. First, *ab initio* calculations indicate that superconductivity originates from the *d* orbitals of what would normally be expected to be pair-breaking magnetic Fe ions, suggesting that new non-phonon pairing mechanisms are responsible for the high-T_c superconducting state[11,12]. Second, F-doped LaFeAsO is a semimetal, which exhibits strong antiferromagnetic fluctuations and a possible spin density wave instability around 150 K in the parent undoped LaFeAsO (refs 5, 13–17). And last, superconductivity may emerge on several disconnected pieces of the Fermi surface[11,12,18,19], thus exhibiting the multi-gap pairing that has recently attracted so much attention in MgB_2 (ref. 20).

Given the importance of magnetic correlations in the doped oxypnictides, transport measurements at very high magnetic fields are vital to probe the mechanisms of superconductivity. Indeed, first measurements of the upper critical magnetic field $B_{c2}(T)$ have yielded a slope $B'_{c2} = dB_{c2}/dT \approx 2\,T\,K^{-1}$ near T_c, for both La- and Sm-based oxypnictides[2–6]. From the conventional one-band Werthamer-Helfand-Hohenberg (WHH) theory[21], such slopes already imply rather high values of the upper critical magnetic field at zero temperature $B_{c2}(0) = 0.69\,T_c\,B'_{c2} \approx 36$ T for $LaFeAsO_{0.89}F_{0.11}$, ~59 T for $SmFeAsO_{0.9}F_{0.1}$, and ~72 T for $PrFeAsO_{0.89}F_{0.11}$, all well above $B_{c2}(0) \approx 30$ T of Nb_3Sn. However, studies of the high-field superconductivity in MgB_2 alloys have shown that the upward curvature of $B_{c2}(T)$ resulting from multiband effects can significantly increase $B_{c2}(0)$, as compared to the WHH one-band extrapolation (see ref. 22 and references therein). Our high-field d.c. transport measurements on $LaFeAsO_{0.89}F_{0.11}$ samples show that $B_{c2}(T)$ indeed exhibits signs of two-gap behaviour similar to that in MgB_2 with a value of $B_{c2}(0)$ that exceeds the WHH extrapolation by ~2 times and which also exceeds the BCS paramagnetic limit, B_p (in tesla) $= 1.84\,T_c$ (in kelvin) $= 51.5$ T for $T_c = 28$ K.

Polycrystalline $LaFeAsO_{0.89}F_{0.11}$ samples were made by solid state synthesis[4]. A sample $\sim 3 \times 1 \times 0.5\,mm^3$ was used for our four-probe transport measurements in the 45 T hybrid magnet at the National High Magnetic Field Laboratory, supplemented by low-field measurements in a 9 T superconducting magnet. Our low-field data agree well with earlier data taken at ORNL on the same sample[4], indicating its good temporal and atmospheric stability. The 45 T hybrid magnet was swept only from 11.5 T to 45 T owing to the static 11.5 T background field generated by the outer superconducting coil of the magnet combination, while lower fields were swept from 0 T to 9 T in a Physical Property Measurement System (PPMS) superconducting magnet with parameters shown in Fig. 1 legend.

The results of our high-field measurements of the sample resistance as a function of magnetic field strength, $R(B)$, are shown in Fig. 1. The broad $R(B)$ transitions are not surprising because the sample consists of misoriented anisotropic crystalline grains. Given the predicted resistivity ratio $\Gamma = \rho_c/\rho_{ab} \approx 10$–$15$ for this layered compound[11], the local $B_{c2}(\theta) \approx B_{c2}(0)[\cos^2\theta + \Gamma^{-1}\sin^2\theta]^{-1/2}$ should vary strongly, depending on the angle θ between the *c*-axis in the grain and the applied field. Thus, we can identify two characteristic fields: the high-field onset B_{max} of the superconducting transition, and the zero-resistance, low-field onset B_{min}, as illustrated by Fig. 1. The in-field transitions are shown in Fig. 1a for **B** perpendicular to the broad 1-mm-wide face of the sample, and in Fig. 1b for **B** applied parallel to the same face (in both cases, transport current was always perpendicular to **B**). Differences in the $R(B)$ curves of Fig. 1 do indicate some grain texture, as the zero-resistance field B_{min} in Fig. 1b (36 T at 4.2 K) is higher than in Fig. 1a (25 T at 4.2 K), suggesting that the plate-like grains tend to align their *a*–*b* planes parallel to the broad face of the sample.

Shown in Fig. 2a are the temperature dependences of the fields B_{min}, B_{mid} and B_{max} (all in T) evaluated at 10%, 50% and 90% of the normal state resistance at the transition temperature, $R_n(T_c)$, respectively. The interpretation of B_{min} and B_{max} in polycrystals can be complicated by strong vortex pinning and the particulars of the grain misorientation distribution. However in our case, the analysis is simplified because: (1) the measured magnetization curves are nearly

[1]National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, USA. [2]Materials Science & Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA.

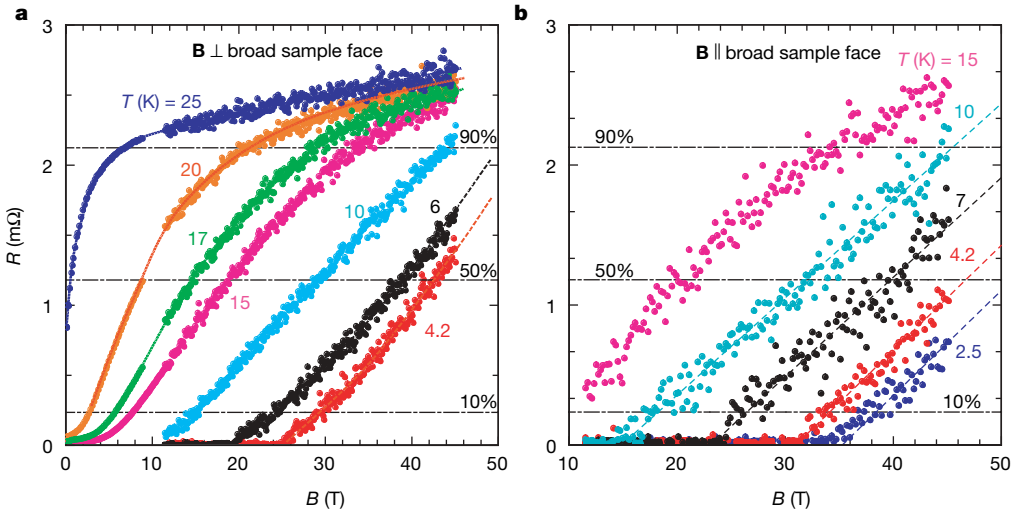


Figure 1 | Variation of $LaFeAsO_{0.89}F_{0.11}$ sample resistance with applied magnetic field at fixed temperatures in the range 4.2 K to 25 K. **a**, The resistance $R(B)$ for different temperatures taken in swept fields in the 45 T hybrid magnet at a measuring current of 1 mA and from 0 to 9 T in a superconducting magnet at a measuring current of 5 mA for **B** applied perpendicular to the broad face of the sample. The current density was $\sim$0.2 A cm^{-2} in the hybrid magnet measurements and 5 times higher in the PPMS superconducting magnet measurements. **b**, $R(B)$ data taken in the hybrid magnet for **B** parallel to the broad sample face, which we believe has a higher fraction of *a*–*b*-oriented grains. The normal state resistivity at 30 K is estimated to be 0.15 mΩ cm with an uncertainty of $\sim$15%. The horizontal dashed lines indicate 10%, 50% and 90% of the resistive transition relative to the normal state resistance at the transition temperature, $R_n(T_c)$, respectively.

reversible, suggesting weak pinning; and (2) the crystalline anisotropy of the compound results in a broad distribution of the local $B_{c2}(\theta)$ values in different grains. Thus, $B_{max}(T)$ is associated with the larger in-plane upper critical field $B_{c2}^{\parallel}(T) \approx B_{c2}^{\perp}(T)\Gamma^{1/2}$ because grains with their *a*–*b* planes oriented along the applied field become superconducting first upon cooling. In turn, the zero-resistance field $B_{min}(T)$ can be interpreted as the field below which those superconducting grains with $B_{c2}(\theta) > B$ form a percolative path at our low measuring current density $\sim$0.2 A cm^{-2}. For $\Gamma \gg 1$ and negligible thermal activation of vortices, both the effective medium and the percolation[23] theories give $B_{min}(T) \approx B_{c2}^{\perp}(T)/p_c$, where $p_c < 1$ is a temperature-independent number determined by the percolation threshold defined by the material anisotropy and the orientational grain distributions of the specific sample. Accordingly, our interpretation of Fig. 2 is that $B_{max}(T)$ and $B_{min}(T)$ reflect the temperature dependences of $B_{c2}^{\parallel}(T)$ and $B_{c2}^{\perp}(T)$, respectively, in the case when thermal activation of vortices is weak, as we now discuss.

The upward curvature of $B_{min}(T)$ in Fig. 2 might also be interpreted as the signature of the irreversibility field $B_m(T)$ resulting from melting of a weakly pinned vortex lattice. Thermal vortex fluctuation effects are quantified by the Ginzburg parameter $Gi = (8\pi^2 k_B T_c \Lambda_a^2/\xi_c \phi_0^2)^2/2$ (ref. 24), where Λ_a is the London penetration depth, ξ_a is the *a*–*b* plane coherence length, ϕ_0 is the magnetic flux quantum, $\xi_c = \xi_a \Gamma^{-1/2}$ is the *c*-axis coherence length, and k_B is the Boltzmann constant. The copper oxide high-temperature superconductors have strong thermal fluctuations resulting in $Gi \approx 1$–10^{-2} and $B_m(T) \ll B_{c2}(T)$. The conventional low-T_c superconductors, in which vortex fluctuations are negligible, have $Gi \approx 10^{-6}$–10^{-10} and $B_{c2} - B_m \ll B_{c2}$. The coherence length $\xi_c = [\phi_0/2\pi B_{c2}(0)\Gamma^{1/2}]^{1/2}$ can be estimated for $\Gamma = 15$ and $B_{c2}^{\parallel}(0) = 60$ T in Fig. 2 as 1.2 nm, which for $T_c = 26$ K and $\Lambda_a = 215$ nm, extracted from recent NMR measurements[25], yields $Gi = 3.4 \times 10^{-4}$, a value close to $Gi = 2.1 \times 10^{-4}$ of clean MgB_2, but a value some 30 times smaller than Gi for the least anisotropic high-temperature superconductor, $YBa_2Cu_3O_{7-x}$, lead-

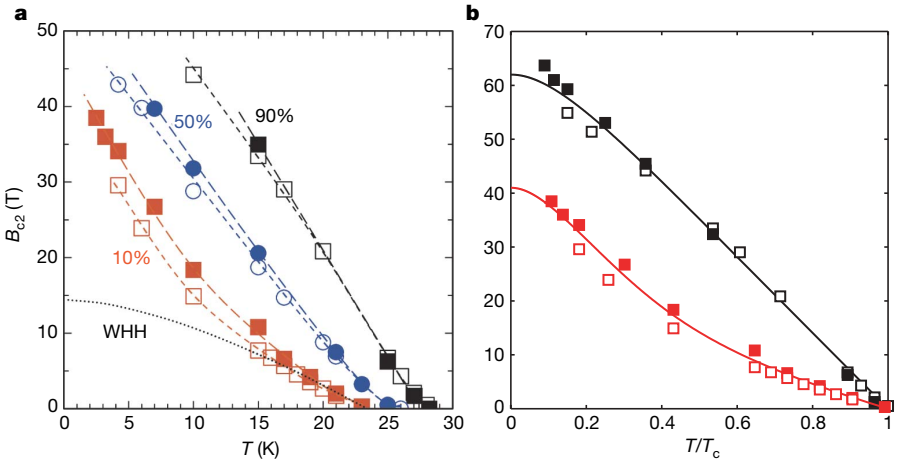


Figure 2 | Upper critical field–temperature phase diagram of $LaFeAsO_{0.89}F_{0.11}$. **a**, The measured fields $B_{max}(T)$ and $B_{min}(T)$ (black and red squares, respectively) along with the midpoint transition fields (blue circles). The filled and open symbols correspond respectively to the parallel and perpendicular field orientations. The dotted line shows the WHH curve defined by the slope of $B_{min}(T)$ at T_c. **b**, $B_{max}(T)$ (black squares) and $B_{min}(T)$ (red squares) plotted as functions of the reduced temperature T/T_c. The data points above 45 T were extracted by linear extrapolation of $R(B)$ at $B < 45$ T to $R(B) = 0.9R_n(T_c)$, as shown by dashed lines in Fig. 1. The lines correspond to $B_{c2}(T)$ calculated from the two-gap theory for the parameters described in the text.

ing us to classify $LaFeAsO_{0.89}F_{0.11}$ as an 'intermediate-T_c' superconductor. We estimate B_m based on a theory of vortex fluctuations in moderately anisotropic superconductors[24] (see Supplementary Information), which shows that $B_{c2}(T)$ and $B_m(T)$ at $T < 0.5T_c$ differ only by ~20%. Therefore, we conclude that the temperature dependence of the resistive transition field $B_{min}(T)$ at $T < 0.5T_c$ reflects the behaviour of $B_{c2}^{\perp}(T)$ rather than that of the melting field $B_m(T)$.

As is evident from Fig. 2, $B_{c2}^{\perp}(T)$ exhibits a significant upward curvature, which is much less pronounced for $B_{c2}^{\parallel}(T)$. Because this behaviour is similar to that observed in dirty MgB_2 films[22], we suggest, in agreement with the *ab initio* calculations[11,12], that superconductivity in $LaFeAsO_{0.89}F_{0.11}$ results from two bands: a nearly two-dimensional electron band with high in-plane diffusivity D_1 and a more isotropic heavy hole band with smaller diffusivity D_2. The upward curvature of $B_{c2}(T)$ is then controlled by the ratio $\eta = D_2/D_1$: for $\eta \ll 1$, the upward curvature is pronounced, while for $\eta \approx 1$, $B_{c2}(T)$ exhibits a more traditional WHH-like behaviour[21]. For fields along the *a*–*b* plane, the parameter η should be replaced with $\eta = D_2/[D_1^{(ab)}D_1^{(c)}]^{1/2}$, allowing strong anisotropy $D_1^{(ab)} \gg D_1^{(c)}$ to significantly increase η for $B||ab$ as compared to $B||c$ (ref. 22). In this case the upward curvature of $B_{c2}^{\parallel}(T)$ does become less pronounced than for $B_{c2}^{\perp}(T)$, in agreement with the data in Fig. 2.

To check the self-consistency of our interpretation, we fit $B_{c2}^{\perp}(T)$ in Fig. 2, using the two-gap theory outlined in Supplementary Information. We took $\eta = D_2/D_1 = 0.08$ and the interband BCS coupling constants $\lambda_{12} = \lambda_{21} = 0.5$, but the results are relatively insensitive to the choice of either λ_{12} and λ_{21} (including their sign) or the intraband coupling constants λ_{11} and λ_{22}. For example, Fig. 2 shows a rather good fit for the case of strong interband repulsion $\lambda_{12}\lambda_{21} \gg \lambda_{11}\lambda_{22}$ suggested in ref. 12. However, this theory also shows that the fit remains nearly as good, even if we assume that intraband pairing is significant, $\lambda_{11}\lambda_{22} > \lambda_{12}\lambda_{21}$, with all coupling constants being of the same order of magnitude. Thus, our experimental data do not yet enable us to unambiguously distinguish between different pairing scenarios suggested in the literature, but they do indicate a significant difference in the effective masses in the electron and hole bands. For $B||ab$, we rescaled the parameter $\eta \rightarrow \eta\Gamma^{1/2} \approx 0.31$, taking the estimate $\Gamma = \rho_c/\rho_{ab} \approx 15$ suggested in ref. 11, which describes $B_{c2}^{\parallel}(T)$ well, as is also evident from Fig. 2. Therefore the two-gap scenario is qualitatively consistent with our experimental data.

The newly discovered $LaFeAsO_{0.89}F_{0.11}$ exhibits exceptionally high B_{c2}, and this obviously non-optimized material has been quickly synthesized in bulk form showing zero resistance above 35 T at 2.5 K. Moreover, given the high dB_{c2}/dT values of 2–3 T K^{-1}, which have also been observed in the Sm-based ($T_c = 43$ K; ref. 6) and Pr- or Nd-based oxypnictides ($T_c = 52$ K), it seems reasonable to expect $B_{c2}^{\parallel}(T)$ and $B_{c2}^{\perp}(T)$ values 1.5–2 times higher than $LaFeAsO_{0.89}F_{0.11}$, which puts $B_{c2}^{\parallel}(0)$ above 100 T. Thus doped oxypnictides appear as a new family of high-field superconductors, for which extensive pulsed-field measurements in the 100 T range will be required to fully reveal the novel physics of competing superconducting and magnetic orders. It is tempting to think that they may also have great importance for high-field applications.

Received 2 April; accepted 5 May 2008.
Published online 28 May 2008.

1. Kamihara, Y., Watanabe, T., Hirano, M. & Hosono, H. Iron-based layered superconductor La[$O_{1-x}F_x$]FeAs (x = 0.05-0.12) with T_c = 26 K. *J. Am. Chem. Soc.* **130,** 3296–3297 (2008).
2. Zhu, X., Yang, H., Fang, L., Mu, G. & Wen, H.-H. Upper critical field, Hall effect and magnetoresistance in the iron-based layered superconductor $LaO_{0.9}F_{0.1-\delta}FeAs$. Preprint at ⟨http://arxiv.org/abs/0803.1288v1⟩ (2008).
3. Chen, G. F. *et al.* Superconducting properties of Fe-based layered superconductor $LaO_{0.9}F_{0.1-\delta}$ FeAs. Preprint at ⟨http://arxiv.org/abs/0803.0128v1⟩ (2008).
4. Sefat, A. S. *et al.* Electron correlations in the superconductor $LaFeAsO_{0.89}F_{0.11}$ with low carrier density. *Phys. Rev. B* **77,** 174503 (2008).
5. Chen, G. F. *et al.* Superconductivity at 41K and its competition with spin-density-wave instability in layered $CeO_{1-x}F_x$ FeAs. Preprint at ⟨http://arxiv.org/abs/0803.3790v2⟩ (2008).
6. Chen, X. H. *et al.* Superconductivity at 43 K in samarium-arsenide oxides $SmFeAsO_{1-x}F_x$. Preprint at ⟨http://arxiv.org/abs/0803.3603v1⟩ (2008).
7. Ren, Z.-A. *et al.* Superconductivity in iron-based F-doped layered quaternary compound Nd[$O_{1-x}F_x$]FeAs. Preprint at ⟨http://arxiv.org/abs/0803.4234⟩ (2008).
8. Ren, Z.-A. *et al.* Superconductivity at 52K in iron-based F-doped layered quaternary compound Pr[$O_{1-x}F_x$]FeAs. Preprint at ⟨http://arxiv.org/abs/0803.4283⟩ (2008).
9. Zimmer, B. I., Jeitschko, W., Albering, J. H., Glaum, R. & Reehuis, M. The rare earth transition metal phosphide oxides LnFePO, LnRuPO and LnCoPO with ZrCuSiAs type structure. *J. Alloy. Comp.* **229,** 238–242 (1995).
10. Quebe, P., Terbuchte, L. J. & Jeitschko, W. Quaternary rare earth transition metal arsenide oxides *RT*AsO (*T*=Fe, Ru, Co) with ZrCuSiAs type structure. *J. Alloy. Comp.* **302,** 70–74 (2000).
11. Singh, D. J. & Du, M. H. $LaFeAsO_{1-x}F_x$: A low carrier density superconductor near itinerant magnetism. Preprint at ⟨http://arxiv.org/abs/0803.0429v1⟩ (2008).
12. Mazin, I. I., Singh, D. H., Johannes, M. D. & Du, M. H. Unconventional sign-reversal superconductivity in $LaFeAsO_{1-x}F_x$. Preprint at ⟨http://arxiv.org/abs/0803.2740v1⟩ (2008).
13. Cao, C., Hirschfeld, P. J. & Cheng, H.-P. Coexistence of antiferromagnetism with superconductivity in $LaO_{1-x}F_xFeAs$: effective Hamiltonian from ab initio studies. Preprint at ⟨http://arxiv.org/abs/0803.3236⟩ (2008).
14. Dai, X., Fang, Z., Zhou, Y. & Zhang, F. Even parity, orbital singlet and spin triplet pairing for superconducting La($O_{1-x}F_x$)FeAs. Preprint at ⟨http://arxiv.org/abs/0803.3982v1⟩ (2008).
15. Dong, J. *et al.* Competing orders and spin-density wave instability in La($O_{1-x}F_x$)FeAs. Preprint at ⟨http://arxiv.org/abs/0803.3426v1⟩ (2008).
16. Ma, F. & Lu, Z.-Y. Iron-based superconductor $LaO_{1-x}F_xFeAs$ superconductor: an antiferromagnetic semimetal. Preprint at ⟨http://arxiv.org/abs/0803.3286v1⟩ (2008).
17. McGuire, M. A. *et al.* Evidence for the spin density wave in LaFeAsO. Preprint at ⟨http://arxiv.org/abs/0804.0796⟩ (2008).
18. Kuroki, K. *et al.* Unconventional superconductivity originating from disconnected Fermi surfaces in $LaO_{1-x}F_xFeAs$. Preprint at ⟨http://arxiv.org/abs/0803.3325v1⟩ (2008).
19. Lebegue, S. Electronic structure and properties of the Fermi surface of the superconductor LaOFeP. *Phys. Rev. B* **75,** 035110 (2007).
20. Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitani, Y. & Akimitsu, J. Superconductivity at 39 K in magnesium diboride. *Nature* **410,** 63–64 (2001).
21. Werthamer, N. R., Helfand, E. & Hohenberg, P. C. Temperature and purity dependence of the superconducting critical field, H_{c2}. III. Electron spin and spin-orbit effects. *Phys. Rev.* **147,** 295–302 (1966).
22. Gurevich, A. Limits of the upper critical field in dirty two-gap superconductors. *Physica C* **456,** 160–169 (2007); Enhancement of the upper critical field by nonmagnetic impurities in dirty two-gap superconductors. *Phys. Rev. B* **67,** 184515 (2003).
23. Glazman, L. I., Koshelev, A. I. & Lebed', A. G. Resistive transition and critical fields of superconducting ceramics. *Sov. Phys. JETP* **67,** 1235–1241 (1988).
24. Blatter, G. & Geshkenbein, V. in *The Physics of Superconductors* Vol. 1 (eds Bennerman, K. H. & Ketterson, J. B.) 800 (Springer, Berlin, 2003).
25. Ahilan, K. *et al.* ^{19}F NMR investigation of $LaAsFeO_{0.89}F_{0.11}$ superconductor. Preprint at ⟨http://arxiv.org/abs/0804.4026⟩ (2008).

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

Acknowledgements Work at the NHMFL was supported by IHRP under NSF Cooperative Agreement, by the State of Florida, by the DOE, by the NSF Focused Research Group on Magnesium Diboride (FRG), and by AFOSR. Work at ORNL was supported by the Division of Materials Sciences and Engineering, Office of Basic Energy Sciences. We are grateful for discussions with G. Boebinger, E. Hellstrom, P. Lee, J. Jiang, and C. Tarantini at the NHMFL.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.H. (hunte@asc.magnet.fsu.edu).

Vol 453|12 June 2008|doi:10.1038/nature07010

LETTERS

Capture of hydroxymethylene and its fast disappearance through tunnelling

Peter R. Schreiner[1], Hans Peter Reisenauer[1], Frank C. Pickard IV[2], Andrew C. Simmonett[2], Wesley D. Allen[2], Edit Mátyus[3] & Attila G. Császár[3]

Singlet carbenes exhibit a divalent carbon atom whose valence shell contains only six electrons, four involved in bonding to two other atoms and the remaining two forming a non-bonding electron pair. These features render singlet carbenes so reactive that they were long considered too short-lived for isolation and direct characterization. This view changed when it was found that attaching the divalent carbon atom to substituents that are bulky and/or able to donate electrons produces carbenes that can be isolated and stored[1]. *N*-heterocyclic carbenes are such compounds now in wide use, for example as ligands in metathesis catalysis[2]. In contrast, oxygen-donor-substituted carbenes are inherently less stable and have been less studied. The pre-eminent case is hydroxymethylene, H–C–OH; although it is the key intermediate in the high-energy chemistry of its tautomer formaldehyde[3–7], has been implicated since 1921 in the photocatalytic formation of carbohydrates[8], and is the parent of alkoxycarbenes that lie at the heart of transition-metal carbene chemistry, all attempts to observe this species or other alkoxycarbenes have failed[9]. However, theoretical considerations indicate that hydroxymethylene should be isolatable[10]. Here we report the synthesis of hydroxymethylene and its capture by matrix isolation. We unexpectedly find that H–C–OH rearranges to formaldehyde with a half-life of only 2 h at 11 K by pure hydrogen tunnelling through a large energy barrier in excess of 30 kcal mol^{-1}.

Apart from Fischer carbenes—named after the chemist who prepared the first metal-carbene compound, $W(CO)_5(C(CH_3)OCH_3)$ (ref. 11)—oxygen-donor-substituted carbenes have received little attention, mostly because of the lack of suitable synthetic precursors. Even the tetra-atomic hydroxymethylene archetype (**1**; Fig. 1) has not been isolated until now because the starting materials necessary for typical carbene generation strategies are too unstable or too difficult to prepare.

We devised a route to **1** using thermal extrusion of CO_2 (refs 12, 13) from glyoxylic acid (**3**; Fig. 1) via high-vacuum flash pyrolysis (HVFP), followed by immediate matrix isolation. This approach is probably of general utility for generating otherwise inaccessible carbenes. Starting material **3** is a commodity that can be readily purified for matrix isolation experiments. It is also volatile enough to enable vaporization and transport through a heated quartz tube. After several attempts and optimization of the reaction parameters (temperature, gas flow, length and diameter of quartz tube), we were successful in the generation and trapping of **1** in an argon matrix at 11 K (see Methods for details).

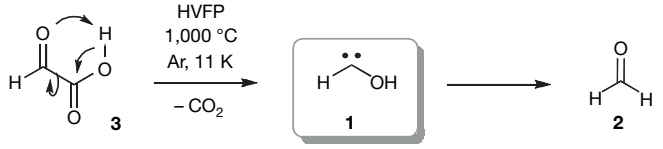


Figure 1 | Generation of hydroxymethylene (1). Thermal extrusion of CO_2 through high-vacuum flash pyrolysis (HVFP) from glyoxylic acid (**3**) and rearrangement of **1** to formaldehyde (**2**).

We characterized H–C–OH and H–C–OD by measuring their infrared spectra (Fig. 2a, b) and comparing them with decisive electronic structure and variational nuclear motion computations, the latter using a high-quality quartic vibrational force field. Inspection of Table 1 shows remarkable agreement between measured and predicted vibrational band origins (VBOs), which provides convincing evidence for the successful preparation of **1**. For 13 of the 15 assigned bands of *trans*-H–C–OH and *trans*-H–C–OD, the mean and standard deviation of the residual between theory and experiment are only 6.0 and 3.3 cm^{-1}, respectively, which is well within the expected range of matrix shifts. We stress that no empirical adjustments were applied to the variationally computed VBOs. The related harmonic frequencies also displayed in Table 1 demonstrate that the full inclusion of vibrational anharmonicity is essential for achieving such a high level of agreement. Particularly impressive is the match for four combination and overtone levels: $\nu_3 + \nu_4$ of H–C–OH, $2\nu_3$ of H–C–OD, and the Fermi resonance pair ($\nu_3 + \nu_4$, ν_1) of H–C–OD. The only notable disparities in Table 1 occur for ν_1(OH stretch) of **1** and ν_2(OD stretch) of [D_1]-**1**, for which the experimental assignments lie 61 and 39 cm^{-1}, respectively, below the theoretical values. Matrix shifts of this magnitude for vibrational fundamentals are not unprecedented[14].

The recorded ultraviolet/visible spectrum of **1** displays one very weak absorption band between 500 and 380 nm (wavelength of maximum absorption, 427 nm) with a distinctive vibrational fine structure (Fig. 3). Irradiation of the matrix with monochromatic light within this spectral range (435, 470 and 500 nm) causes rapid rearrangement of **1** to formaldehyde (**2**; Fig. 1) and partial fragmentation into CO and H_2. On the basis of rigorous multireference coupled cluster computations within the Mk-MRCCSD formalism[15], we attribute the signals in Fig. 3 to the lowest-lying open-shell singlet excited electronic state (S_1) of carbene **1**. Upon geometry optimization, the S_1 state of **1** relaxes to a non-planar structure with a widened H–C–O angle of 127.4° and a dihedral angle of 108.4°, consistent with the extensive vibrational progression observed in the electronic absorption spectrum. Our best theoretical vertical and adiabatic excitation energies are 2.99 eV (415 nm) and 2.40 eV (516 nm), respectively, in complete accord with experiment.

For further theoretical characterization of **1,** we optimized all singlet structures related to **1** at the rigorous AE-CCSD(T)/cc-pCVQZ

[1]Institut für Organische Chemie der Justus-Liebig-Universität, Heinrich-Buff-Ring 58, D-35392 Giessen, Germany. [2]Center for Computational Chemistry and Department of Chemistry, University of Georgia, Athens, Georgia 30602, USA. [3]Laboratory of Molecular Spectroscopy, Institute of Chemistry, Eötvös University, H-1518 Budapest 112, PO Box 32, Hungary.

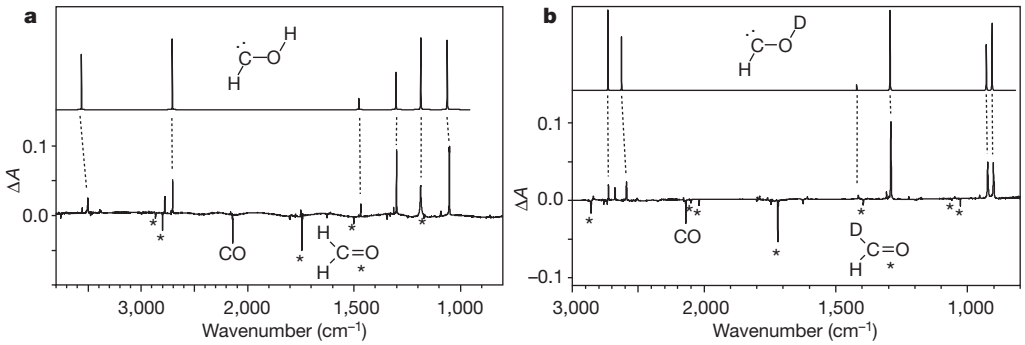


Figure 2 | Infrared spectra of 1 and [D_1]-1. a, Lower trace, hydroxymethylene (Ar matrix, 11 K) difference infrared spectrum (absorbance change ΔA) obtained by subtracting the spectra of unirradiated and irradiated (2 min at 470 nm) matrix-isolated pyrolysis products of glyoxylic acid. Upper trace, computed AE-CCSD(T)/cc-pCVQZ variational anharmonic VBOs of *trans*-hydroxymethylene (**1t**) and double-harmonic intensities from Table 1. Asterisks indicate peaks assigned to **2**. **b**, Lower trace, [D_1]-hydroxymethylene (Ar matrix, 11 K) difference infrared spectrum obtained by subtracting the spectra of unirradiated and irradiated (2 min at 470 nm) matrix-isolated pyrolysis products of [D_1]-glyoxylic acid. Upper trace, computed AE-CCSD(T)/cc-pCVQZ variational anharmonic VBOs of [D_1]-*trans*-hydroxymethylene and double-harmonic intensities from Table 1.

level and pinpointed their relative energies to around 0.1 kcal mol^{-1} by means of exhaustive focal-point analyses (FPA)[16]. (For structural depictions, optimized *xyz* coordinates, and FPA energetic tables, see Supplementary Figs 1 and 2, Supplementary Tables 2 and 3, and Supplementary Tables 4–10, respectively; for full details of the theoretical approach, see Methods.) The potential energy surface[17] surrounding **1** is shown schematically in Fig. 4. Hydroxymethylene exhibits a *trans*-planar equilibrium structure (**1t**) with an H–C–O angle of 102.3°. The C–O bond length of 1.311 Å is much shorter than the corresponding 1.427 Å distance in methanol[18], indicative of the expected π-type stabilization of the electron-deficient carbene centre through the adjacent oxygen *p* lone pair. Our FPA computations place the lowest triplet state of H–C–OH much higher in energy (T_0 = 28.0 kcal mol^{-1}) than the closed-shell singlet **1t**. This large separation increases at the transition state (**TS2**; Fig. 4) for isomerization to formaldehyde, where AE-CCSD(T)/cc-pCVQZ theory gives a singlet–triplet energy difference of 37.6 kcal mol^{-1}. These results indicate that intersystem crossing to triplet surfaces should not occur in our experiments. In addition, we find no evidence for the *cis*-**1** isomer (**1c**; Fig. 4) that lies 4.4 kcal mol^{-1} above **1t**, presumably because of the high barrier (**TS1**, 26.8 kcal mol^{-1}; Fig. 4) for internal rotation and because the CO_2 extrusion process of Fig. 1 is likely to yield exclusively **1t** (similar processes yielding only specific conformers have been reported before[19]). The barrier computed for the rearrangement of **1t** to **2** is even higher (**TS2**, 29.7 kcal mol^{-1}), which reinforces the earlier theoretical prediction[10] that hydroxymethylene should be observable under matrix isolation conditions.

Despite being located at an energy minimum on the potential energy surface that is surrounded by high enthalpic barriers, matrix-isolated **1** disappears quickly with a half-life ($t_{1/2}$) of ~2 h in Ar, Kr and Xe matrices. The magnitude of $t_{1/2}$ is virtually independent of temperature in the 11–20 K range (see Supplementary Table 1). In contrast, the monodeuterated species ([D_1]-**1**) is completely stable under the same conditions. Thermal rearrangement through either **TS1** or **TS2** is not conceivable at such low temperatures. Because of the low concentration of **1**, the very limited mobility of molecules in noble-gas matrices, and the observed lack of temperature dependence of $t_{1/2}$, bimolecular reactions are also an unlikely cause. Therefore, quantum mechanical tunnelling appears to be the most viable mechanism behind the rapid disappearance of *trans*-H–C–OH as well as the persistence of H–C–OD. It is relevant that strong tunnelling effects in unimolecular reactions on the potential energy

Table 1 | Comparison of computed and experimental vibrational frequencies

Vibrational description*	ω (I_{rel})	ν	Expt. (I_{rel})
trans-H–C–OH (**1t**)			
78% ν_1(OH str.) + 8% [$\nu_3 + 2\nu_6$]	3,765.4 (57)	3,561.6	3,500.6 (43)
84% [$\nu_4 + \nu_5 + \nu_6$]		3,520.8	
83% [$3\nu_5$]		3,516.0	
72% [$\nu_3 + \nu_4$] + 15% ν_2(CH str.) + 14% [$\nu_3 + \nu_5$]		2,785.5	2,776.2 (16)
55% ν_2(CH str.) + 18% [$\nu_3 + \nu_4$]	2,876.6 (99)	2,706.5	2,703.3 (37)
96% ν_3(HOC def. + HCO def.)	1,513.7 (13)	1,475.1	1,465.5 (10)
92% ν_4(CO str.)	1,334.1 (39)	1,300.5	1,297.1 (53)
97% ν_5(HOC def. − HCO def.)	1,220.0 (100)†	1,183.5	1,183.2 (100)
97% ν_6(twist)	1,094.3 (82)	1,058.9	1,048.5 (88)
trans-H–C–OD ([D_1]-**1t**)			
87% [$2\nu_3$] + 7% ν_1		2,852.6	2,841.3 (8)
51% [$\nu_3 + \nu_4$] + 33% ν_1(CH str.) + 5% [$2\nu_3$]	2,878.9 (100)‡	2,729.5	2,726.1 (30)
45% ν_1(CH str.) + 37% [$\nu_3 + \nu_4$] + 5% ν_2		2,682.8	2,675.9 (20)
83% ν_2(OD str.)	2,739.5 (57)	2,626.8	2,588.1 (58)
78% [$2\nu_4$] + 5% ν_4		2,566.4	
97% ν_3(HCO def.)	1,451.0 (8)	1,420.8	1,414.7 (4)
92% ν_4(CO str.)	1,326.4 (84)	1,294.1	1,290.8 (100)
98% ν_5(DOC def.)	953.6 (58)	928.7	923.1 (66)
98% ν_6(twist)	933.8 (70)	907.1	901.6 (59)

Theoretical (AE-CCSD(T)/cc-pCVQZ) harmonic (ω) and variational anharmonic (ν) vibrational band origins (in cm^{-1}) relative to experimental bands (Ar matrix, 11 K) for *trans*-hydroxymethylene and its monodeuterated isotopologue. Relative infrared absorption intensities (I_{rel}, in %) from the double-harmonic approximation are included from AE-CCSD(T)/cc-pVTZ computations.
* Distribution of the converged vibrational wavefunctions over the normal mode basis states; all contributions larger than 5% are listed.
† Absolute intensity, 142 km mol^{-1}.
‡ Absolute intensity, 116 km mol^{-1}.

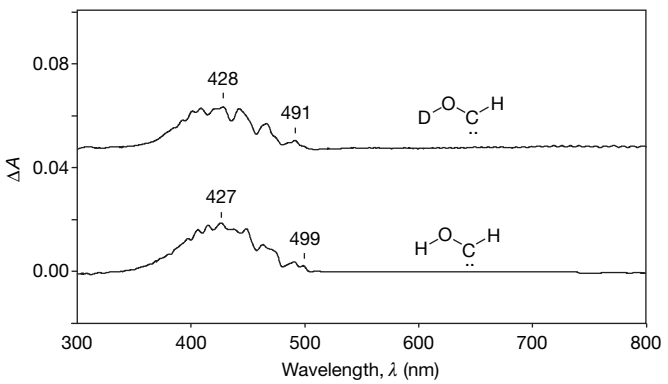


Figure 3 | Ultraviolet/visible spectrum of 1 and [D_1]-1. Shown are difference ultraviolet/visible spectra (unirradiated − irradiated) of **1** and [D_1]-**1** in an argon matrix at 11 K.

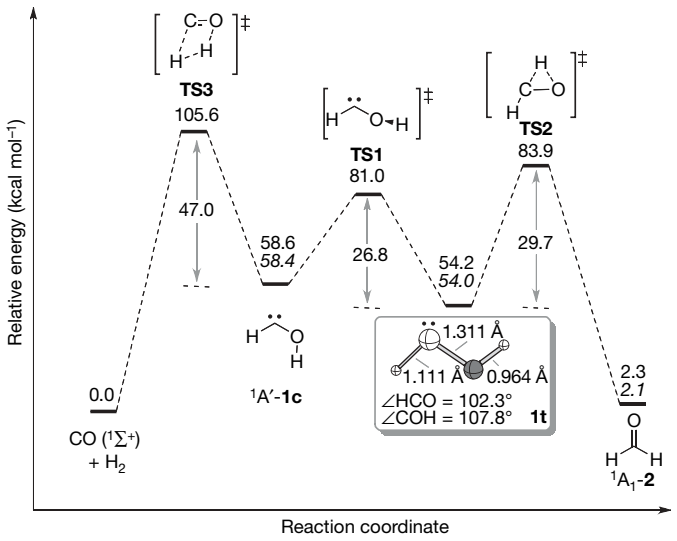


Figure 4 | Schematic H–C–O–H potential energy hypersurface. Key features of the computed energy profile (high-level coupled cluster focal point analyses + harmonic zero-point vibrational energies (ZVPE); values in italics use anharmonic ZPVEs) for the rearrangement of singlet **1** (**1c**, *cis* form; **1t**, *trans* form, inset) to formaldehyde and CO + H_2.

surface shown in Fig. 4 have been noted in a number of previous experimental and theoretical studies[20,21].

To obtain a theoretical tunnelling rate for **1t** isomerizing to **2**, we used the AE-CCSD/cc-pCVTZ method to precisely map out the intrinsic reaction path (IRP) connecting **TS2** to the reactant and product and to determine harmonic vibrational frequencies along this steepest-descent route. A final potential energy curve for the isomerization IRP was then constructed from high-quality AE-CCSD(T)/cc-pCVQZ energy points. The quantum dynamics was treated within a simple reaction-path hamiltonian model[22] with tunnelling probabilities given by the standard WKB (Wentzel-Kramers-Brillouin) formula in terms of barrier penetration integrals, which we computed numerically from our electronic structure data.

The vibrational 'reaction' mode of **1t** that leads towards **TS2** has a harmonic frequency of $\omega_0 = 1{,}220\ \mathrm{cm}^{-1}$. The tunnelling lifetime of *trans*-H–C–OH near 0 K is obtained by ascribing a 'collision' energy (ε) equivalent to the zero-point vibrational energy (ZPVE) of the reaction mode ($\omega_0/2$), evaluating the WKB transmission coefficient at this energy, $\kappa(\varepsilon)$, and multiplying by the classical rate (ω_0) at which the reactant hits the barrier[23]. This theoretical analysis yields a half-life of 2.0 h (122 min) for **1t** in its ground vibrational level, implying that the observed $t_{1/2}$(H–C–OH) ≈ 2 h can be fully accounted for by a tunnelling mechanism. The same computational procedure yields a tunnelling half-life for *trans*-H–C–OD of over 1,200 yr, which is also consistent with experiment. Evaluating $\kappa(\varepsilon)$ by fitting the asymmetric Eckart potential[21,24] to the imaginary **TS2** barrier frequency ($\omega^* = 2{,}174i\ \mathrm{cm}^{-1}$) as well as the ZPVE-corrected reaction energy ($-52.05\ \mathrm{kcal\ mol}^{-1}$) and barrier height ($31.60\ \mathrm{kcal\ mol}^{-1}$) yields qualitatively similar tunnelling lifetimes for both isotopologues of hydroxymethylene.

Pending full-dimensional quantum dynamics computations on a semi-global potential energy surface, we conclude that the disappearance of **1t** is attributable to pure quantum mechanical tunnelling under a barrier in excess of $30\ \mathrm{kcal\ mol}^{-1}$. The occurrence of such an event near 0 K on a timescale of a few hours is a striking chemical phenomenon with little precedent[25–28].

It has been proposed that **1** may exist in interstellar space and that its reaction with **2** would be a possible source of simple sugars[8]. The properties of hydroxymethylene revealed in our experiments raise considerable problems for the astronomical detection of **1t**, as well as its participation in interstellar chemistry at very low pressures. That is, molecule **1t** will be inherently unstable at all energies unless it is formed in an almost perfect quantum mechanical stationary state whose vibrational wavefunction is highly localized, despite the propensity of this species, at least in dissipative surroundings, to tunnel to **2** and become trapped there. Extensive theoretical work[29] on the acetylene/vinylidene isomerization suggests that such long-lived hydroxymethylene states may in principle exist *in vacuo*, but populating and characterizing them would certainly be a formidable experimental challenge.

METHODS SUMMARY

Several HVFP matrix isolation experiments were performed to determine the optimal pyrolysis temperature (1,000 °C) for producing **1**. Condensing the decarboxylation products of **3** on a cold window at 11 K in an Ar matrix ultimately achieved a **1**:**2** ratio of 1:5.5. Comparable results were obtained from experiments in Kr, Xe and N_2 matrices and at temperatures above 11 K (compare Supplementary Table 1). In addition to the parent isotopologue **1**, monodeuterated H–C–OD ([D_1]-**1**) was prepared from HC(O)-C(O)OD and was subjected to the same detailed experimental and computational scrutiny.

The electronic structure computations used correlation-consistent families [(aug)-cc-p(C)V*X*Z] of atomic-orbital basis sets. Most electronic wavefunctions were determined with single-reference coupled-cluster theory, incorporating all single and double excitations (CCSD), and with perturbative inclusion of connected triple excitations [CCSD(T)]. Higher-order correlation effects were evaluated similarly with the CCSDT(Q) method. Optimum geometric structures for all species on the ground-state singlet surface were obtained with all-electron (AE) CCSD(T) theory using the cc-pCVQZ basis set. Final energetics were determined by means of valence focal-point extrapolations[16] with the cc-pV*X*Z series of basis sets from $X = 2$ through to 6. The valence FPA results were appended with core correlation shifts, as well as one-electron mass-velocity and Darwin relativistic corrections, both computed at the CCSD(T)/cc-pCVQZ level.

For anharmonic vibrational computations, AE-CCSD(T)/cc-pCVQZ complete quartic force fields were evaluated for all minima on the ground-state singlet surface. VBOs for hydroxymethylene were computed with a variational approach called DEWE that constructs a discrete variable representation of the full normal-coordinate Eckart-Watson hamiltonian but handles any internal-coordinate potential function via an analytical transformation[30]. The VBOs in Table 1 were computed from an optimal Simons-Parr-Finlan representation of the AE-CCSD(T)/cc-pCVQZ quartic force field, using sufficient grid points to converge the eigenvalues to better than $0.1\ \mathrm{cm}^{-1}$.

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

Received 25 November 2007; accepted 16 April 2008.

1. Bourissou, D., Guerret, O., Gabbaï, F. P. & Bertrand, G. Stable carbenes. *Chem. Rev.* **100,** 39–91 (2000).
2. Nolan, S. P. *N-Heterocyclic Carbenes in Synthesis* (Wiley-VCH, Weinheim, 2006).
3. Kemper, M. J. H., Vandijk, J. M. F. & Buck, H. M. Ab initio calculation on the photochemistry of formaldehyde. The search for a hydroxycarbene intermediate. *J. Am. Chem. Soc.* **100,** 7841–7846 (1978).
4. Lucchese, R. R. & Schaefer, H. F. Metal-carbene complexes and the possible role of hydroxycarbene in formaldehyde laser photochemistry. *J. Am. Chem. Soc.* **100,** 298–299 (1978).
5. Hoffmann, M. R. & Schaefer, H. F. Hydroxycarbene (HCOH) and protonated formaldehyde: Two potentially observable interstellar molecules. *Astrophys. J.* **249,** 563–565 (1981).
6. Reid, D. L., Hernández-Trujillo, J. & Warkentin, J. A theoretical study of hydroxycarbene as a model for the homolysis of oxy- and dioxycarbenes. *J. Phys. Chem. A* **104,** 3398–3405 (2000).
7. Goddard, J. D. & Schaefer, H. F. The photodissociation of formaldehyde: Potential energy surface features. *J. Chem. Phys.* **70,** 5117–5134 (1979).
8. Baly, E. C. C., Heilbron, I. M. & Barker, W. F. CX.—Photocatalysis. Part I. The synthesis of formaldehyde and carbohydrates from carbon dioxide and water. *J. Chem. Soc. Trans.* **119,** 1025–1035 (1921).
9. Sierra, M. A. Di- and polymetallic heteroatom stabilized (Fischer) metal carbene complexes. *Chem. Rev.* **100,** 3591–3637 (2000).
10. Pau, C.-F. & Hehre, W. J. Relative thermochemical stabilities of hydroxymethylene and formaldehyde by ion cyclotron double resonance spectroscopy. *J. Phys. Chem.* **86,** 1252–1253 (1982).
11. Fischer, E. O. & Maasböl, A. On existence of tungsten carbonyl carbene complex. *Angew. Chem. Int. Edn Engl.* **3,** 580–581 (1964).
12. Weiner, B. R. & Rosenfeld, R. N. Pyrolysis of pyruvic acid in the gas phase. A study of the isomerization mechanism of a hydroxycarbene intermediate. *J. Org. Chem.* **48,** 5362–5364 (1983).

13. Rosenfeld, R. N. & Weiner, B. Energy disposal in the photofragmentation of pyruvic acid in the gas phase. *J. Am. Chem. Soc.* **105,** 3485–3488 (1983).
14. Jacox, M. E. The spectroscopy of molecular reaction intermediates trapped in the solid rare gases. *Chem. Soc. Rev.* **31,** 108–115 (2002).
15. Evangelista, F. A., Allen, W. D. & Schaefer, H. F. Coupling term derivation and general implementation of state-specific multireference coupled cluster theories. *J. Chem. Phys.* **127,** 024102 (2007).
16. Császár, A. G., Allen, W. D. & Schaefer, H. F. In pursuit of the *ab initio* limit for conformational energy prototypes. *J. Chem. Phys.* **108,** 9751–9764 (1998).
17. Schreiner, P. R. & Reisenauer, H. P. The ''non-reaction'' of ground-state triplet carbon atoms with water revisited. *ChemPhysChem* **7,** 880–885 (2006).
18. Venkateswarlu, P. & Gordy, W. Methyl alcohol II. Molecular structure. *J. Chem. Phys.* **23,** 1200–1202 (1955).
19. Reisenauer, H. P., Romanski, J., Mloston, G. & Schreiner, P. R. Dimethoxycarbene: Conformational analysis of a reactive intermediate. *Eur. J. Org. Chem.* 4813–4818 (2006).
20. Sodeau, J. R. & Lee, E. K. C. Intermediacy of hydroxymethylene (HCOH) in the low temperature matrix photochemistry of formaldehyde. *Chem. Phys. Lett.* **57,** 71–74 (1978).
21. Miller, W. H. Tunneling corrections to unimolecular rate constants, with application to formaldehyde. *J. Am. Chem. Soc.* **101,** 6810–6814 (1979).
22. Miller, W. H., Handy, N. C. & Adams, J. E. Reaction path Hamiltonian for polyatomic molecules. *J. Chem. Phys.* **72,** 99–112 (1980).
23. Carrington, T. Jr, Hubbard, L. M., Schaefer, H. F. & Miller, W. H. Vinylidene: Potential energy surface and unimolecular reaction dynamics. *J. Chem. Phys.* **80,** 4347–4354 (1984).
24. Johnston, H. S. *Gas Phase Reaction Rate Theory* (Ronald Press, New York, 1966).
25. McMahon, R. J. & Chapman, O. L. Direct spectroscopic observation of intramolecular hydrogen shifts in carbenes. *J. Am. Chem. Soc.* **109,** 683–692 (1987).
26. Zuev, P. S. & Sheridan, R. S. Tunneling in the C–H insertion of a singlet carbene: *tert*-Butylchlorocarbene. *J. Am. Chem. Soc.* **116,** 4123–4124 (1994).
27. Pettersson, M. *et al.* *Cis* → *trans* conversion of formic acid by dissipative tunneling in solid rare gases: Influence of environment on the tunneling rate. *J. Chem. Phys.* **117,** 9095–9098 (2002).
28. Maçôas, E. M. S., Khriachtchev, L., Pettersson, M., Fausto, R. & Räsänen, M. Rotational isomerism of acetic acid isolated in rare-gas matrices: Effect of medium and isotopic substitution on IR-induced isomerization quantum yield and *cis* → *trans* tunneling rate. *J. Chem. Phys.* **121,** 1331–1338 (2004).
29. Zou, S., Bowman, J. M. & Brown, A. Full-dimensionality quantum calculations of acetylene–vinylidene isomerization. *J. Chem. Phys.* **118,** 10012–10023 (2003).
30. Mátyus, E., Czakó, G., Sutcliffe, B. T. & Császár, A. G. Vibrational energy levels with arbitrary potentials using the Eckart-Watson Hamiltonians and the discrete variable representation. *J. Chem. Phys.* **127,** 084102 (2007).

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

Acknowledgements We are grateful for support from the Fonds der Chemischen Industrie, the US Department of Energy, and the Hungarian Scientific Research Fund. We thank J. Bowman and W. Miller for comments on the tunnelling analysis.

Author Contributions P.R.S. and H.P.R. formulated the initial working hypothesis and provided, analysed and interpreted all experimental data. F.C.P. and A.C.S. performed all the electronic structure computations under the direction of W.D.A. The variational vibrational computations were executed by E.M. under the guidance of A.G.C. and W.D.A. The tunnelling analysis was performed by W.D.A., with input from F.C.P. and A.C.S. The manuscript was primarily written by P.R.S. and W.D.A.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.R.S. (prs@org.chemie.uni-giessen.de) or W.D.A. (wdallen@uga.edu).

METHODS

Matrix isolation studies. The cryostat used for the matrix isolation studies was an APD Cryogenics HC-2 closed-cycle refrigerator system fitted with CsI windows for infrared (IR) and BaF_2 windows for ultraviolet/visible (UV/Vis) measurements. IR spectra were recorded with a Bruker IFS 55 FTIR spectrometer (4,500–300 cm^{-1}, resolution 0.7 cm^{-1}), and UV/Vis spectra were recorded with an Agilent HP 8453 diode-array spectrometer and a JASCO V-670 spectrophotometer. For the combination of HVFP with matrix isolation we used a home-built, water-cooled oven directly connected to the vacuum shroud of the cryostat. The pyrolysis zone consisted of an empty quartz tube (inner diameter 8 mm, length of heating zone 50 mm) resistively heated by a wire. The temperature was controlled by a Ni/CrNi thermocouple. Water-free glyoxylic acid was prepared by heating the commercially available monohydrate (Sigma-Aldrich) under vacuum for several days; monodeuterated glyoxylic acid was obtained by dissolving the monohydrate in D_2O and evaporating to dryness three times. The precursors were evaporated from a heated storage bulb (70–80 °C) into the quartz pyrolysis tube. Immediately after leaving the tube, at a distance of ~50 mm, the pyrolysis products were co-condensed with a large excess of either argon, krypton, xenon, or nitrogen on the surface of the cold matrix window. For irradiations a mercury high-pressure lamp (HBO 200, Osram) with a monochromator (Bausch & Lomb) was used (band width ~10 nm).

Several experiments were performed in order to determine the optimal reaction temperature (1,000 °C). Under these conditions the decarboxylation reactions were not complete, and unreacted precursors were always present in the matrices. Moreover, the main reaction products in the matrix were CO_2, formaldehyde (**2**), and small amounts of CO. The yield of matrix-isolated hydroxymethylene (**1**) compared to formaldehyde was about 20%. By measuring the spectral differences of irradiated and unirradiated matrices, it was possible to elaborate the IR and UV/Vis spectral properties of **1**, despite its low concentration.

The kinetics of the thermal rearrangement of **1** to **2** at 11 K (lowest achievable temperature), 15 K and 20 K were investigated by taking IR spectra every 30 min while keeping the cryostat at the respective temperature and carefully shielding it from all external light sources (compare Supplementary Table 1). Decrease of the two most intense peaks of **1** at ~1,050 and 1,300 cm^{-1} clearly followed first-order kinetics with a half-life of around 2 h in Ar, Kr and Xe, whereas the lifetime of **1** in an N_2 matrix was prolonged by a factor larger than three. The bands of $[D_1]$-**1t** did not change under identical conditions for extended periods of time.

Electronic structure computations. The Supplementary Information contains extensive details of the electronic structure computations. The AE-CCSD(T)/cc-pCVQZ optimum geometric structures for all species on the ground-state singlet surface are depicted in Supplementary Fig. 1, and corresponding atomic cartesian coordinates are given in Supplementary Tables 2 and 3. All of the explicitly computed electronic energies entering into the subsequent focal-point analyses are listed in Supplementary Table 4. The focal-point extrapolations, with auxiliary core correlation, relativistic, and ZPVE terms, are laid out in Supplementary Table 5 for the seven relative energies pinpointed in this study. Supplementary Table 6 provides a succinct collection of the final FPA results. Our FPA procedures have been described in earlier publications[16,31–34] and validated on a large number of chemical systems. For completeness, we provide references detailing the cc-pV*X*Z ($X = 2$–6)[35] and cc-pCVQZ[36] basis sets, our relativistic methods[37–39], and the recently formulated CCSDT(Q) scheme for incorporating connected quadruple excitations into coupled cluster computations[40,41].

The first open-shell singlet excited electronic state (S_1) of H–C–OH was investigated by our recently developed multireference coupled cluster (Mk-MRCCSD) methods[15,42]. For planar geometries, the S_1 state is of $^1A''$ symmetry, arising from the $7a' \rightarrow 2a''$ molecular orbital excitation. The reference determinants for the Mk-MRCCSD computations included all possible distributions of two electrons in the ($7a'$, $2a''$) HOMO/LUMO active space. For planar geometries only two references are of the correct symmetry, but four references must be incorporated when the S_1 state twists out of plane. The optimum (twisted, C_1) structure of the S_1 state, as well as the planar transition state for internal rotation, is shown in Supplementary Fig. 2, as given by our four-reference, all-electron Mk-MRCCSD computations with the aug-cc-pCVTZ basis set. Comparative structures for the ground-state singlet (S_0) and analogous triplet (T_1) states are also given in Supplementary Fig. 2. Cartesian coordinates of the Mk-MRCCSD optimized structures appear in Supplementary Table 7. Supplementary Table 5 gives the total energies involved in our Mk-MRCCSD determination of the vertical and adiabatic excitation energies for the S_1 state. All multireference coupled cluster computations for the S_1 state were carried out using ROHF orbitals optimized on the corresponding triplet state (T_1).

Vibrational computations and tunnelling analysis. Anharmonic ZPVEs of all minima on the ground-state singlet surface were computed by applying vibrational perturbation theory (VPT2)[43–48] to our complete AE-CCSD(T)/cc-pCVQZ quartic force fields, including the G_0 terms[49]. ZPVEs of all transition states were evaluated from harmonic frequencies at the AE-CCSD(T)/cc-pCVQZ level. The quartic force fields were determined from a grid of tightly-converged energy points (263 in the case of **1t**) using established finite-difference techniques[50]. The methodology is implemented in INTDIF2005, which is an abstract program written by W.D.A. for Mathematica to perform general numerical differentiations to high orders of electronic structure data.

For the variational vibrational computations, the potential energy surfaces of the *trans* conformations of H–C–OH and H–C–OD were constructed by transforming the AE-CCSD(T)/cc-pCVQZ quartic force fields to an optimal SPF coordinate representation[51]. The final variational procedure used 531,441 grid points to build the vibrational hamiltonian matrix, which allowed convergence of the eigenvalues to better than 0.1 cm^{-1}. To give a quantitative interpretation of the computed vibrational levels, overlaps of the numerically exact wavefunctions were evaluated over multidimensional normal-mode harmonic oscillator basis functions. The squares of these coefficients comprise the normal mode distributions (NMDs) that describe the vibrational states in Table 1 of the text. Convergence of the NMDs presented in Table 1 is better than 1%.

Our tunnelling analysis employed IRPs[52] and mathematical algorithms[53] for generating them that are widely employed in theoretical chemistry. The application of a reaction path hamiltonian along an IRP requires the computation of vibrational frequencies for modes orthogonal to the path, a task which demands attention to numerous subtle issues[54]. The WKB method[55,56] for evaluating transmission probabilities through a potential energy barrier in terms of barrier penetration integrals is a classic theory that has long been applied to isomerization problems[57].

31. Allen, W. D., East, A. L. L. & Császár, A. G. in *Structures and Conformations of Non-Rigid Molecules* (eds Laane, J., Dakkouri, M., van der Veken, B. & Oberhammer, H.) 343–373 (NATO ASI Series C, Kluwer, Dordrecht, 1993).
32. Császár, A. G. *et al.* in *Spectroscopy from Space* (eds Demaison, J., Sarka, K. & Cohen, E. A.) 317–339 (NATO Science Series II, Vol. 20, Kluwer, Dordrecht, 2001).
33. Schuurman, M. S., Muir, S. R., Allen, W. D. & Schaefer, H. F. Toward subchemical accuracy in computational thermochemistry: Focal point analysis of the heat of formation of NCO and [H,N,C,O] isomers. *J. Chem. Phys.* **120,** 11586–11599 (2004).
34. Gonzales, J. M. *et al.* Definitive ab initio studies of model S_N2 reactions $CH_3X + F^-$ (X = F, Cl, CN, OH, SH, NH_2, PH_2). *Chem. Eur. J.* **9,** 2173–2192 (2003).
35. Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90,** 1007–1023 (1989).
36. Woon, D. E. & Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *J. Chem. Phys.* **103,** 4572–4585 (1995).
37. Perera, A. S. & Bartlett, R. J. Relativistic effects at the correlated level. An application to interhalogens. *Chem. Phys. Lett.* **216,** 606–612 (1993).
38. Balasubramanian, K. *Relativistic Effects in Chemistry* Part A, *Theory and Techniques* (Wiley, New York, 1997).
39. Tarczay, G., Császár, A. G., Klopper, W. & Quiney, H. M. Anatomy of relativistic energy corrections in light molecular systems. *Mol. Phys.* **99,** 1769–1794 (2001).
40. Bomble, Y. J., Stanton, J. F., Kállay, M. & Gauss, J. Coupled-cluster methods including noniterative corrections for quadruple excitations. *J. Chem. Phys.* **123,** 054101 (2005).
41. Kállay, M. & Gauss, J. Approximate treatment of higher excitations in coupled-cluster theory. *J. Chem. Phys.* **123,** 214105 (2005).
42. Evangelista, F. A., Allen, W. D. & Schaefer, H. F. High-order excitations in state-universal and state-specific multireference coupled cluster theories: Model systems. *J. Chem. Phys.* **125,** 154113 (2006).
43. Watson, J. K. G. in *Vibrational Spectra and Structure*, Vol. 6 (ed. Durig, J. R.) 1–89 (Elsevier, New York and Amsterdam, 1977).
44. Mills, I. M. in *Molecular Spectroscopy: Modern Research* (eds Rao, K. N. & Mathews, C. W.) 1–115 (Academic, New York, 1972).
45. Papoušek, D. & Aliev, M. R. *Molecular Vibrational-Rotational Spectra* (Elsevier, Amsterdam, 1982).
46. Nielsen, H. H. The vibration-rotation energies of molecules. *Rev. Mod. Phys.* **23,** 90–136 (1951).
47. Clabo, D. A. Jr, Allen, W. D., Remington, R. B., Yamaguchi, Y. & Schaefer, H. F. A systematic study of molecular vibrational anharmonicity and vibration-rotation interaction by self-consistent-field higher-derivative methods. Asymmetric top molecules. *Chem. Phys.* **123,** 187–239 (1988).
48. Allen, W. D. *et al.* A systematic study of molecular vibrational anharmonicity and vibration-rotation interaction by self-consistent-field higher-derivative methods. Linear polyatomic molecules. *Chem. Phys.* **145,** 427–466 (1990).
49. Schuurman, M. S., Allen, W. D. & Schaefer, H. F. The *ab initio* limit quartic force field of BH_3. *J. Comput. Chem.* **26,** 1106–1112 (2005).
50. DeKock, R. L. *et al.* The electronic structure and vibrational spectrum of *trans*-HNOO. *J. Phys. Chem. A* **108,** 2893–2903 (2004).

51. Czakó, G., Furtenbacher, T., Császár, A. G. & Szalay, V. Variational vibrational calculations using high-order anharmonic force fields. *Mol. Phys.* **102,** 2411–2423 (2004).
52. Fukui, K. A formulation of the reaction coordinate. *J. Phys. Chem.* **74,** 4161–4163 (1970).
53. Gonzales, C. & Schlegel, H. B. Reaction path following in mass-weighted internal coordinates. *J. Phys. Chem.* **94,** 5523–5527 (1990).
54. Allen, W. D., Bodi, A., Szalay, V. & Császár, A. G. Adiabatic approximations to internal rotation. *J. Chem. Phys.* **124,** 224310 (2006).
55. Liboff, R. L. *Introductory Quantum Mechanics* (Addison-Wesley, Reading, Massachusetts, 2003).
56. Razavy, M. *Quantum Theory of Tunneling* (World Scientific, Singapore, 2003).
57. Gray, S. K., Miller, W. H., Yamaguchi, Y. & Schaefer, H. F. Reaction path Hamiltonian: Tunneling effects in the unimolecular isomerization HNC → HCN. *J. Chem. Phys.* **73,** 2733–2739 (1980).

Vol 453|12 June 2008|doi:10.1038/nature07073

LETTERS

Continental mantle signature of Bushveld magmas and coeval diamonds

Stephen H. Richardson[1] & Steven B. Shirey[2]

The emplacement of the 2.05-billion-year-old Bushveld complex, the world's largest layered intrusion and platinum-group element (PGE) repository[1], is a singular event in the history of the Kaapvaal craton of southern Africa[2–4], one of Earth's earliest surviving continental nuclei. In the prevailing model for the complex's mineralization, the radiogenic strontium and osmium isotope signatures of Bushveld PGE ores are attributed to continental crustal contamination of the host magmas[5–11]. The scale of the intrusion and lateral homogeneity of the PGE-enriched layers[1], however, have long been problematical for the crustal contamination model, given the typically heterogeneous nature of continental crust. Furthermore, the distribution of Bushveld magmatism matches that of seismically anomalous underlying mantle[3,12], implying significant interaction before emplacement in the crust. Mineral samples of the ancient 200-km-deep craton keel, encapsulated in macrodiamonds and entrained by proximal kimberlites, reveal the nature of continental mantle potentially incorporated by Bushveld magmas[13,14]. Here we show that sulphide inclusions in ~2-billion-year-old diamonds from the 0.5-billion-year-old Venetia and 1.2-billion-year-old Premier kimberlites (on opposite sides of the complex) have initial osmium isotope ratios even more radiogenic than those of Bushveld sulphide ore minerals[6,15]. Sulphide Re–Os and silicate Sm–Nd and Rb–Sr isotope compositions indicate that continental mantle harzburgite and eclogite components, in addition to the original convecting mantle magma, most probably contributed to the genesis of both the diamonds and the Bushveld complex. Coeval diamonds provide key evidence that the main source of Bushveld PGEs is the mantle rather than the crust.

Archaean cratons contain a preponderance of Earth's precious-metal and diamond mineralization. In particular, the Kaapvaal craton hosts not only the largest known gold and platinum deposits, namely the Witwatersrand basin and the Bushveld complex, but also some of the economically most significant diamond deposits, in surrounding kimberlite pipes (Fig. 1). The evolutionary history of the Kaapvaal craton includes amalgamation of the ~3.7-Gyr-old Witwatersrand and 3.2-Gyr-old Kimberley blocks (Fig. 1) by subduction–accretion and continent–continent collision processes around 2.9 Gyr ago[16,17]. The Kaapvaal craton subsequently collided with the Zimbabwe craton some 2.5–2.7 Gyr ago[4], to produce the Limpopo belt, a highly strained metamorphic terrane that sutures the cratons. Archaean crust in the central zone of the Limpopo belt then experienced a second major tectonothermal event ~2.0 Gyr ago[18]. The age of the latter event overlaps the well-constrained 2.054-Gyr emplacement age (ref. 4; also J. S. Scoates and R. M. Friedman, manuscript in preparation) of the Bushveld complex.

The Bushveld PGE mineralization is concentrated in the Merensky reef and UG2 chromitite within the critical zone of the Rustenburg layered suite (RLS), which has an overall volume of around 500,000 km^3 and a MgO-rich (picritic to boninitic) bulk composition[1,19,20]. The critical zone represents the transition between the lower and main zones of the RLS and is marked by Sr isotopic gradients[5,21–23], inter-mineral Nd and Pb isotopic disequilibrium[9,10,24] and PGE sulphide minerals—laurite (RuS_2) and erlichmanite (OsS_2)—with distinctly different Os isotope signatures[15] (Fig. 2). These attributes suggest at least two sources for the PGEs that were mixed into successive Bushveld magmas, as well as late-stage modi-

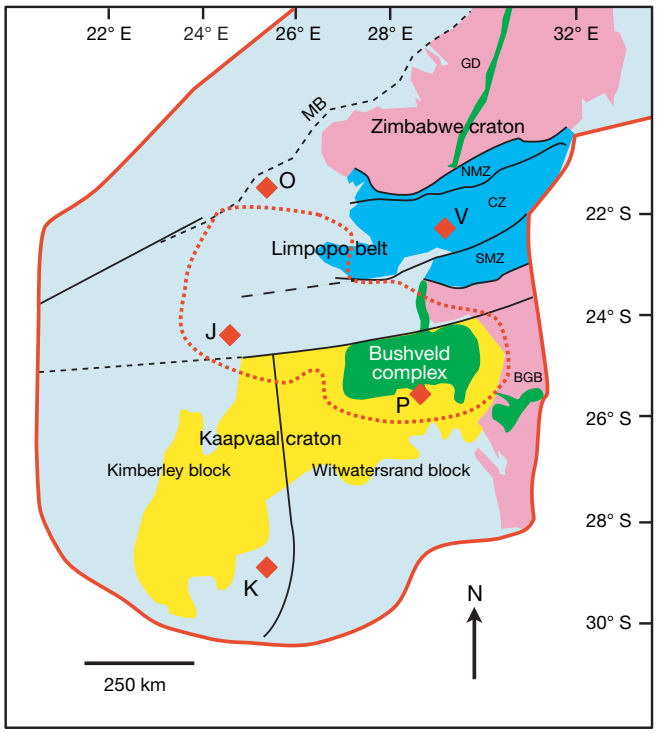


Figure 1 | Sketch map of the Kaapvaal and Zimbabwe cratons (light blue) and intervening Limpopo belt (dark blue) of southern Africa (following ref. 33). Positions of the 0.09-Gyr-old Kimberley (denoted K), 0.24-Gyr-old Jwaneng (J), 0.09-Gyr-old Orapa (O), 0.52-Gyr-old Venetia (V) and 1.2-Gyr-old Premier (P) kimberlites are shown in relation to the 2.05-Gyr-old Bushveld complex and other crustal units including exposed Archaean basement gneisses (pink), Transvaal and other supracrustal sequences older than 2.0 Gyr (yellow), the Barberton greenstone belt (BGB), the Magondi belt (MB), the Great Dyke (GD), and the northern marginal zone (NMZ), the central zone (CZ) and the southern marginal zone (SMZ) of the Limpopo belt. The red dotted line surrounds the region of seismically slow cratonic mantle at a depth of 150 km resulting from Bushveld–Molopo magmatism[3,12].

[1]Department of Geological Sciences, University of Cape Town, Rondebosch 7701, South Africa. [2]Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015, USA.

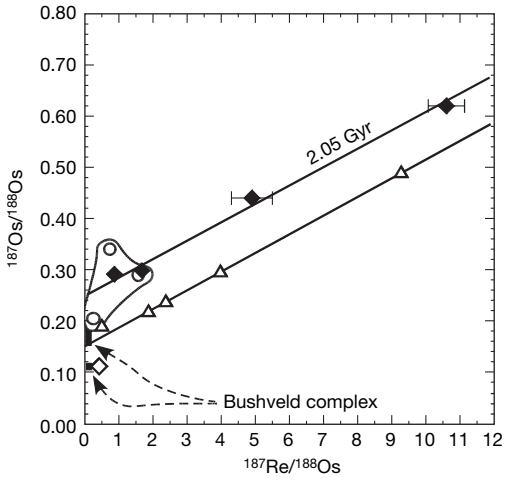


Figure 2 | Re–Os isochron diagram for sulphide inclusions in Premier and Venetia diamonds. Re–Os isotope data for Premier eclogitic (open circles) and Venetia eclogitic (filled diamonds) and peridotitic (open diamond) sulphides are shown in relation to a 2.05-Gyr reference isochron corresponding to the emplacement age of the Bushveld complex. Error bars are the sum of Re and Os run precisions after isotope-dilution error magnification and blank corrections (see Table 1), and are smaller than the size of the plotted points except where otherwise indicated. Also shown are the fields for Bushveld PGE sulphide ore minerals[6,15] and a 2.04-Gyr isochron defined by pyroxenites (open triangles)[8] from the Bastard unit overlying the Merensky reef of the Bushveld complex.

fication of the PGE-rich layers by deuteric fluids. Even so, individual layers within the RLS are well correlated over a distance of at least 300 km between the eastern and western limbs of the complex, exemplified by a remarkable consistency in Sr isotopic composition with stratigraphic level. In particular, the PGE-rich Merensky reef, near the top of the critical zone, has the same initial $^{87}Sr/^{86}Sr$ ratio (0.7064) at the base of the horizon on opposite sides of the complex[5,23].

Radiogenic isotope studies of sulphide- and silicate-inclusion-bearing diamonds and related xenolith minerals, entrained by kimberlite magmas at 150–200-km depths in ancient continental mantle, show that major craton evolution events may be marked by coeval diamond crystallization[12,14,16]. Whereas xenoliths are open systems subject to metasomatism during residence in the continental mantle and alteration during kimberlite eruption[25], inclusions remain closed systems with negligible diffusion of radiogenic isotope tracers through the host diamond, even at mantle temperatures on a billion-year timescale[13,26]. Furthermore, studies of mantle xenoliths show that sulphide minerals are the main carrier phase for PGEs in upper-mantle rocks[27]. Therefore, sulphide inclusions in ~2.0-Gyr-old diamonds (Fig. 2) from the Premier (on the southern margin of the Bushveld complex) and Venetia (in the central zone of the Limpopo belt) kimberlites provide a way of characterizing the continental mantle PGE input to Bushveld magmas that avoids the secondary effects seen in xenoliths.

The three Premier and six Venetia sulphides (Table 1) are all assemblages of pyrrhotite and chalcopyrite plus/minus pentlandite derived from monosulphide solid solution by exsolution during cooling *en route* to the surface. The smallest sample (VEN-10; 4 μg) is a relatively high-Ni (>20 wt%) sulphide typical of peridotitic sulphide inclusions that are rarely seen in southern Africa but are more abundant in harzburgitic diamonds from the Siberian and Slave cratons[26]. The other eight samples (10–89 μg) are sulphides with low Ni contents (<10 wt%) typical of eclogitic inclusion assemblages from other Kaapvaal localities[16,28], but with an unusually wide range of initial ('common') Os concentrations (5–10,000 p.p.b.) extending to much higher values than normal for eclogitic sulphides.

The Venetia eclogitic sulphide with the lowest common Os content and highest Re/Os ratio has an extremely radiogenic present-day $^{187}Os/^{188}Os$ ratio (~52) giving a depleted mantle model age (T_{DM})[25] of 4.16 ± 0.26 Gyr that is highly sensitive to blank correction (see Table 1). Even so, this age suggests precursor formation and residence in the oldest part of the local continental lithospheric mantle keel, which has a minimum age of 3.7 Gyr on the basis of the oldest rhenium depletion model age (T_{RD}) for Venetia peridotite xenoliths[25]. Similarly old ages have been reported for the ancient gneiss complex (south of the Barberton greenstone belt; see Fig. 1) in the eastern Kaapvaal (3.4–3.7 Gyr)[4] and the lithospheric mantle beneath the Zimbabwe craton (3.2–3.9 Gyr)[29], supporting the idea that very old subcontinental lithospheric mantle (SCLM) occurs in this region.

The other four Venetia eclogitic sulphides show a range of lower Re/Os ratios and less radiogenic $^{187}Os/^{188}Os$ ratios (0.293–0.621) that describe an array approximating a 2.05-Gyr reference isochron (Fig. 2), but with scatter outside analytical errors (Isoplot Model 3 age of 2,053 ± 520 Myr (uncertainty, ±2 s.d.); initial $^{187}Os/^{188}Os$ ratio of 0.256 ± 0.049 (uncertainty, ±2 s.d.); mean sum of weighted deviates equals 17), implying significant initial $^{187}Os/^{188}Os$ variation. Even so, the initial $^{187}Os/^{188}Os$ ratio of the array is convincingly radiogenic relative to the contemporaneous convecting mantle value (0.1145), on the basis of the global mantle growth curve[25] and the single Venetia peridotitic sulphide (Fig. 2). Three Premier sulphides with a more limited spread in low Re/Os ratio scatter above and below the 2.05-Gyr Venetia array (Fig. 2), suggesting that a radiogenic initial $^{187}Os/^{188}Os$ ratio was a feature of part of the regional sub-Bushveld mantle at this time.

The range of initial $^{187}Os/^{188}Os$ values for 2.05-Gyr-old Premier and Venetia sulphide inclusions (0.199–0.315; Table 1) is even more elevated, relative to convecting mantle, than the equivalent values for pyroxenites (0.151–0.177)[8] and laurites (0.153–0.192)[6,15] from the Merensky reef and immediately overlying Bastard unit of the Bushveld complex (Fig. 2). The radiogenic Os and Sr isotope signatures of the RLS rocks and minerals (Fig. 3), together with their elevated $\delta^{18}O$ values (~1‰ higher than the convecting mantle value; $\delta^{18}O = [(^{18}O/^{16}O)_{sample}/(^{18}O/^{16}O)_{SMOW} - 1] \times 10^3$; SMOW, standard mean ocean water)[11], have been attributed to crustal contamination[5–11]. Although Archaean upper crust appears to satisfy both the Re–Os and the Rb–Sr isotope systems in the crustal contamination model, more than 20 wt% upper crust is required because

Table 1 | Re–Os isotopic composition of sulphide inclusions in diamonds from the Premier and Venetia kimberlites

	Wt (μg)	Re (p.p.b.)	Os (p.p.b.)	Os_i (p.p.b.)	$^{187}Re/^{188}Os$	$^{187}Os/^{188}Os$	$^{187}Os/^{188}Os_i$
PREM-01	12.8	289	905	891	1.57(18)	0.2918(7)	0.2373
PREM-13	10.5	387	2,593	2,537	0.739(64)	0.3411(6)	0.3154
PREM-17	25.2	536	10,560	10,510	0.247(19)	0.2072(1)	0.1986
VEN-1	88.7	667	35.07	4.6	728(12)	52.3(24)	27.0
VEN-2	24.3	464	1,368	1,346	1.67(20)	0.2995(9)	0.2415
VEN-3	18.2	101	48.78	46.1	10.61(53)	0.6208(40)	0.2522
VEN-6	47.1	1,180	1,208	1,167	4.9(6)	0.4405(3)	0.2703
VEN-7	17.3	266	1,515	1,492	0.865(76)	0.2933(7)	0.2632
VEN-10	4.4	257	2,928	—	0.4222(34)	0.1127(3)	0.0980

Wt, weight. Re and Os concentrations are in ng g^{-1} after correction for minimum blanks: Re, 60 fg; Os, 4 fg. Common Os concentrations (Os_i) are calculated assuming chondritic $^{187}Os/^{188}Os$ ratios. Errors in $^{187}Re/^{188}Os$ (least significant digits in parentheses) are the sum of Re and Os run precisions (2 $s.d._{mean}$) after isotope-dilution error magnification. Errors in $^{187}Os/^{188}Os$ are the sum of run precision and uncertainty in Os blank (4 ± 2 fg) correction. $^{187}Os/^{188}Os$ ratios projected back 2.05 Gyr ($^{187}Os/^{188}Os_i$) are calculated using a ^{187}Re decay constant of 1.666×10^{-11} yr^{-1}.

of the low PGE content of the crust[8]. Furthermore, the upper crust is heterogeneous on the outcrop scale and implausibly thorough mixing of local contaminants would be required to explain the 300-km-scale lateral homogeneity of features like the Merensky reef[5,23]. Indeed, demonstrable local effects such as hydrothermal circulation are more likely to produce isotopic heterogeneity[24] than homogeneity. Contamination involving lower crust in a sub-RLS magma chamber is another way to homogenize magmas before emplacement in the upper crust[7]. However, lower-crustal rocks are unlikely to be able to supply the necessary radiogenic Sr and Os on the required scale[8].

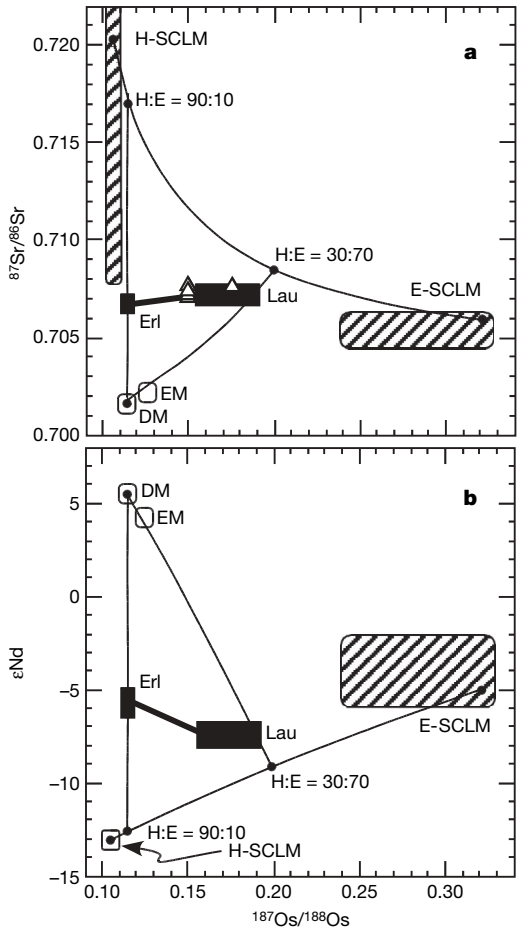


Figure 3 | Isotope correlation diagrams comparing Bushveld rocks and PGE sulphide ore minerals[6,8,9,15] to potential mantle source components 2.05 Gyr ago. a, Sr–Os. The harzburgitic SCLM (H-SCLM) field is defined by $^{187}Os/^{188}Os$ ratios of selected Kaapvaal peridotite xenoliths[25] projected back 2.05 Gyr coupled with initial $^{87}Sr/^{86}Sr$ ratios (0.708–0.780) of subcalcic garnet macrocrysts (disaggregated harzburgitic diamond host rocks) from Premier and Venetia kimberlite heavy-mineral concentrates[14] (S.H.R., unpublished). The eclogitic SCLM (E-SCLM) field is based on initial $^{187}Os/^{188}Os$ ratios of sulphide inclusions in Premier and Venetia diamonds (Table 1) coupled with initial $^{87}Sr/^{86}Sr$ of lherzolitic garnet and clinopyroxene inclusions in Premier diamonds[14]. The pyroxenites (open triangles) are Bushveld rocks with low Os concentrations[8]. **b**, Nd–Os. The Nd isotopic composition ($\varepsilon Nd = [(^{143}Nd/^{144}Nd)_{sample}/(^{143}Nd/^{144}Nd)_{CHUR} - 1] \times 10^4$; CHUR, chondritic uniform reservoir) of H-SCLM is modelled on Sm–Nd isotopic evolution of harzburgitic garnet inclusions in Kimberley and Finsch diamonds[13] from western Kaapvaal SCLM unaffected by Bushveld modification[3,12,14]. The εNd of E-SCLM is approximated by $^{143}Nd/^{144}Nd$ ratios of lherzolitic inclusions in Premier diamonds[14] projected back 2.05 Gyr. In **a** and **b**, selected mixing curves (based on inferred Kaapvaal SCLM component[13,14,25,28] melt and Bushveld magma[8,9] Sr, Nd and Os concentrations) are drawn between convecting mantle melt (depleted mantle (DM) or enriched mantle (EM): 240 p.p.m. Sr, 10 p.p.m. Nd, 0.1 p.p.b. Os) and H- and E-SCLM melts (H-SCLM: 150 p.p.m. Sr, 25 p.p.m. Nd, 0.3 p.p.b. Os; E-SCLM: 300 p.p.m. Sr, 10 p.p.m. Nd, 0.1 p.p.b. Os). Simple mass-balance calculations suggest that the early erlichmanites (Erl) represent a mix of >60% DM melt and <40% SCLM melt comprising harzburgitic (H) and eclogitic (E) components in the ratio 90:10, whereas the later laurites (Lau) represent a mix of <25% DM melt and >75% SCLM melt comprising H and E in the ratio 30:70.

An alternative explanation for the radiogenic isotope compositions of the Bushveld complex involves the equilibration of Bushveld magmas with the Kaapvaal SCLM before emplacement at higher levels in the crust[21,22,30]. The Os, Sr and Nd isotopic compositions of the local sub-Bushveld mantle can then be estimated using sulphide and silicate minerals in diamonds and mantle xenoliths (Fig. 3). In this model, a melt of picritic or boninitic composition, as required by the major element content of the RLS[19], is derived from the convecting mantle below the SCLM. This magma heats and preferentially removes the low-temperature-melting sulphide and other metasomatic phases from harzburgite and eclogite comprising the Archaean mantle keel of the Kaapvaal craton. Seismic tomography of the Kaapvaal and Zimbabwe cratons shows that P-wave velocities at a depth of 150 km beneath the Bushveld complex (Fig. 1) are 1.0% to 1.5% lower than those in surrounding lithospheric mantle[3]. In addition, eclogitic diamonds predominate over peridotitic diamonds in this seismically slow lithosphere[12]. These features indicate that the SCLM was profoundly modified by interaction with Bushveld magmas during emplacement of the RLS.

Numerous mantle xenolith and diamond studies indicate that the Kaapvaal SCLM is composed of harzburgite with unradiogenic Os and Nd and highly radiogenic Sr isotope signatures, interspersed with lesser eclogites with characteristically radiogenic Os isotope signatures[13,25,28]. The corresponding Sr and Nd isotopic compositions of the eclogitic SCLM are less well known but can be inferred using ~2.0-Gyr-old lherzolitic inclusions in Premier diamonds[14] as proxies for this component. Such inclusions represent the addition of a basaltic component into the Archaean SCLM in the vicinity of the Bushveld complex[12,14].

In the mantle contamination model, parental magmas of the RLS originating in convecting mantle incorporate variable proportions of harzburgitic and eclogitic SCLM components (Fig. 3) *en route* to emplacement in the crust. The more magnesian character of the lower zone indicates a greater contribution from the harzburgitic SCLM component, whereas the more aluminous character of the main zone is consistent with a greater contribution from the eclogitic SCLM component. The generally anomalous oxygen isotope composition of the RLS[11] may be attributable to variable incorporation of the latter component, given the positively skewed distribution of mantle eclogite $\delta^{18}O$ values reflecting seafloor alteration of Archaean oceanic crustal precursors[31]. Even so, the elevated $\delta^{18}O$ values of at least some RLS rocks are almost certainly due in part to the effects of local continental crustal contamination[11]. In this case, their Sr, Nd and Os compositions would also be locally affected[8,9].

The multicomponent SCLM contamination model (with or without crust) may also apply to other PGE-enriched intrusions such as the 2.7-Gyr-old Stillwater complex (Wyoming craton), the 2.5-Gyr-old Great Dyke (Zimbabwe craton) and the 1.3-Gyr-old Muskox intrusion (Slave craton), given that their parental magma types (ultramafic versus tholeiitic (or anorthositic))[21,32] are comparable in the same way as the lower and main zones of the RLS. The new constraints from coeval diamonds indicate that continental mantle harzburgite and eclogite are underappreciated yet critical components in the genesis of the Bushveld complex and similar large PGE-ore-bearing intrusions.

METHODS SUMMARY

Sulphide inclusions in diamonds invariably show large rosette fracture systems around each grain caused by late-stage differential decompression during

kimberlite eruption. Such fracture systems may extend to the exterior of the diamond, exposing the inclusion to alteration by kimberlitic fluids. This is common in Premier diamonds, resulting in a paucity of uncompromised specimens from the Premier kimberlite. For example, pilot study sample PREM-1 contained two sulphide inclusions with surface alteration to hematite (though the Re–Os systematics of the interior of the larger inclusion do not appear to have been seriously affected). Therefore, all diamond samples with possible sulphide inclusions were cut and polished into plates to check the integrity of the inclusions. Selected plates were then cleaved in a tungsten-carbide-lined steel cracker to release the inclusions.

Sulphide inclusions often break up during liberation from the host diamond, allowing for characterization of exterior and interior surfaces before recombination for Re–Os isotope analysis. All macroscopic black fragments recovered from the cracker were mounted directly onto carbon conductive tape for positive identification in a scanning electron microscope using back-scattered electron imaging and energy-dispersive X-ray fluorescence spectrometry. After inclusion characterization and cleaning, a mixed ^{185}Re–^{190}Os tracer solution was added and Re and Os analysed using microchemistry and negative thermal ionization techniques outlined in refs 16 and 26.

Received 21 December 2007; accepted 28 April 2008.

1. Eales, H. V. & Cawthorn, R. G. in *Layered Intrusions* (ed. Cawthorn, R. G.) 181–229 (Elsevier Science, Amsterdam, 1996).
2. de Wit, M. J. *et al.* Formation of an Archaean continent. *Nature* **357,** 553–562 (1992).
3. James, D. E., Fouch, M. J., VanDecar, J. C., van der Lee, S. & Group, K. S. Tectospheric structure beneath southern Africa. *Geophys. Res. Lett.* **28,** 2485–2488 (2001).
4. Eglington, B. M. & Armstrong, R. A. The Kaapvaal Craton and adjacent orogens, southern Africa: a geochronological database and overview of the geological development of the craton. *South Afr. J. Geol.* **107,** 13–32 (2004).
5. Kruger, F. J. & Marsh, J. S. Significance of ^{87}Sr/^{86}Sr ratios in the Merensky cyclic unit of the Bushveld Complex. *Nature* **298,** 53–55 (1982).
6. McCandless, T. E. & Ruiz, J. Osmium isotopes and crustal sources for platinum-group mineralization in the Bushveld Complex, South Africa. *Geology* **19,** 1225–1228 (1991).
7. McCandless, T. E., Ruiz, J., Adair, B. I. & Freydier, C. Re-Os isotope and Pd/Ru variations in chromitites from the Critical Zone, Bushveld Complex, South Africa. *Geochim. Cosmochim. Acta* **63,** 911–923 (1999).
8. Schoenberg, R., Kruger, F. J., Nägler, T. F., Meisel, T. & Kramers, J. D. PGE enrichment in chromitite layers and the Merensky Reef of the western Bushveld Complex; a Re-Os and Rb-Sr isotope study. *Earth Planet. Sci. Lett.* **172,** 49–64 (1999).
9. Maier, W. D., Arndt, N. T. & Curl, E. A. Progressive crustal contamination of the Bushveld Complex; evidence from Nd isotopic analyses of the cumulate rocks. *Contrib. Mineral. Petrol.* **140,** 316–327 (2000).
10. Prevec, S. A., Ashwal, L. D. & Mkaza, M. S. Mineral disequilibrium in the Merensky Reef, western Bushveld Complex, South Africa; new Sm-Nd isotopic evidence. *Contrib. Mineral. Petrol.* **149,** 306–315 (2005).
11. Harris, C., Pronost, J. J. M., Ashwal, L. D. & Cawthorn, R. G. Oxygen and hydrogen isotope stratigraphy of the Rustenburg Layered Suite, Bushveld Complex; constraints on crustal contamination. *J. Petrol.* **46,** 579–601 (2005).
12. Shirey, S. B. *et al.* Diamond genesis, seismic structure, and evolution of the Kaapvaal-Zimbabwe craton. *Science* **297,** 1683–1686 (2002).
13. Richardson, S. H., Gurney, J. J., Erlank, A. J. & Harris, J. W. Origin of diamonds in old enriched mantle. *Nature* **310,** 198–202 (1984).
14. Richardson, S. H., Harris, J. W. & Gurney, J. J. Three generations of diamonds from old continental mantle. *Nature* **366,** 256–258 (1993).
15. Hart, S. R. & Kinloch, E. D. Osmium isotope systematics in Witwatersrand and Bushveld ore deposits. *Econ. Geol.* **84,** 1651–1655 (1989).
16. Richardson, S. H., Shirey, S. B., Harris, J. W. & Carlson, R. W. Archean subduction recorded by Re-Os isotopes in eclogitic sulfide inclusions in Kimberley diamonds. *Earth Planet. Sci. Lett.* **191,** 257–266 (2001).
17. Schmitz, M. D., Bowring, S. A., de Wit, M. J. & Gartz, V. Subduction and terrane collision stabilized the western Kaapvaal craton tectosphere 2.9 billion years ago. *Earth Planet. Sci. Lett.* **222,** 363–376 (2004).
18. Boshoff, R. *et al.* Geological history of the Central Zone of the Limpopo Complex: the West Alldays area. *J. Geol.* **114,** 699–716 (2006).
19. Davies, G., Cawthorn, R. G., Barton, J. M. Jr & Morton, M. Parental magma to the Bushveld Complex. *Nature* **287,** 33–35 (1980).
20. Cawthorn, R. G. & Walraven, F. Emplacement and crystallization time for the Bushveld Complex. *J. Petrol.* **39,** 1669–1687 (1998).
21. Hamilton, P. J. Sr. isotope and trace element studies of the Great Dyke and Bushveld mafic phase and their relation to early Proterozoic magma genesis in southern Africa. *J. Petrol.* **18,** 24–52 (1977).
22. Harmer, R. E. & Sharpe, M. R. Field relations and strontium isotope systematics of the marginal rocks of the eastern Bushveld Complex. *Econ. Geol.* **80,** 813–837 (1985).
23. Lee, C. A. & Butcher, A. R. Cyclicity in the Sr isotope stratigraphy through the Merensky and Bastard Reef units, Atok section, eastern Bushveld Complex. *Econ. Geol.* **85,** 877–883 (1990).
24. Mathez, E. A. & Kent, A. J. R. Variable initial Pb isotopic compositions of rocks associated with the UG2 chromitite, eastern Bushveld Complex. *Geochim. Cosmochim. Acta* **71,** 5514–5527 (2007).
25. Carlson, R. W. *et al.* in *The J. B. Dawson Volume (Proc. 7th Internat. Kimberlite Conf.)* (eds Gurney, J. J., Gurney, J. L., Pascoe, M. D. & Richardson, S. H.) 99–108 (Red Roof Design, Cape Town, 1999).
26. Westerlund, K. J. *et al.* A subduction wedge origin for Paleoarchean peridotitic diamonds and harzburgites from the Panda kimberlite, Slave craton: evidence from Re-Os isotope systematics. *Contrib. Mineral. Petrol.* **152,** 275–294 (2006).
27. Mitchell, R. H. & Keays, R. R. Abundance and distribution of gold, palladium and iridium in some spinel and garnet lherzolites: implications for the nature and origin of precious metal-rich intergranular components in the upper mantle. *Geochim. Cosmochim. Acta* **45,** 2425–2433 (1981).
28. Shirey, S. B. *et al.* Archean emplacement of eclogitic components into the lithospheric mantle during formation of the Kaapvaal Craton. *Geophys. Res. Lett.* **28,** 2509–2512 (2001).
29. Nägler, T. F., Kramers, J. D., Kamber, B. S., Frei, R. & Prendergast, M. D. A. Growth of subcontinental lithospheric mantle beneath Zimbabwe started at or before 3.8 Ga; Re-Os study on chromites. *Geology* **25,** 983–986 (1997).
30. Allègre, C. J. & Luck, J. M. Osmium isotopes as petrogenetic and geological tracers. *Earth Planet. Sci. Lett.* **48,** 148–154 (1980).
31. Sharp, Z. D. *Principles of Stable Isotope Geochemistry* Ch. 11 (Prentice Hall, Englewood Cliffs, New Jersey, 2006).
32. Lambert, D. D. *et al.* Re-Os and Sm-Nd isotope geochemistry of the Stillwater Complex, Montana; implications for the petrogenesis of the J-M Reef. *J. Petrol.* **35,** 1717–1753 (1994).
33. McCourt, S., Kampunzu, A. B., Bagai, Z. & Armstrong, R. A. The crustal architecture of Archaean terranes in northeastern Botswana. *South Afr. J. Geol.* **107,** 147–158 (2004).

Acknowledgements We thank De Beers for donation of rare diamond specimens; R. W. Carlson and J. W. Harris for collaboration on related diamond projects; R. J. Hemley, C. Yan and J. Lai for access to diamond cutting and polishing facilities; and T. Mock, M. Horan and C. Hadidiacos for analytical laboratory help. This work was supported by the Carnegie Institution of Washington and NSF grant EAR-0310059 to S.B.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.H.R. (steve.richardson@uct.ac.za).

Vol 453|12 June 2008|doi:10.1038/nature06719

LETTERS

Ultrasonic frogs show hyperacute phonotaxis to female courtship calls

Jun-Xian Shen[1], Albert S. Feng[2], Zhi-Min Xu[1], Zu-Lin Yu[1], Victoria S. Arch[3], Xin-Jian Yu[5] & Peter M. Narins[3,4]

Sound communication plays a vital role in frog reproduction[1,2], in which vocal advertisement is generally the domain of males. Females are typically silent, but in a few anuran species they can produce a feeble reciprocal call[3] or rapping sounds[4] during courtship. Males of concave-eared torrent frogs (*Odorrana tormota*) have demonstrated ultrasonic communication capacity[5]. Although females of *O. tormota* have an unusually well-developed vocal production system[6], it is unclear whether or not they produce calls or are only passive partners in a communication system dominated by males. Here we show that before ovulation, gravid females of *O. tormota* emit calls that are distinct from males' advertisement calls, having higher fundamental frequencies and harmonics and shorter call duration. In the field and in a quiet, darkened indoor arena, these female calls evoke vocalizations and extraordinarily precise positive phonotaxis (a localization error of $<1^\circ$), rivalling that of vertebrates with the highest localization acuity (barn owls[7,8], dolphins, elephants and humans[9]). The localization accuracy of *O. tormota* is remarkable in light of their small head size (interaural distance of <1 cm), and suggests an additional selective advantage of high-frequency hearing beyond the ability to avoid masking by low-frequency background noise[5].

Odorrana tormota (formerly *Amolops tormotus*)[10] is an arboreal, nocturnal frog living near noisy streams in Huangshan Hot Springs, China. On rainy nights when ovulation is imminent, gravid females can be found in numbers in their natural habitat. We recorded the vocalizations of captive females kept in a quiet darkened room using an ultrasonic microphone and PCTape, a computer-based data acquisition system[11]. In captivity, gravid females vocalize rarely (<1 call per hour on average), but they stop calling altogether after ovulation. This suggests that these vocalizations probably serve to advertise their gravid reproductive state.

A representative female call consists of an upward and downward frequency modulated (FM) tone pip having a fundamental frequency from 7.2 to 9.8 kHz with multiple harmonics that extend into the ultrasonic range (Fig. 1a). The calls are moderately intense, with a peak intensity of ~80 dB sound pressure level (SPL), measured 1 m away from a calling female frog. The call duration is generally short (<150 ms), and the fundamental frequency of female calls is higher than that of the males' advertisement calls[11,12] (see below) despite females having an unusually larger body size and more robust vocal cords[6]. Also, unlike males' advertisement calls, female calls do not display nonlinear characteristics (for example, chaos, subharmonics or frequency jumps).

To determine the functional significance of female calls, we carried out acoustic playback experiments in the field and in a quiet room indoors. In the field, broadcasting female calls at the rate of 1 call per 15 s at ~90 dB SPL (measured at 50 cm from the playback loudspeaker) consistently induced nearby males to call antiphonally or to increase their calling activities. Moreover, most of these males displayed positive phonotaxis—they rapidly approached the loudspeaker with high-precision. Results of the playback experiments in the field suggest that a female call functions as a courtship call; emission of such calls not only signals a female's receptivity to males but also her location.

To validate the above hypothesis and to assess the acuity of sound localization quantitatively, we repeated the acoustic playback experiments in a quiet arena in a darkened room, this time to individual captive males. In response to female calls (intensity ~80 dB SPL measured at 1 m from the loudspeaker, presented at a rate of 1 call per 15 s), males produced four types of vocal responses during phonotaxis (see Supplementary Videos). The most prominent was the antiphonal response, which was time-locked (within 60.4 ± 31.7 ms from 459 antiphonal responses from 48 males) to the stimulus offset. The majority of antiphonal responses comprised a short note (duration ≤ 40 ms), with prominent upward and downward FM sweeps and multiple harmonics (Fig. 1b). The fundamental frequency (F_0) of the antiphonal responses varied with responding individuals, from 6.2 to 7.8 kHz (Fig. 1b). This variation was a function of the ambient noise level; in the field where it is noisy, males increased the F_0 of their antiphonal responses presumably in order to decrease acoustic masking by low-frequency background interference[11]. The population mean of F_0 is 7.24 ± 0.96 kHz ($n = 43$) in the field—this is significantly higher than the value observed indoors, 5.96 ± 1.01 kHz ($n = 48$; $t = 6.21$, $P < 0.0001$).

Whereas the antiphonal response typically occurred before males jumped towards the loudspeaker broadcasting the female call, responding males switched to another type of vocalization midway during the phonotaxis—namely, the staccato call, which consisted of up to 60 triple pulses repeated at the rate of 14 Hz, lasting 3–4 s, each pulse with a duration of ~3 ms, and a dominant frequency of ~4.0 kHz (Fig. 1c). Following production of a staccato call, the male moved rapidly or hopped towards the loudspeaker. After an antiphonal response and before the next female call stimulus, males often produced two or three short (~50 ms) shallow FM pips. When close to the loudspeaker (20 cm), males sometimes produced a low-level long (250–300 ms) call having a fundamental frequency of ~4–5 kHz and many subharmonics.

In addition to evoked vocal responses, the female call also consistently triggered positive phonotaxis. Analysis of video recordings of males' phonotactic paths in the indoor arena showed that males' localization acuity was extraordinary. On hearing a female call, a male usually oriented his body towards the loudspeaker — this was followed by a long-distance hop towards the loudspeaker (Fig. 2; see Supplementary Video). The precision of the long-distance (range: 30–75 cm) hops was remarkable, with an average azimuthal error of just $0.7^\circ \pm 3.3^\circ$ ($n = 41$). Males' localization acuity rivals that of

[1]State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. [2]Department of Molecular & Integrative Physiology and Beckman Institute, University of Illinois, Urbana, Illinois 61801, USA. [3]Department of Ecology & Evolutionary Biology, [4]Department of Physiological Science, University of California, Los Angeles, California 90095, USA. [5]Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China.

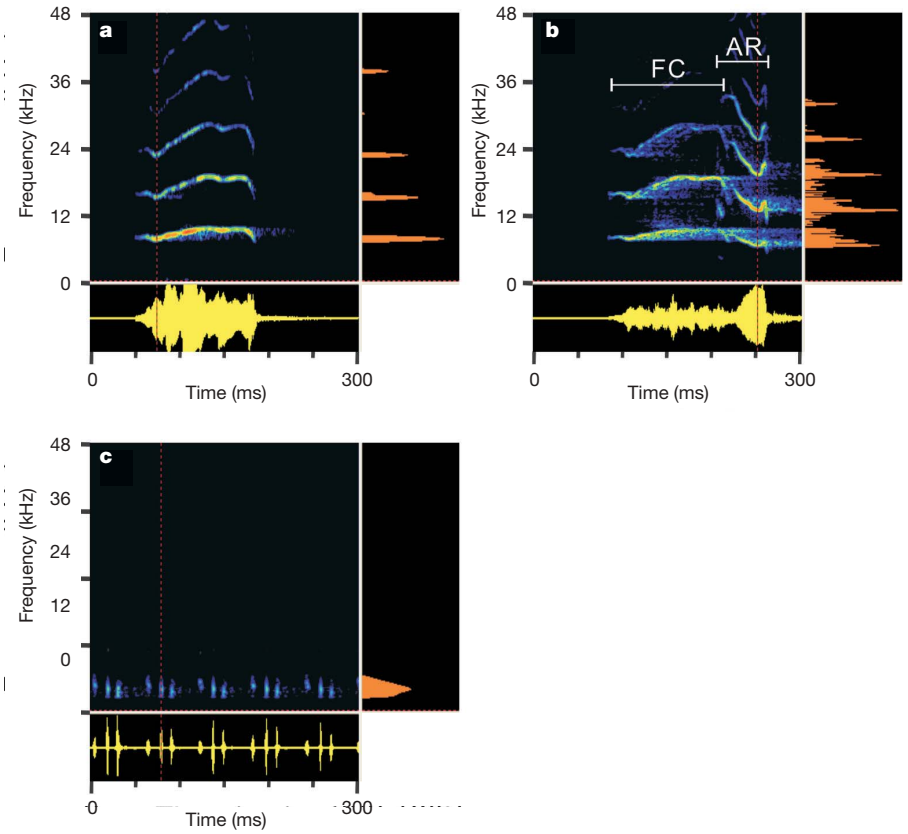


Figure 1 | Female's courtship call and male's evoked vocal responses. **a**, A female's courtship call (FC) (duration 134 ms), showing its high-fundamental frequency and rich harmonics. **b**, A male's antiphonal response (AR) emitted 8 ms before the offset of the FC. **c**, A staccato call, comprising a string of soft clicks at low frequencies. In each panel of this figure, the sound spectrogram is shown at top left, the signal waveform at bottom left, and the instantaneous amplitude spectrum at the right.

the best vertebrate performers (barn owls[7,8], elephants, dolphins and humans[9]), despite their very short interaural distance (<1 cm).

O. tormota inhabits the vegetation along noisy streams, rendering communication problematic. Males have been shown to produce calls with spectral energy extending into the ultrasonic range to avoid masking from the broadband ambient noise; the fundamental frequency of male calls is around 5–7 kHz on average[5,6,11,12]. That females of *O. tormota* produce moderate-level calls having a significantly higher F_0 (7.2–9.8 kHz) than male calls is surprising. First, females of *O. tormota* not only possess distinct vocal cords, but their vocal cords are bigger and their medial vocal ligaments are thicker compared to their male counterparts[6]. Generally, the mass of the vocal cords tends to be correlated to body size, so the larger female frogs would be expected to produce calls with a lower dominant frequency[13–17]. Further studies are therefore necessary to determine the mechanisms underlying high-frequency sound production in females. Second, for most anuran species, females are silent, and those that vocalize generally possess a rudimentary larynx and thus can emit low-level calls or rapping, or simple release calls[3,4,18–22]. The high-frequency and moderately intense mating calls of female *O. tormota* are thus exceptional. Our results show that they promote reproduction by unambiguously conveying receptive state and location in the presence of a noisy background. This finding further supports the hypothesis that the frogs produce high-frequency components to overcome the masking effects of ambient noise, to increase the salience of the communication signal, to facilitate amplexus and to ensure successful reproduction in the field. Little is known, however, of the mechanisms that underlie production of the higher F_0 in the females.

Another surprising finding is that the female call is a universal signal used to both stimulate vocal responses from and attract conspecific individual males distributed at different sectors along the Tau

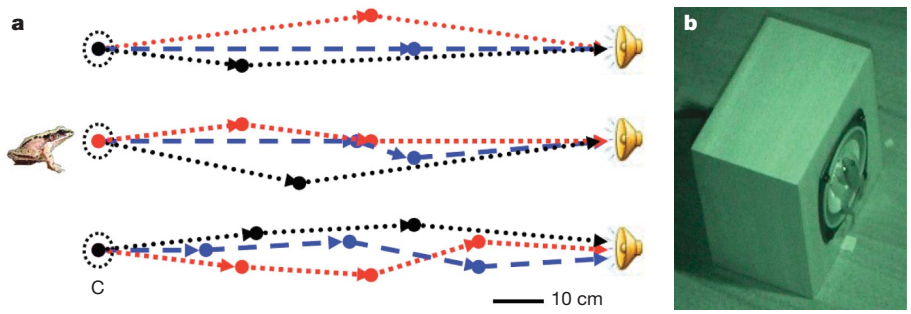


Figure 2 | Males' phonotactic responses to a female's courtship call. **a**, Representative phonotactic orienting responses from six males. Circles represent the landing site of a hop; the arrow indicates the hop's direction towards the loudspeaker broadcasting the female's courtship call. C, glass cover. **b**, Image showing a male frog reaching the centre of the diaphragm of the loudspeaker (taken from a video recording of the frog's phonotaxic responses).

Hua Creek in Huangshan Hot Springs, China. This implies that males of *O. tormota* have conservative characters for female call pattern recognition and perception, although the neural mechanisms and genetic bases of these characters are unknown. Furthermore, how the female call of *O. tormota* has evolved as a sex-specific and potent signal model, how the auditory system of receivers generates perceptual biases that direct female call evolution, and how the high-frequency mating calls are produced in the female frog, are but three of the unanswered questions about this remarkable animal[5,17,23].

The third surprising finding is that the male *O. tormota* can localize female calls with an extraordinary acuity of just 1° despite their small head size (interaural distance ~1 cm). It is known that barn owls that prey upon small mammals can locate sound with an accuracy of 1–2° (refs 7, 8); they and dolphins, elephants and humans have the highest localization acuity among the vertebrates[9]. Amphibians are less well endowed, generally showing an acuity of ~16–23° (ref. 24), as they locate a sound source based on low-frequency perception. In contrast, ultrasonic males of *O. tormota* have the capacity to perceive higher-frequency sounds as an adaptation to their noisy habitats[5], which may underlie their hyperacute sound localization. Additional mechanisms that underlie localization hyperacuity in these animals remain to be studied.

METHODS SUMMARY

Females of *O. tormota* were caught on rainy nights in Huangshan Hot Springs, China. In a quiet, darkened room at midnight, female frog vocalizations were recorded using an ultrasonic microphone and PCTape, a computer-based data-acquisition system. Data were saved as WAV files and analysed, and displayed using SELENA software. Acoustic playback experiments were carried out at night in the field with a peak ambient noise level of 70–88 dB SPL or in a quiet and darkened indoor room under dim infrared illumination. The WAV file of a recorded female call of *O. tormota* was stored either on an MP3 player for field tests, or on the Play unit for the indoor study. Calls were broadcast through a loudspeaker at ~80–90 dB SPL and a rate of one call per 15 s. The vocal responses of 48 males indoors and 43 males in the field, as well as their spontaneous calling (no stimulus, as control) were recorded using PCTape indoors and a digital audio recorder (frequency range: 10 Hz–96 kHz) in the field with an ultrasonically-sensitive microphone placed 10 cm from the loudspeaker to monitor the frogs' vocalizations as well as the acoustic playback stimulus for later analysis. During the phonotaxis experiments in the indoor arena, each male frog was placed on the floor under a removable glass cover, 1 m in front of the loudspeaker. Video recordings of males' phonotactic behaviours were made under infrared illumination. The trajectories for each male frog were obtained from the video recordings. The hop distance (D, in cm) and azimuthal jump angle (α, in degrees) were measured.

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

Received 6 October 2007; accepted 19 March 2008.
Published online 11 May 2008.

1. Capranica, R. R. *The Evoked Vocal Response of the Bullfrog: A Study of Communication by Sound* (Research Monograph 33, MIT, Cambridge, Massachusetts, 1965).
2. Glaw, F. & Vences, M. *A Fieldguide to the Amphibians and Reptiles of Madagascar* 2nd edn (Vences & Glaw, Köln, Germany, 1994).
3. Roy, D. Communication signals and sexual selection in amphibians. *Curr. Sci.* **72,** 923–927 (1997).
4. Tobias, M. L., Viswanathan, S. & Kelley, D. B. Rapping, a female receptive call, initiates male/female duets in the South African clawed frog. *Proc. Natl Acad. Sci. USA* **95,** 1870–1875 (1998).
5. Feng, A. S. *et al.* Ultrasonic communication in frogs. *Nature* **440,** 333–336 (2006).
6. Suthers, R. A. *et al.* Voices of the dead: Complex nonlinear vocal signals from the larynx of an ultrasonic frog. *J. Exp. Biol.* **209,** 4984–4993 (2006).
7. Knudsen, E. I., Blasdel, G. G. & Konishi, M. Sound localization by the barn owl (*Tyto alba*) measured with search coil technique. *J. Comp. Physiol. A* **133,** 1–11 (1979).
8. Bala, A. D. S., Spitzer, M. W. & Takahashi, T. T. Prediction of auditory spatial acuity from neural images on the owl's auditory space map. *Nature* **424,** 771–774 (2003).
9. Heffner, R. S. & Masterton, R. B. in *Comparative Perception* Vol. I, *Basic Mechanisms* (eds Berkley, M. A. & Stebbins, W. C.) 285–314 (John Wiley & Sons, New York, 1990).
10. Cai, H.-X., Che, J., Pang, J.-F., Zhao, E.-M. & Zhang, Y.-P. Paraphyly of Chinese *Amolops* (Anura, Ranidae) and phylogenetic position of the rare Chinese frog, *Amolops tormotus. Zootaxa* **1531,** 49–55 (2007).
11. Narins, P. M. *et al.* Old world frog and bird vocalizations contain prominent ultrasonic harmonics. *J. Acoust. Soc. Am.* **115,** 910–913 (2004).
12. Feng, A. S., Narins, P. M. & Xu, C. H. Vocal acrobatics in a Chinese frog, *Amolops tormotus. Naturwissenschaften* **89,** 352–356 (2002).
13. Narins, P. M. & Smith, S. L. Clinal variation in anuran advertisement calls: Basis for acoustic isolation? *Behav. Ecol. Sociobiol.* **19,** 135–141 (1986).
14. Marquez, R. Female choice in the midwife toads (*Alytes obstetricans* and *A. cisternasii*). *Behaviour* **132,** 151–161 (1995).
15. McClelland, B. E., Wilczynski, W. & Ryan, M. J. Correlations between call characteristics and morphology in male cricket frogs (*Acris crepitans*). *J. Exp. Biol.* **199,** 1907–1919 (1996).
16. Gerhardt, H. C. & Huber, F. *Acoustic Communication in Insects and Anurans* (Univ. Chicago Press, Chicago, 2002).
17. Narins, P. M., Feng, A. S., Fay, R. R. & Popper, A. N. *Hearing and Sound Communication in Amphibians* (Springer, New York, 2007).
18. Duellman, W. E. & Trueb, L. *Biology of Amphibians* (McGraw-Hill, New York, 1986).
19. Roy, D., Borah, B. & Sarma, A. Analysis and significance of female reciprocal call in frogs. *Curr. Sci.* **69,** 265–270 (1995).
20. Kelley, D. B. & Tobias, M. L. in *The Design of Animal Communication* (eds Konishi, M. & Hauser, M.) 9–35 (MIT Press, Cambridge, Massachusetts, 1999).
21. Emerson, S. B. & Boyd, S. K. Mating vocalizations of female frogs: Control and evolutionary mechanisms. *Brain Behav. Evol.* **53,** 187–197 (1999).
22. Kelley, D. B. Vocal communication in frogs. *Curr. Opin. Neurobiol.* **14,** 751–757 (2004).
23. Rand, A. S. Is the canary singing? *Froglog* **48,** 3 (2001).
24. Christensen-Dalsgaard, J. in *Sound Source Localization* (eds Popper, A. N. & Fay, R. R.) 67–123 (Springer, New York, 2005).

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

Acknowledgements This work was supported by the National Natural Sciences Foundation of China (J.-X.S.), the National Institute on Deafness and Other Communication Disorders (A.S.F. and P.M.N.), the UCLA Academic Senate and the Paul S. Veneklasen Research Foundation (P.M.N.), and the National Science Foundation (A.S.F.).

Author Contributions J.-X.S., A.S.F. and P.M.N. were responsible for project planning. All authors (except P.M.N.) conducted the behavioural experiments, J.-X.S. and Z.-M.X. analysed the behavioural data and video recordings, and J.-X.S. and A.S.F. analysed the acoustic data. All authors contributed 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 J.-X.S. (shenjx@sun5.ibp.ac.cn).

METHODS

Animal preparation. Females of *O. tormota* ($n = 4$) were caught on rainy nights along Tau Hua Creek in Huangshan Hot Springs, China, between 22 and 28 April, 2007. In a quiet, darkened room at midnight to early morning hours, female frog vocalizations were recorded using an ultrasonic microphone and PCTape, a computer-based data-acquisition system. Data were saved as WAV files and analysed (fast Fourier transform, 1,024 points), and displayed using SELENA software[5,11,25]. Three to five males were captured each night and kept individually in small plastic cages for acoustic playback experiments.

Acoustic playback experiments. Acoustic playback experiments were carried out between 13 and 29 May 2007 at night along Tau Hua Creek with a peak ambient noise level of 70–88 dB SPL[12] or in a quiet and darkened indoor room, ~1 km from the frog's natural habitat. The WAV file of a recorded female call of *O. tormota* was stored either on an MP3 player for field tests, or on the Play unit, a computer-based playback system[5,25], for indoor study. The MP3 player broadcasts the calls through a power amplifier and a loudspeaker (FE87E). Both playback systems have the means to adjust the level of acoustic playback to ~80–90 dB SPL and the rate of stimulus presentation to one call per 15 s. The playback stimulus was presented over a 5-min period. The loudspeaker was positioned 1 m away from a calling frog in the field, as well as in an indoor arena. Playback experiments in the indoor arena were made under dim infrared illumination.

Audio and video recording, and data analysis. We observed the vocal responses of 48 males indoors and 43 males in the field, as well as their spontaneous calling (no stimulus, as control); for this we used a PCTape system indoors and a digital audio recorder (Sound Devices, model 702; frequency range: 10 Hz–96 kHz) in the field. During the phonotaxis experiments in the indoor arena (3.5 × 4.5 m), each male frog was placed on the floor (that is, the release site) under a removable glass cover (inside diameter 8.5 cm), 1 m in front of the loudspeaker. The peak sound pressure level of the playback stimulus was measured with a ¼-inch wide-band condenser microphone (G.R.A.S. 40BE) placed above the cage or at the release site; the sound was recorded on the digital audio recorder (Sound Devices 702) and this was compared with the output of a calibrator (Bruel and Kjaer 4231) that produces a 1-kHz tone at 94 dB SPL. Audio recordings were made with either a PCTape system[5,6] or the digital audio recorder, with a microphone mounted on a tripod and placed 10 cm from the loudspeaker to monitor the frog's vocalizations as well as the acoustic playback stimulus for later analysis. Video recordings of males' phonotactic behaviours were made under infrared illumination, using an infrared video camera (Sony DCR-TRV30E). The trajectories for each male frog were obtained from the video recordings. We measured the hop distance (D, in cm) and azimuthal jump angle (α, in degrees); the latter was calculated using the formula $\alpha = \arcsin d/D$, where d is the shortest distance from the animal's present position to the straight line between the animal's initial position and the centre of the loudspeaker.

25. Siemers, B. M. & Schnitzler, H.-U. Echolocation signals reflect niche differentiation in five sympatric congeneric bat species. *Nature* **429**, 657–661 (2004).

LETTERS

Perceptual accuracy and conflicting effects of certainty on risk-taking behaviour

Sharoni Shafir[1], Taly Reich[2], Erez Tsur[1], Ido Erev[2] & Arnon Lotem[3]

The 'certainty effect'[1,2] is a notable violation of expected utility theory by decision makers[3–6]. It shows that people's tendency to select the safer of two prospects increases when this prospect provides a good outcome with certainty (for example, people prefer a monetary gain of 3 with certainty over 4 with a probability of 0.8, but do not prefer 3 with a probability of 0.25 over 4 with a probability of 0.2). Subsequent work on experience-based decision making in rats[7] extended the certainty effect to other animals, suggesting its generality across different species and different decision-making mechanisms. However, an attempt to replicate this study with human subjects showed a surprising 'reversed certainty effect'[8,9], namely, the tendency to prefer the safer option decreases when this prospect is associated with certainty (and people now prefer 4 with a probability of 0.8 over 3 with certainty). Here we show that these conflicting results can be explained by perceptual noise and that the certainty effect can be restored experimentally by reducing perceptual accuracy. Using complementary experiments in humans and honeybees (*Apis mellifera*), we show that by manipulating perceptual accuracy in experience-based tasks, both the certainty and the reversed certainty effects can be exhibited by humans and other animals: the certainty effect emerges when it is difficult to discriminate between the different rewards, whereas the reversed certainty effect emerges when discrimination is easy. Our results fit a simple process-based model of matching behaviour[10,11], capable of explaining the certainty effect in humans and other animals that make repeated decisions based on experience. This mechanism should probably be distinguished from those involved in the original certainty effect that was exhibited by human subjects in single description-based problems[1,2].

Experimental studies of choice behaviour in repeated settings reveal a surprising difference between humans and rats. It has been demonstrated[7] that rats deviate from the predictions of expected utility theory[3,4] in the direction of the certainty effect[2]: their tendency to select the safer of two prospects increases when this prospect provides good outcomes with certainty. Human decision makers exhibit the certainty effect when they make a single choice based on a description of the possible outcomes[1,2]. However, in repeated choice tasks when the decision makers have to rely on personal experience (like the rats in ref. 7) human subjects exhibit the opposite tendency, namely, the reversed certainty effect[8,9]. Thus, current data suggest that experience has a different effect on rats than it does on humans and that, paradoxically, experience-based decisions of rats are more similar to human description-based decisions than to experience-based ones.

To resolve this paradox, we hypothesized that these conflicting results represent different experimental conditions. We suspected that rats that received the rewards in the form of cups of water[7] could assess differences in reward less accurately than humans who received the rewards in the form of digital numbers[8,9]. If true, this hypothesis implies that humans, like rats, should exhibit the certainty effect if tested with harder-to-assess rewards. We start with a theoretical illustration showing how a simple experience-based mechanism of choice behaviour can produce the certainty and reversed certainty effects when differences in perceptual accuracy are considered. To test our predictions, we then present complementary experiments in humans and honeybees.

For the theoretical illustration we use the problem studied by ref. 8, which is an experience-based version of the demonstration of the original certainty effect[2]. In this problem, students were asked to choose repeatedly (in 400 trials) between two unmarked buttons of a 'computerized money machine'. They received no previous information concerning the pay-off distribution of each button, but could learn about it from experience (after each choice the immediate pay-off appeared on the screen, given as numerals representing Israeli currency). Each participant was tested in one of two problems. In problem 1, the R (risky) button provided 4 points with a probability of 0.8 (0 otherwise), whereas the S (safe) button provided 3 points with certainty. In problem 2, all previous probabilities were divided by four, giving R providing 4 with probability 0.2 (0 otherwise), and S providing 3 with probability 0.25 (0 otherwise). Note that because each participant was tested in only one problem, the task was a simple repeated binary choice between two alternatives. Previous studies of such tasks in humans and other animals reveal a robust tendency to deviate from maximization towards matching the probability of future choices to that of past success[10–12]. The cognitive mechanism of matching behaviour is not fully understood but it is suggested to be based on acquiring a value representation of each alternative that is dynamically adjusted by experience[10–13]. A decision rule is then applied to determine the action taken in each step based on these representations. A cognitively plausible decision rule for alternatives with variable outcomes has been suggested to be in the form of a competition between samples drawn from the internal representations of each alternative[14–17]. Accordingly, each alternative is represented in memory by a distribution of expected outcomes (rather than by a single aggregated value); in each step, the decision maker picks a sample from the internal representation of each alternative and chooses the action associated with the most favourable sample.

In Tables 1 and 2 we show how applying this mechanism to the above set of problems results in the reversed certainty effect when discrimination is easy (Table 1), and in a behaviour that resembles the original certainty effect when discrimination is difficult (Table 2). Table 1 considers the extreme case of perfect discrimination. Under the specified choice mechanism, the proportion of R choices in problems 1 and 2 should reflect the proportion of cases in which a sample picked from the representation of R is greater than that picked from

[1]B. Triwaks Bee Research Center, Department of Entomology, Faculty of Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, Rehovot 76100, Israel. [2]Max Wertheimer Minerva Center for Cognitive Studies, Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel. [3]Department of Zoology, Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel.

Table 1 | Expected proportion of R choices when discrimination is easy

Problem 1			
		R	
	Pay-offs and probabilities	4 ($P = 0.8$)	0 ($P = 0.2$)
S	3 ($P = 1$)	R is best ($P = 0.8$)	S is best ($P = 0.2$)
	0 ($P = 0$)	–	–
Proportion of R choices: 0.8			
Problem 2			
		R	
	Pay-offs and probabilities	4 ($P = 0.2$)	0 ($P = 0.8$)
S	3 ($P = 0.25$)	R is best ($P = 0.05$)	S is best ($P = 0.2$)
	0 ($P = 0.75$)	R is best ($P = 0.15$)	No difference ($P = 0.6$)
Proportion of R choices: $0.05 + 0.15 + 0.6 \times 0.5 = 0.5$			

Calculation of the expected proportion of R choices in problems 1 and 2, when discrimination is easy and R is chosen when perceived as the best alternative. When there is no perceived difference, R and S are chosen at random (each with a probability of 0.5).

Table 2 | Expected proportion of R choices when discrimination is difficult

Problem 1			
		R	
	Pay-offs and probabilities	4 ($P = 0.8$)	0 ($P = 0.2$)
S	3 ($P = 1$)	No perceived difference ($P = 0.8$)	S is best ($P = 0.2$)
	0 ($P = 0$)	–	–
Proportion of R choices: $0.8 \times (0.5) = 0.4$			
Problem 2			
		R	
	Pay-offs and probabilities	4 ($P = 0.2$)	0 ($P = 0.8$)
S	3 ($P = 0.25$)	No perceived difference ($P = 0.05$)	S is best ($P = 0.2$)
	0 ($P = 0.75$)	R is best ($P = 0.15$)	No difference ($P = 0.6$)
Proportion of R choices: $0.05 \times 0.5 + 0.15 + 0.6 \times 0.5 = 0.475$			

Calculation of the expected proportion of R choices in problems 1 and 2, when discrimination is difficult (3 and 4 are indistinguishable) and R is chosen when perceived as the best alternative. When there is no perceived difference, R and S are chosen at random (each with a probability of 0.5).

the representation of S. If the internal representations reflect past experiences accurately, the probability of such cases can be derived directly from the distribution of joint pay-offs described in Table 1. It is easy to see from Table 1 that, under these conditions, the proportion of R choices should be 0.8 in problem 1 but only 0.5 in problem 2. This implies a strong preference for the risky alternative when the safer alternative provides good outcomes with certainty, namely, a strong reversed certainty effect[8,9]. Note that the strength of this result depends critically on the frequency of cases in which a sample of 4 competes with a sample of 3 (Table 1; top-left cell of the four 'inner' cells). This is most common in problem 1, but extremely rare in problem 2 (0.8 versus 0.05). Furthermore, the assumption of a complete dominance of 4 over 3 is based on the assumption of perfect perception and memory. Clearly, errors that cause 4 to be perceived (or remembered) as similar to 3 should diminish the preference of the risky alternative in problem 1, thereby attenuating the reversed certainty effect. If discrimination becomes so difficult that subjects can tell the difference between zero and 3 (or zero and 4), but not between 3 and 4, the strong reversed certainty effect is replaced by a weak certainty effect (Table 2, 0.4 R choices in problem 1; that is, subjects are expected to shift towards risk aversion). The perceptual constraints assumed in Table 2 may be considered simply as poor perceptual resolution, or also as an extreme case of the Weber–Fechner law[18], according to which diminishing sensitivity to increasing magnitudes can increase perceptual noise. Theoretically, it can also emerge from a sharply concaved subjective value function that increases sharply from zero to 3 but remains virtually the same between 3 and 4. Finally, we remark that qualitatively similar results are expected from realistically refined versions of the above mechanism in which some level of averaging is involved in shaping the distribution of expected outcomes in memory representation: for example, when 4 is remembered as less than its perceived value because it is also associated with experiencing zeros at the same key. Such devaluation of the risky alternative would make it more susceptible to perceptual noise, allowing the certainty effect to emerge even if a reward of 4 was never perceived as equal to 3 but only close to it. Conversely, a tendency to underweight rare events (zero pay-offs) may increase the preference for the risky alternative, producing a reversed certainty effect that may be stronger than expected from Table 1. As long as zero is clearly distinguishable from 3, difficult discrimination between 3 and 4 reduces the attractiveness of the risky option whereas easy discrimination increases it.

If our proposed mechanism is correct, it should be possible to produce the certainty and the reversed certainty effects experimentally, in both humans and other animals, by manipulating the level of perceptual accuracy in the presentation of rewards. To test this prediction with human subjects we carried out a set of experiments in which we replicated the experimental set-up of ref. 8, but replaced the clear feedback of digital numbers with patches of red dots representing the amount of reward (see Methods and Fig. 1). The results were striking. When the same rewards of 3 and 4 monetary units were presented by random patches of 30 and 40 dots, respectively, the reversed certainty effect was replaced by the original certainty effect (compare Fig. 1a with b). This result cannot be explained by the mere use of dots, because in a second experiment, when the same number of dots was presented in a more regular manner (Fig. 1c), the certainty effect diminished considerably. It would seem that the regular presentation of dots (Fig. 1c) makes discrimination easier than in the random patch condition (Fig. 1b), but not to the extent that will produce the reversed certainty effect (as in Fig. 1a). Finally, a third experiment (Fig. 1d) shows that the presentation of rewards in patches of dots can still produce the reversed certainty effect provided that the difference between the two rewards is sufficiently large to facilitate easy discrimination (30 versus 60 dots).

To test our predictions with an animal model, we carried out a set of experiments with honeybees (*Apis mellifera*) that were trained to choose between two syringe tips from which rewards with different sugar concentrations or no reward (empty tip) were provided. The probabilities of receiving a reward (sugar solution) or no reward in the certain and risky alternatives were the same as in problems 1 and 2. To manipulate perceptual accuracy we provided the high and low rewards (analogous to 4 and 3 in the human set-up) as sugar concentrations of 10% and 5% (easy discrimination), 6.7% and 5% (medium discrimination), and identical 6.7% concentrations (impossible discrimination). The results were highly consistent with our predictions and with the human data. We found a strong reversed certainty effect when discrimination was easy (Fig. 2a), a medium reversed certainty effect when discrimination was medium (Fig. 2b), and a weak certainty effect when discrimination was impossible (Fig. 2c). Note that in the case of medium discrimination, the expected values of the two options were virtually the same ($0.8 \times 6.7\% = 5.3\% \approx 5\%$) and yet, a significant reversed certainty

effect was clearly indicated (Fig. 2b). This strong effect cannot be attributed to the ability of honeybees to discriminate between expected values of 5% and 5.3% because honeybees show a weaker effect in a much simpler discrimination test between constant rewards of 5% and 6.7% (reported in the legend to Fig. 2). The risk indifference observed in all discrimination conditions of problem 2 is also inconsistent with expected utility theory (especially in Fig. 2a where discrimination was easy and the expected value of the risky option was clearly higher), but was predicted by Tables 1 and 2. Finally, because the certainty effect exhibited by the honeybees in the present study (Fig. 2c) was expected to be weak (see Table 2), it is not surprising that it failed to reach statistical significance. There is good experimental evidence, however, that honeybees are strongly risk averse when the differences between certain and risky alternatives are more pronounced[19].

In light of our experimental results, previous conflicting evidence[7–9] can be resolved by the simple assertion that humans and other animals deviate from maximization towards a preference for the alternative perceived to lead to better outcomes most of the time (thereby underweighting rare events[9]). This tendency leads to the reversed certainty effect when discrimination is easy, and to behaviour that resembles the original certainty effect when

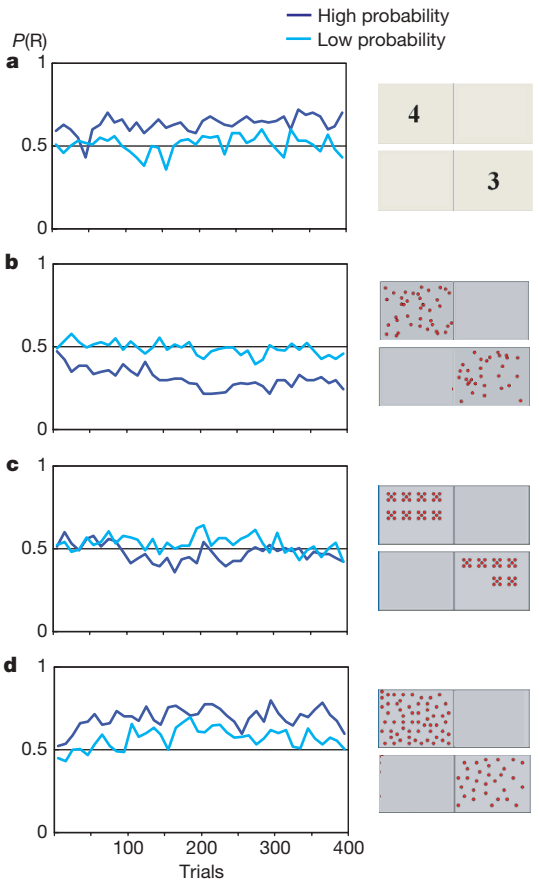


Figure 1 | Experiments with human subjects. Average proportion of human subjects' selection of the risky alternative (R) in 400 trials (presented in 40 blocks of 10 trials) when the safe alternative (S) provides reward with certainty (dark blue line: problem 1, high probability condition), and after both probabilities of receiving reward (for S and R) were divided by four (light blue line: problem 2, low probability condition). A typical reward feedback in each experiment is shown to the right of each panel. **a**, The original reversed certainty effect[8] (problems 1 and 2 as in Table 1) and feedback given as numeral digits representing rewards in 0.01 new Israeli shekels (the difference between problems was significant[8]). **b**, The same set-up and problems as **a** but feedbacks given as random patches of 30 and 40 dots. (The original certainty effect is restored: $P = 0.001$, Mann–Whitney U-test, $z = 3.27$, $n1 = n2 = 30$.) **c**, Same as **b** but the 30 and 40 dots are organized in groups ($P = 0.093$, Mann–Whitney U-test, $z = 1.68$, $n1 = n2 = 20$). **d**, Dots feedback as in **b** but problems 1 and 2 are modified by increasing the reward difference between S and R, represented by 30 and 60 dots, respectively (the difference between problems is significant and in the direction of the reversed certainty effect, $P = 0.007$, Mann–Whitney U-test, $z = 2.69$, $n1 = n2 = 20$). Statistical differences between participants' proportion of R choices in problems 1 and 2 were tested using a two-tailed non-parametric Mann–Whitney U-test (using Statistica version 6.0). Except for **a**, which is based on ref. 8, all data in **b**–**d** are original and are not reported elsewhere.

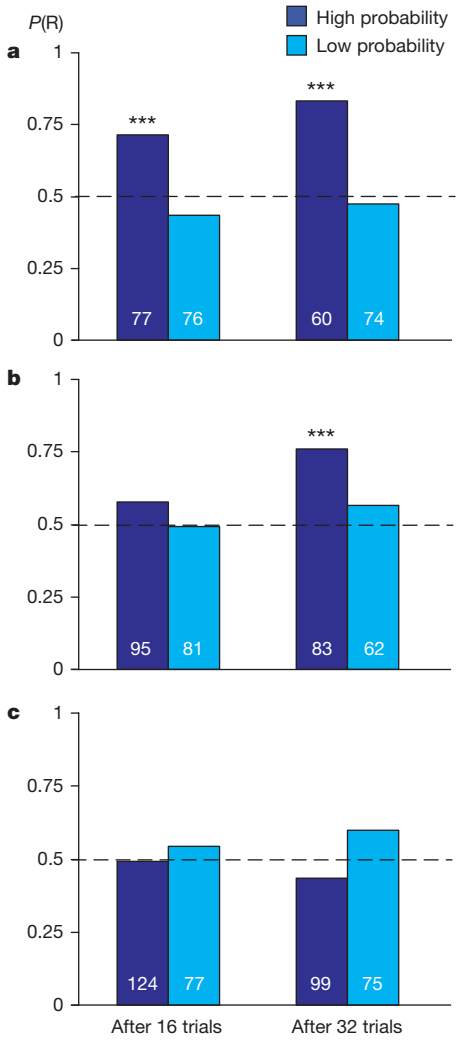


Figure 2 | Experiments with honeybees. The proportion of honeybees that chose the risky alternative (R) after 16 and 32 conditioning trials in the high probability conditions of problem 1 (high probability, dark blue bars: the safe alternative provides reward with certainty and the risky alternative with a probability of 0.8), and in the low probability conditions of problem 2 (low probability, light blue bars: after both probabilities were divided by four), when discrimination between the riskier and safer rewards was easy (**a**; 10% versus 5% sugar concentration), medium (**b**; 6.7% versus 5% sugar concentration) and impossible (**c**; 6.7% versus 6.7% sugar concentration). The dashed line marks risk indifference (that is, $P(R) = 0.5$). Numbers in bars are sample sizes. Three asterisks indicate highly significant deviation from risk indifference (χ^2 test, degrees of freedom $= 1$ (χ_1^2), $P \leq 0.0001$), whereas unmarked bars show no significant deviation from risk indifference. A control experiment (not shown) confirmed that honeybees were able to discriminate well between a 5% sugar solution and a drop of water (choice proportions after 16 and 32 conditioning trials were 0.78 ($N = 36$, $\chi_1^2 = 11.8$, $P = 0.0006$) and 0.77 ($N = 31$, $\chi_1^2 = 9.9$, $P = 0.0017$), respectively), and could also discriminate, but less well, between 6.7% and 5% sugar solutions (choice proportions after 16 and 32 conditioning trials were 0.61 ($N = 82$, $\chi_1^2 = 4.0$, $P = 0.046$) and 0.64 ($N = 55$, $\chi_1^2 = 11.8$, $P = 0.0006$), respectively).

discrimination between the two attractive rewards is difficult. We emphasize that this latter case may only appear to resemble the original certainty effect, because our study suggests that the similarity to the original certainty effect is in fact misleading. The mechanism considered in our study is relevant to experienced-based decisions, and can produce the observed certainty effect only when perceptual noise is involved. This was clearly not the case in the original paradox[1] and its reformulation by ref. 2, where subjects made a single choice based on an accurate description of possible outcomes. Notably, a recent study of the original description-based Allais-type problem showed that the introduction of a certainty reference has a marked effect on decision-makers' brain activation[20]. In other words, the original certainty effect seems to be associated with a strong response to the presentation of the concept of certainty. In experience-based tasks, on the other hand, the concept of certainty is never presented to the participants. The knowledge that one alternative is certain can only be inferred indirectly after many repeated choices (and may be acquired much more easily when the subjects are forced to sample each alternative separately[21]). It is therefore quite unlikely that the same type of brain activation responsible for the original certainty effect is also responsible for the certainty effect found in our study (although it might have only activated after many trials, when some sense of certainty could be acquired, and contributed to the increasing effect of certainty observed in Fig. 1b after about 200 trials). Considering all of the above, both the certainty and the reversed certainty effects studied here seem to represent a robust and independent violation of expected utility theory, which is specific to repeated-choice behaviour. We suggest calling this phenomenon 'perceptual matching' because it is derived from matching choices to the perceived best option in memory representation. Considering the prevalence of repeated choices in real life situations, this phenomenon may be as important as the original (but different) certainty effect that inspired its discovery.

METHODS SUMMARY

Experiments with human subjects were based on paid participants (students) that received 20 new Israeli shekels for showing up and contingent pay-off in 400 repeated choices between alternatives of problems 1 and 2 (that is, 0.04 new Israeli shekels with a probability of 0.8 versus 0.03 new Israeli shekels with certainty, and 0.04 new Israeli shekels with a probability of 0.2 versus 0.03 new Israeli shekels with a probability of 0.25, respectively). To generate perceptual difficulty, the pay-off obtained after each choice was presented on the computer screen in the form of red dots rather than numerical digits (see Fig. 1). No previous information on pay-off structure or length of experiment was provided to participants.

Three experiments with honeybees were conducted to test the effect of perceptual accuracy on a similar set of problems as with human subjects. The pay-off obtained after each choice was a 2-μl drop of sucrose solution of different concentrations. The reward 0 was represented by an empty syringe tip. In the first experiment (Fig. 2a), the pay-off in the risky and safe alternatives was 10% and 5% sugar concentration, respectively. In the second experiment (Fig. 2b), pay-off was 6.7% and 5%, and in the third experiment (Fig. 2c), the pay-off in both alternatives was 6.7%. Owing to differences in sensitivity to sucrose between nectar- and pollen-collecting honeybees[22], only foragers with full pollen baskets were captured, on their return to the hive. Honeybees were tested in a binary-choice proboscis-extension-response conditioning assay[19]. There were 32 conditioning trials with test trials after trials 16 and 32, in which both odours were presented for 1 s twice, in an alternating sequence.

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

Received 4 January; accepted 15 February 2008.

1. Allais, M. Le comportement de l'homme rationel devant le risque, critique des postulates et axiomes de l'ecole americaine. *Econometrica* **21,** 503–546 (1953).
2. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **47,** 263–291 (1979).
3. Von Neumann, J. & Morgenstern, O. *The Theory of Games and Economic Behavior* (John Wiley and Sons, New York, 1944).
4. Savage, L. J. *The Foundations of Statistics* (Wiley, New York, 1954).
5. Tversky, A. & Wakker, P. Risk attitudes and decision weights. *Econometrica* **63,** 1255–1280 (1995).
6. Real, L. A. Paradox, performance, and the architecture of decision-making in animals. *Am. Zool.* **36,** 518–529 (1996).
7. MacDonald, D. H., Kagel, J. H. & Battalio, R. C. Animals' choices over uncertain outcomes: Further experimental results. *Econ. J.* **101,** 1065–1084 (1991).
8. Barron, G. & Erev, I. Small feedback-based decisions and their limited correspondence to description-based decisions. *J. Behav. Decis. Making* **16,** 215–233 (2003).
9. Hertwig, R., Barron, G., Weber, E. U. & Erev, I. Decisions from experience and the weighting of rare events. *Psychol. Sci.* **15,** 534–539 (2004).
10. Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.* **57,** 94–107 (1950).
11. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304,** 1782–1787 (2004).
12. Herrnstein, R. J. On the law of effect. *J. Exp. Anal. Behav.* **13,** 243–266 (1970).
13. March, J. G. Learning to be risk averse. *Psychol. Rev.* **103,** 309–319 (1996).
14. Gibbon, J., Church, R. M., Fairhurst, S. & Kacelnik, A. Scalar expectancy theory and choice between delayed rewards. *Psychol. Rev.* **95,** 102–114 (1988).
15. Osborne, M. J. & Rubinstein, A. Games with procedurally rational players. *Am. Econ. Rev.* **88,** 834–847 (1998).
16. Marsh, B. & Kacelnik, A. Framing effects and risky decisions in starlings. *Proc. Natl Acad. Sci. USA* **99,** 3352–3355 (2002).
17. Erev, I. & Barron, G. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychol. Rev.* **112,** 912–931 (2005).
18. Stevenes, S. S. *Psychophysics* (John Wiley & Sons, New-York, 1975).
19. Shafir, S., Menda, G. & Smith, B. H. Caste-specific differences in risk sensitivity in honeybees, *Apis mellifera. Anim. Behav.* **69,** 859–868 (2005).
20. Dickhaut, J. *et al.* The impact of the certainty context on the process of choice. *Proc. Natl Acad. Sci. USA* **100,** 3536–3541 (2003).
21. Gottlieb, D. A., Weiss, T. & Chapman, B. G. The format in which uncertainty information is presented affects decision biases. *Psychol. Sci.* **18,** 240–246 (2007).
22. Scheiner, R., Page, R. E. & Erber, J. Sucrose responsiveness and behavioral plasticity in honey bees (*Apis mellifera*). *Apidologie (Celle)* **35,** 133–142 (2004).

Acknowledgements We thank M. Tennenholtz and A. Rubinstein for comments. This work was supported by grants from the Israel Science Foundation (ISF) to A.L., I.E. and S.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.L. (lotem@post.tau.ac.il) or I.E. (erev@techunix.technion.ac.il).

METHODS

Experiments with human subjects. Undergraduate students served as paid participants in the experiments (30 were assigned to each problem in the first experiment and 20 to each of the problems in the other experiments). In addition to the performance-contingent pay-off described below, each student received 20 new Israeli shekels for showing up. The total pay-off ranged from 23 to 37 new Israeli shekels (1 new Israeli shekel ≈ 0.22 US dollars).

Three experiments were conducted to test the effect of perceptual accuracy on the set of problems studied by refs 2 and 8 (that is, R versus S: in problem 1, 4(0.8) versus 3(1); in problem 2, 4(0.2) versus 3(0.25)). To generate perceptual difficulty, the pay-off obtained after each choice was presented in the form of red dots rather than numerical digits (see Fig. 1). In the first experiment (Fig. 1b) the numerical feedback of 3 and 4 was replaced by patches of 30 and 40 red dots. The location of each dot in a patch was randomly assigned at each trial. The reward 0 was represented by a single white dot in the centre of the display area. The second experiment (Fig. 1c) resembled the first one but, in order to improve perception, the 30 and 40 dots were organized into regular groups of five dots. In the third experiment, the feedback was the same as in the first experiment (patches of randomly distributed red dots) but problems 1 and 2 were modified to increase the difference between favourable outcomes from 30 and 40 to 30 and 60 dots, respectively (Fig. 1d: problem 1, 60(0.8) versus 30(1); problem 2, 60(0.2) versus 30(0.25)).

The task in all experiments was a binary choice between two button keys on a computer screen, followed by an immediate feedback that appeared for one second on the surface of the selected key. After this second, the selected key returned to its original state and the feedback reappeared at the centre of a display panel located just below the two keys until the next key was selected. The participants repeated this task 400 times, and usually completed the experiment within 10 to 30 min. They were informed that they were playing on a 'computerized dot machine', and that each dot signified a pay-off of 0.001 new Israeli shekels. They received no previous information as to the game's pay-off structure, and to avoid an 'end of task' effect, they were not informed that the experiment would comprise 400 trials. The time interval between repeated choices was determined by the participants, but was too short to enable counting of the dots (and the participants reported that they never tried to do so).

Experiments with honeybees. Three experiments were conducted to test the effect of perceptual accuracy on a similar set of problems as with human subjects. The pay-off obtained after each choice was a 2-μl drop of sucrose solution of different concentrations. The reward 0 was represented by an empty syringe tip. In the first experiment (Fig. 2a), the pay-off in the risky and safe alternatives was 10% and 5% sugar concentration, respectively. In the second experiment (Fig. 2b) pay-off was 6.7% and 5%, and in the third experiment (Fig. 2c), the pay-off in both alternatives was 6.7%.

Owing to differences in sensitivity to sucrose between nectar- and pollen-collecting honeybees[22], only foragers with full pollen baskets were captured, on their return to the hive. Honeybees were tested in a binary-choice proboscis-extension-response conditioning assay[19]. For each subject, an odour (benzyl acetate or geranyl acetate) was presented from either left or right and associated with either R or S pay-offs. The other odour was presented from the other side and associated with the other pay-off, in a balanced design across subjects. During conditioning trials, a harnessed honeybee was presented with an odour for 5 s. Pay-off was delivered 3 s after initiation of odour. There were 32 conditioning trials, with an inter-trial interval of 12 or 15 min (for the low- and high-probability conditions, respectively). Inter-trial interval was smaller in the low-probability condition because the probability of being fed was smaller. A test trial was conducted after trials 16 and 32, in which both odours were presented for 1 s twice, in an alternating sequence. A photograph of the honeybee was then taken. Photographs were scored blindly for the direction in which the honeybee extended its proboscis, which represented its choice[19].

An automated proboscis-extension-response conditioning machine was used, consisting of a carousel loaded with 24 honeybees mounted on a stepper motor. Miniature pumps controlled delivery of odours. Hydraulic arms brought the appropriate syringe tip above the honeybee and lowered it towards the honeybee's antennae and proboscis. Syringe pumps administered sugar solution of the appropriate concentration. Image analysis software monitored the size of the drop on the syringe tip before each delivery.

LETTERS

Towards a transgenic model of Huntington's disease in a non-human primate

Shang-Hsun Yang[1,2,3]*, Pei-Hsun Cheng[1,2]*, Heather Banta[1]*, Karolina Piotrowska-Nitsche[1,2,8], Jin-Jing Yang[1,2], Eric C. H. Cheng[1,2], Brooke Snyder[1,2], Katherine Larkin[1], Jun Liu[1,2,4], Jack Orkin[1], Zhi-Hui Fang[2], Yoland Smith[1,4,5], Jocelyne Bachevalier[1,6,7], Stuart M. Zola[1,4,7,9], Shi-Hua Li[2], Xiao-Jiang Li[2,3,4] & Anthony W. S. Chan[1,2,3,4]

Non-human primates are valuable for modelling human disorders and for developing therapeutic strategies; however, little work has been reported in establishing transgenic non-human primate models of human diseases. Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder characterized by motor impairment, cognitive deterioration and psychiatric disturbances followed by death within 10–15 years of the onset of the symptoms[1–4]. HD is caused by the expansion of cytosine-adenine-guanine (CAG, translated into glutamine) trinucleotide repeats in the first exon of the human huntingtin (*HTT*) gene[5]. Mutant *HTT* with expanded polyglutamine (polyQ) is widely expressed in the brain and peripheral tissues[2,6], but causes selective neurodegeneration that is most prominent in the striatum and cortex of the brain. Although rodent models of HD have been developed, these models do not satisfactorily parallel the brain changes and behavioural features observed in HD patients. Because of the close physiological[7], neurological and genetic similarities[8,9] between humans and higher primates, monkeys can serve as very useful models for understanding human physiology and diseases[10,11]. Here we report our progress in developing a transgenic model of HD in a rhesus macaque that expresses polyglutamine-expanded *HTT*. Hallmark features of HD, including nuclear inclusions and neuropil aggregates, were observed in the brains of the HD transgenic monkeys. Additionally, the transgenic monkeys showed important clinical features of HD, including dystonia and chorea. A transgenic HD monkey model may open the way to understanding the underlying biology of HD better, and to the development of potential therapies. Moreover, our data suggest that it will be feasible to generate valuable non-human primate models of HD and possibly other human genetic diseases.

We injected 130 mature rhesus oocytes with high titre lentiviruses expressing exon 1 of the human *HTT* gene with 84 CAG repeats (*HTT*-84Q; Fig. 1c) and lentiviruses expressing the green fluorescent protein (*GFP*) gene (Fig. 1c), under the control of the human polyubiquitin-C promoter, into the perivitelline space. After fertilization by intracytoplasmic sperm injection, 82% (89 out of 108) of the zygotes developed to the 4–8-cell stage and 30 embryos were transferred to eight surrogates. Of those surrogates, six pregnancies (75%; 6 out of 8) were established and five live newborns (22%, 5 out of 23; rHD-1, rHD-2, rHD-3, rHD-4 and rHD-5) were delivered at full term (Table 1). Among these five newborns, two sets of twins (rHD-1 and rHD-2; rHD-4 and rHD-5) were delivered by caesarean section, and a singleton (rHD-3) was delivered naturally.

All of the transgenic monkeys carried the transgenic mutant *HTT* and *GFP* genes. Fig. 1 shows two monkeys (rHD-1 and rHD-2, fraternal twins; Fig. 1a) expressing GFP at 4 weeks of age (Fig. 1b). Transgene integration was confirmed by polymerase chain reaction (PCR) (Fig. 1d, e) and Southern blot analysis (Supplementary Fig. 1). We estimated that rHD-1 and rHD-2 carried a mutant *HTT* gene in samples of umbilical cord, whereas placental tissue of extraembryonic origin from rHD-1 had two mutant *HTT* genes of different size (Fig. 1d). Two or more integration sites were identified in rHD-3, rHD-4 and rHD-5 (Table 1 and Supplementary Fig. 1), suggesting that these monkeys may have expressed the transgenic mutant *HTT* at a higher level and thus showed more severe phenotypes. Owing to the limited access of tissues from the live HD infants, detailed post-mortem analyses of transgene expression were focused on the HD monkeys rHD-4 and rHD-5.

Variations in the repeat length were found in the transgenic monkeys (Table 1). Specifically, rHD-1 carries 29 CAGs whereas the other four monkeys carried CAG repeats ranging from 27 to 88. Among these five newborns, one (rHD-3) survived for 1 month and two (rHD-4 and rHD-5) survived for less than a day. The latter two newborns had respiratory difficulties and showed signs of motor impairment soon after birth. Because of their early death we were unable to perform thorough testing of rHD-4 and rHD-5, and we could not establish whether their symptoms were caused by the mutant *HTT* gene. rHD-3 showed dystonia and chorea 2 days after birth; rHD-2 displayed mild involuntary motor movement that appeared sporadically at 1 week of age. As expected, rHD-1—which has CAGs in the normal range for non-HD humans—seems to be a normal, healthy monkey.

The expression of transgenic mutant *HTT* was confirmed by western blot and immunohistochemistry with mEM48, a monoclonal antibody whose reaction with human HTT is enhanced by polyQ expansion[12]. Western blotting of the placental tissues of all infants demonstrated the presence of oligomeric HTT at high molecular mass (>250 kDa) in the upper portion of a gradient polyacrylamide gel (Fig. 1f). The intensity of mEM48 immunostaining is determined by the polyQ length and expression levels of transgenic *HTT*. Given that rHD-1 and rHD-2 carry a single copy of transgenic *HTT*, the more intense mEM48 staining in rHD-2 placental tissue than in rHD-1 tissues supports the CAG repeat analysis showing that rHD-2 carries a larger repeat (83) than rHD-1 (29) (Fig. 1d and Table 1). Consistently, placental tissues from rHD-3, rHD-4 and rHD-5, which have more copy numbers of transgenic *HTT* with larger repeats (69–88), show much more intense mEM48 immunostaining.

[1]Yerkes National Primate Research Center, [2]Department of Human Genetics, [3]Genetics and Molecular Biology Program, [4]Neuroscience Program, [5]Department of Neurology, [6]Department of Psychology, [7]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia 30329, USA. [8]Department of Experimental Embryology, Institute of Genetics and Animal Breeding, Polish Academy of Sciences, 05-552 Wolka-Kosowska, Poland. [9]Veterans Affairs Medical Center, Atlanta, Georgia 30033, USA.
*These authors contributed equally to this work.

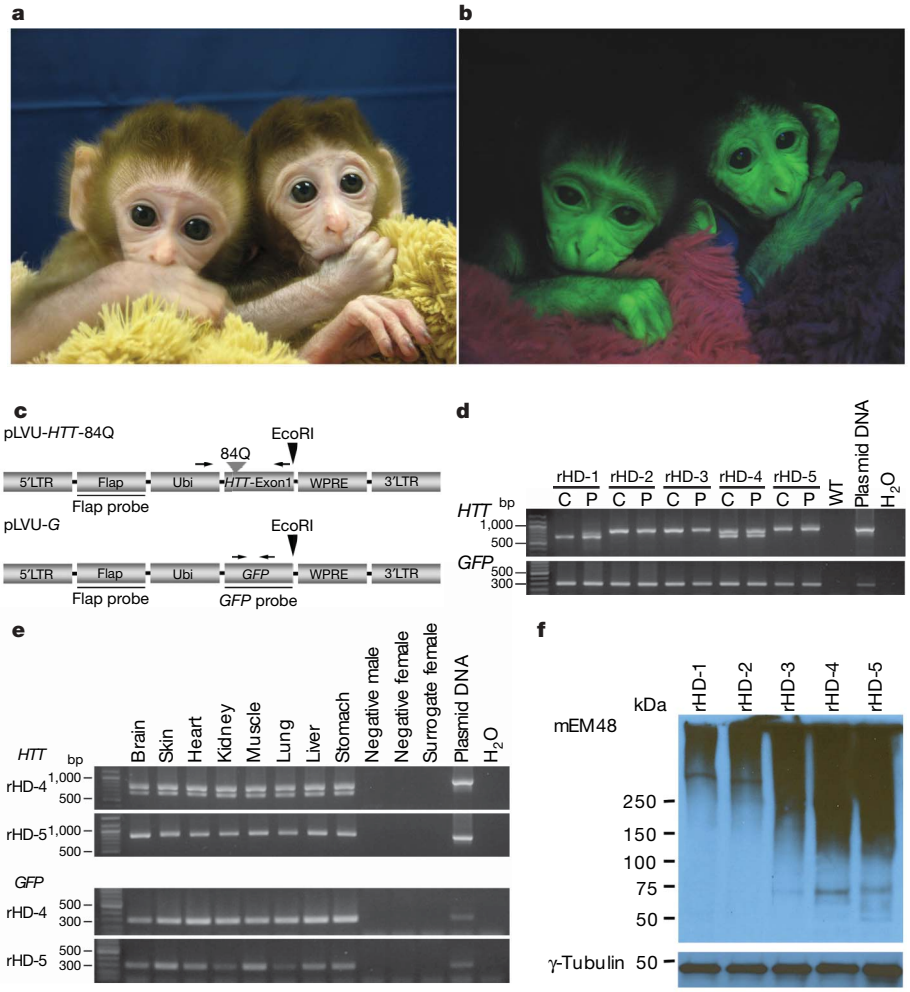


Figure 1 | Generation of a transgenic model in monkeys of HD. **a**, **b**, The transgenic HD monkeys rHD-1 (left) and rHD-2 (right) are shown. Transmission light image (**a**) and fluorescent image (**b**) showing GFP expression in HD monkeys. **c**, Top panel: lentiviral vector that carries exon 1 of the *HTT* gene with 84 CAG repeats (pLVU-*HTT*-84Q). Bottom panel: lentiviral vector that carries the *GFP* gene (pLVU-G). Arrows indicate the positions of PCR primers; arrowheads denote restriction digest sites. Flap, HIV-flap sequence; *GFP*, green-fluorescent-protein gene; *HTT*, huntingtin gene; LTR, long terminal repeat; Ubi, ubiquitin promoter; WPRE, woodchuck post-transcriptional regulatory element. **d**, **e**, The presence of transgenes in HD monkeys was confirmed by PCR analysis using primer sets specifically for mutant *HTT* (top panels) and for the *GFP* gene (bottom panels). PCR of the cord (C) and placental (P) tissues of all HD monkeys (**d**), and PCR of different tissues collected from rHD-4 and rHD-5 (**e**). **f**, Expression of the transgenic mutant *HTT* was confirmed by western blot analysis using the placental tissues. Immunostaining was performed using mouse-monoclonal-mEM48 antibody (top panel) and an antibody against γ-tubulin (bottom panel).

Thus, the levels of mutant HTT in the tissues seem to be associated with the early death of monkeys rHD-3, rHD-4 and rHD-5, and the severity of phenotypes seen in infants rHD-2 and rHD-3, whereas a short CAG repeat (29) in rHD-1 did not result in detectable phenotypes (Table 1).

Expression of mutant HTT was also determined by mEM48 western blotting of post-mortem peripheral (Fig. 2a, b) and brain tissues (Fig. 2c, d) of rHD-4 and rHD-5. As seen in western blots of the placental tissues (Fig. 1f), oligomeric mutant HTT is present in the peripheral tissues (Fig. 2a, b) and brain (Fig. 2c, d) of transgenic

Table 1 | Summary of transgenic HD monkeys

rHD	Sex	CAG repeats	*HTT* copy number	Expression of mutant HTT/GFP	Body weight ratio	Movement dysfunction				
						Onset/ frequency	Dystonia	Chorea	Swallowing difficulty	Respiratory difficulty
1	M	29	1	+/+	2.85*	–	–	–	–	–
2	M	83	1	+/+	2.16*	1 week/ sporadic	+	+	–	+
3	F	84	2	+/+	1.11†	2–3 days/ persistent	++++	++++	++++	++
4	F	27, 65	2	+/+	ND	At birth/ persistent	++	++	ND	++++
5	F	88	4	+/+	ND	At birth/ ND	ND	ND	ND	++++

* The ratio of body weight at 24 weeks after birth to birth weight.
† The ratio of body weight at 28 days after birth to birth weight.
F, female; M, male; ND, not determined.

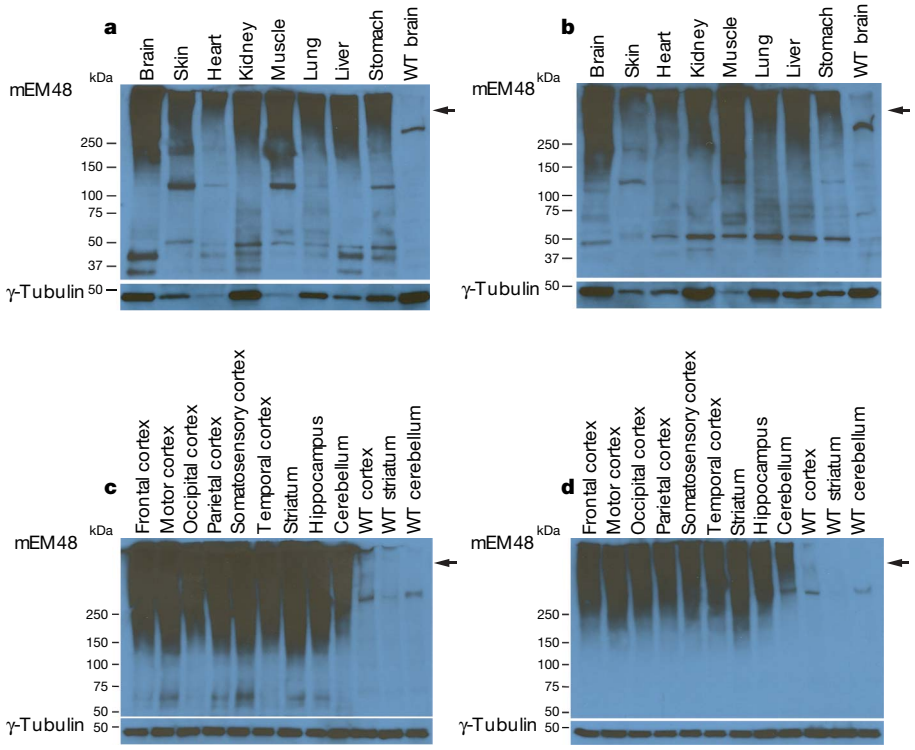


Figure 2 | Expression of mutant HTT in HD monkey peripheral tissues and brains. **a**–**d**, mEM48 immunoblot of peripheral and brain tissues reveals high-molecular-mass oligomeric HTT (arrow) and soluble HTT products. The blot was also probed with an antibody to γ-tubulin as an internal control. Shown are immunoblots of peripheral tissues collected from monkey rHD-4 (**a**) and rHD-5 (**b**), and samples from different brain regions of monkey rHD-4 (**c**) and rHD-5 (**d**), with antibody to mEM48 (top panel) and γ-tubulin (bottom panel). WT, wild-type non-transgenic monkey.

monkeys but not in non-transgenic or wild-type monkeys, indicating that the oligomeric HTT specifically represents transgenic mutant HTT. Variations in the extent of aggregation were observed in peripheral tissues (Fig. 2a, b), suggesting that aggregation of mutant HTT or its expression levels vary in different tissues. In contrast to peripheral tissues, different brain regions show equivalent levels or extent of transgenic mutant HTT (Fig. 2c, d).

Consistent with the western blotting analysis, immunohistochemistry with mEM48 and 1C2—antibodies specific to expanded polyQ domain—showed HTT aggregates or inclusions in the striatum and cortex of rHD-4 and rHD-5 (Fig. 3a, b). Transgenic HTT is widespread in the various brain regions but with no significant difference between regions. Expression of transgenic mutant HTT is extensive in the brain (Fig. 3a, b; upper panel; ×100). At higher magnification (Fig. 3a, b; lower panel; ×630), mutant HTT accumulates in the nuclei and also forms abundant small aggregates that do not seem to be in the cell body. Both antibodies illustrated neuropil aggregates[13] located in an array manner, suggesting that they were originally localized in neuronal processes such as axons and dendrites[14–19]. The size of these neuropil aggregates can be larger than that of axons or dendrites (arrowhead in Fig. 3a, b); this is consistent with neuropil aggregates found in the brains of HD patients[13,20] and in HD mouse models[15–18] in which neuropil HTT aggregates are associated with axonal degeneration. However, Nissl staining did not reveal any obvious degeneration in the striatum of rHD-5 (data not shown).

Variable extents of motor dysfunction were observed in four (rHD-2, rHD-3, rHD-4 and rHD-5) of the five HD monkeys, although we were unable to examine rHD-4 and rHD-5 in detail before their death. rHD-2 and rHD-3 showed different degrees of difficulty in movement coordination and involuntary movement such as chorea and dystonia, whereas rHD-1 has no signs of clinical symptoms (Table 1). On the basis of our preliminary observations, the severity, frequency and onset of the involuntary movements seemed to depend on the length of the CAG repeats and the number of integration sites (Table 1). Monkey rHD-3 survived for 1 month and showed severe chorea, dystonia and difficulty in swallowing. At the time of writing, the monkeys rHD-1 and rHD-2 are 6 months old. Behavioural and cognitive testing and in-depth characterizations are underway; thorough testing, including brain imaging, will be performed periodically throughout development until adulthood.

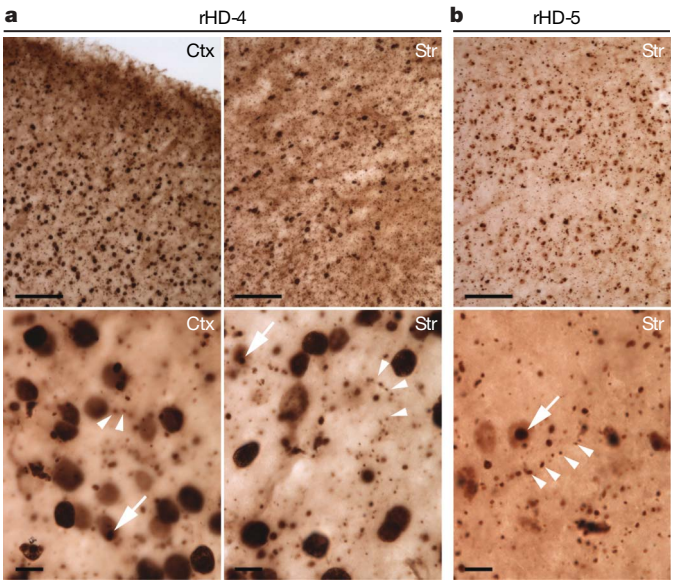


Figure 3 | Histopathology of HD monkey brain. **a**, **b**, Brain sections of rHD-4 (**a**) and rHD-5 (**b**) were immunostained with mEM48 and 1C2, respectively. Low magnification (upper panels; scale bars, 100 μm) shows the abundant distribution of transgenic mutant HTT in the cortex (Ctx) and striatum (Str). High magnification (bottom panels; scale bars, 10 μm) demonstrates that transgenic mutant HTT is distributed in neuronal nuclei and forms neuropil aggregates (arrowheads). Nuclear inclusions (arrows) are evident in sections stained with mEM48 or 1C2.

As there is no standard rating scale for HD in monkeys—especially for infants—our evaluation is based on the unified Huntington's disease rating scale, which has been modified for the use in infant monkeys. On the basis of our preliminary assessment using the Huntington's disease primate model rating scale, in which 80 is the highest score reflecting the most severe symptom, rHD-1 has the score of '0' at 2 and 6 months, whereas rHD-2 has the scores 12 and 26, respectively. In addition, rHD-3 had the score of 60 at 1 month of age.

The early death of transgenic HD monkeys carrying a higher copy number of transgenes expressing a small amino-terminal HTT fragment supports the idea that N-terminal mutant HTT fragments are pathogenic. Although detailed and periodically repeated behavioural characterization of the surviving transgenic HD monkeys remains to be done, the infant HD monkeys already showed some features similar to those found in HD patients, including chorea and dystonia. The transgenic HD monkeys provide the opportunity for a wide range of behavioural and cognitive assessments that are identical to or close to the assessments used for human patients. Furthermore, the two male HD monkeys could serve as founders, and their semen can be collected throughout life non-invasively and cryopreserved. Thus, a cohort of HD monkeys could then be established using these semen and reproductive techniques in a reasonable time frame. Our progress in establishing a HD transgenic monkey also bodes well for the generation of non-human primate models of other important neurodegenerative diseases. Establishing such models is invaluable for understanding disease pathogenesis and for the development of early diagnostic and treatment strategies.

METHODS SUMMARY

Lentiviruses carrying exon 1 of the human *HTT* gene with 84 CAGs (*HTT*-84Q) and the *GFP* gene under the control of human polyubiquitin-C promoter were microinjected into the perivitelline space of monkey metaphase-II-arrested oocytes followed by *in vitro* fertilization, *in vitro* culture and embryo transfer into surrogate females. Transgenic status of HD monkeys was confirmed by PCR and Southern blot analyses on available tissues. The expression of the mutant *HTT* gene was confirmed by immunohistochemistry and western blot. The expression of *GFP* gene was visualized under a Sky-blue II epifluorescent light. Motor impairment was evaluated by the Huntington's diseases primate model rating scale.

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

Received 7 January; accepted 9 April 2008.
Published online 18 May 2008.

1. Cummings, C. J. & Zoghbi, H. Y. Trinucleotide repeats: mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.* **1,** 281–328 (2000).
2. Davies, S. & Ramsden, D. B. Huntington's disease. *Mol. Pathol.* **54,** 409–413 (2001).
3. Myers, R. H., Marans, K. S. & MacDonald, M. E. *Huntington's Disease* in *Genetic Instabilities and Hereditary Neurological Diseases* (eds Wells, R. D. & Warren, S. T.) 301–324 (Academic, San Diego, 1998).
4. Rubinsztein, D. C. Lessons from animal models of Hungtington's disease. *Trends Genet.* **18,** 202–209 (2002).
5. MacDonald, M. E. *et al.* HD research collaborative groups. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntintgton's disease chromosome. *Cell* **72,** 971–983 (1993).
6. Sharp, A. H. *et al.* Widespread expression of Huntington's disease gene (IT15) protein product. *Neuron* **14,** 1065–1074 (1995).
7. Lane, M. A. Nonhuman primate models in biogerontology. *Exp. Gerontol.* **35,** 533–541 (2000).
8. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
9. McConkey, E. H. & Varki, A. A primate genome project deserves high priority. *Science* **289,** 1295–1296 (2000).
10. Chan, A. W. S., Chong, K. Y., Martinovich, C., Simerly, C. & Schatten, G. Transgenic monkeys produced by retroviral gene transfer into mature oocytes. *Science* **291,** 309–312 (2001).
11. Chan, A. W. S., Chong, K. Y. & Schatten, G. in *Transgenic Animal Technology: A Laboratory Handbook* (ed. Pinkert, C. A.) (Academic, San Diego, 2002).
12. Zhou, H. *et al.* Huntingtin forms toxic NH_2-terminal fragment complexes that are promoted by the age-dependent decrease in proteasome activity. *J. Cell Biol.* **163,** 109–118 (2003).
13. Gutekunst, C. A. *et al.* Nuclear and neuropil aggregates in Huntington's disease: relationship to neuropathology. *J. Neurosci.* **19,** 2522–2534 (1999).
14. Bates, G. P., Mangiarini, L., Wanker, E. E. & Davies, S. W. Polyglutamine expansion and Huntington's disease. *Biochem. Soc. Trans.* **26,** 471–475 (1998).
15. Lin, C. H. *et al.* Neurological abnormalities in a knock-in mouse model of Huntington's disease. *Hum. Mol. Genet.* **10,** 137–144 (2001).
16. Li, H. *et al.* Ultrastructural localization and progressive formation of neuropil aggregates in Huntington's disease transgenic mice. *Hum. Mol. Genet.* **8,** 1227–1236 (1999).
17. Li, H., Li, S. H., Yu, Z. X., Shelbourne, P. & Li, X. J. Huntingtin aggregate-associated axonal degeneration is an early pathological event in Huntington's disease mice. *J. Neurosci.* **21,** 8473–8481 (2001).
18. Yu, Z. X. *et al.* Mutant Huntingtin causes context-dependent neurodegeneration in mice with Huntington's disease. *J. Neurosci.* **23,** 2193–2202 (2003).
19. Davies, S. W. *et al.* Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell* **90,** 537–548 (1997).
20. DiFiglia, M. *et al.* Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neuritis in brain. *Science* **277,** 1990–1993 (1997).

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

Acknowledgements We thank J. Fanton (deceased), M. Zelinski-Wooten, M. Sparman and D. Wolf for consultation, C. Lois for lentivirus backbone and C. Testa for critical review of the manuscript. We also thank F. Zhang, T. Caspary, S. Warren, J. Greene, T. Wichmann, M. Wilson, S. Chikazawa, K. Gould, L. Walker, K. Layug, E. Strobert, J. Else, J. Ksiazek, K. Strait, F. Stroud, J. Jenkins, J. Cohen, J. Pare, S. Jenkins, K. Paul, S. Lackey, J. Johnson-Ward, the veterinary staff, the animal resource and the endocrine core laboratory at the Yerkes National Primate Research Center. Acryline was provided by NICHD/NIH. All transgenic HD monkeys were housed under the guideline of the IACUC approved procedures and the support of YNPRC Division of Animal Resources. All newborn monkeys were closely monitored by the veterinary staff and infant care personnel. All procedures were approved by YNPRC/Emory Animal Care and Biosafety Committees. The YNPRC is supported by NIH/NCRR. A.W.S.C., S.H.L. and X.J.-L are supported by grants awarded by the NIH.

Author Contributions S.-H.Y. carried out assisted reproductive technique (ART) in monkeys, viral gene transfer, construct design and molecular analysis; P.-H.C., construct design and evaluation; K.P.-N., ART in monkeys; H.B., animal management; behavioural testing and all animal procedures; K.L., animal care and behavioural testing; E.C.H.C., molecular analysis; J.-J.Y., preparation of high titre lentiviruses; B.S., J.L. and Z.H.F., neuropathological analysis; J.O., surgical procedures and animal care; Y.S., neuropathological analysis; J.B., design of behavioural and cognitive testing; S.M.Z., experimental design and manuscript preparation; S.H.L. and X.J.-L., construct design, analysis and manuscript preparation; A.W.S.C., ART in monkey, viral gene transfer, experimental design, construct design, molecular analysis and manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.W.S.C. (achan@genetics.emory.edu).

METHODS

Construction and preparation of lentiviruses carrying the *HTT*-84Q and *GFP* genes[21]**.** Exon 1 of the human *HTT* gene with 84 CAGs (*HTT*-84Q) was inserted into a lentiviral vector, which was regulated by the human polyubiquitin-C promoter (pLVU-*HTT*-84Q). For GFP lentiviruses, the *HTT*-84Q gene in the vector was replaced by the *GFP* gene (pLVU-*G*). High titre lentiviruses were generated by co-transfection of lentiviral vectors coding for *HTT*-84Q or *GFP*, pΔ8.9, and pVSV-G (Invitrogen) into a 293FT packaging cell (Invitrogen). Viruses were then concentrated by ultracentrifugation using the method described previously[21].

Production of rhesus monkey oocytes, embryos and babies. Assisted reproductive techniques in rhesus monkey have been described elsewhere[11]. In brief, adult females were hormone stimulated and their oocytes were recovered for *in vitro* fertilization and culture. Embryos at the 4–8-cell stage were selected for embryo transfer based on morphological appearance. Surrogate females at synchronized reproductive cycles were identified based on their hormone profile.

Gene delivery in mature rhesus monkey oocytes. Metaphase-II-arrested oocytes were selected for perivitelline space injection followed by intracytoplasmic sperm injection. Lentiviral solution was loaded into the injection needle by micropipette and the viral solution was injected into the perivitelline space. After virus injection, the oocytes were fertilized by intracytoplasmic sperm injection followed by *in vitro* culture until embryo transfer at the 4–8-cells stage[10,11].

Monitoring GFP expression in infants. Live HD monkey infants were placed under a Sky-blue II epifluorescent light (475 nm; Youlum), and images were captured by a digital camera equipped with an emission filter at an emission wavelength of 520 nm.

PCR analysis. To detect the *HTT*-84Q gene, ubiquitin forward primer (5′-GAGGCGTCAGTTTCTTTGGTC-3′) and *HTT*-84Q-R reverse primer (5′-GCTGGGTCACTCTGTCTCTG-3′) was used to yield an 818-base-pair product after amplification of genomic DNA from the HD monkey tissues. However, variations in size resulted because of the number of CAG repeats. Genomic DNA (100 ng) from different tissues was subjected to PCR for 35 cycles at 96 °C for 5 min, 96 °C for 45 s, 62 °C for 45 s, 72 °C for 150 s, and then 72 °C for 7 min. To determine the number of CAG repeats in HD monkeys, the PCR products were sequenced using HD exon 1 forward primer (5'-GGCGACCCTGGAAAAGC-TGA-3′). To detect the *GFP* gene, *GFP*-F forward primer (5'- TTCAAGGA-CGACGGCAACTAC-3′) and *GFP*-R reverse primer (5'-TAGTGGTTGTC-GGGCAGCAG-3′) were used for amplification for 35 cycles at 94 °C for 5 min, 94 °C for 30 s, 64 °C for 30 s, 72 °C for 20 s, and then 72 °C for 5 min, yielding a product of 302 bp. DNA from wild-type monkeys was used as the negative control, and plasmid *HTT*-84Q and *GFP* were used as the positive controls.

Southern blot analysis. Eight micrograms of genomic DNA were digested overnight using EcoRI, which only cut once within the transgene. The digested genomic DNA was then separated by gel electrophoresis on a 0.8% agarose gel and transferred to Hybond-N^+ nylon membranes (Amersham). To determine the number of integration events of the two viruses (*HTT*-84Q or *GFP*) in HD monkeys, a subtraction approach was used because both constructs have identical lentiviral backbones and the short fragment of *HTT*-84Q is not sufficient to distinguish between the mutant *HTT* and endogenous *HTT*. We hybridized two identical membranes with a [^{32}P]-labelled probe that specifically binds to the *GFP* gene and to the flap sequence of the lentiviral vector. The number of integration events of the *GFP* gene and the flap sequence was calculated by subtracting the number of *GFP* integration events from the total number of integration events determined by the flap probe. The number of transgenic mutant *HTT* gene integration events could then be estimated. Integration sites were determined by exposing [^{32}P]-hybridized members to the phosphor screen and scanning using the Typhoon phosphorimager (GE). Plasmid DNA (*HTT*-84Q and *GFP*) digested with EcoRI was used as a positive control; genomic DNA from a non-transgenic monkey was used as a negative control.

Western blot analysis. Total proteins were extracted from different tissues and their concentration was determined using the Bradford assay (Pierce). Equal amounts (20–30 μg) of protein extract with loading dye were boiled before loading into 4–15% gradient polyacrylamide gels (Bio-Rad). After electrophoresis, proteins were transferred onto a PVDF membrane (Bio-Rad) using Bio-Rad's transblot followed by blocking in 5% skimmed milk for 2 h. The membrane was incubated with the primary antibodies mouse monoclonal mEM48[19] (1:50 dilution) and γ-tubulin (Sigma; 1:2,000 dilution), followed by secondary peroxidase-conjugated antibodies (Jackson ImmunoResearch laboratories) for detecting proteins with an Amersham ECL kit (PerkinElmer).

Immunohistochemistry of transgenic monkey tissues. Post-mortem brain tissues of transgenic monkeys were fixed in 4% paraformaldehyde overnight, transferred to 30% sucrose, stored at 4 °C, embedded and then cut into 50 μm sections. For mEM48 staining, the sections were incubated with 0.3% hydrogen peroxide for 15 min, blocked for 1 h at room temperature, and incubated with the primary monoclonal antibody mEM48 (1:50) at 4°C overnight. For 1C2 (1:4000 dilution) staining, the tissues were treated with formic acids (88%) for 10 min.

21. Yang, S. H. *et al.* Enhanced transgenesis by intracytoplasmic injection of envelope-free lentivirus. *Genesis* **45**, 177–183 (2007).

LETTERS

Substrate-targeting γ-secretase modulators

Thomas L. Kukar[1], Thomas B. Ladd[1], Maralyssa A. Bann[1], Patrick C. Fraering[2,3], Rajeshwar Narlawar[4], Ghulam M. Maharvi[1], Brent Healy[1], Robert Chapman[1], Alfred T. Welzel[5], Robert W. Price[1], Brenda Moore[1], Vijayaraghavan Rangachari[1], Bernadette Cusack[1], Jason Eriksen[1], Karen Jansen-West[1], Christophe Verbeeck[1], Debra Yager[1], Christopher Eckman[1], Wenjuan Ye[3], Sarah Sagi[6], Barbara A. Cottrell[6], Justin Torpey[6], Terrone L. Rosenberry[1], Abdul Fauq[1], Michael S. Wolfe[3], Boris Schmidt[4], Dominic M. Walsh[5], Edward H. Koo[6] & Todd E. Golde[1]

Selective lowering of Aβ42 levels (the 42-residue isoform of the amyloid-β peptide) with small-molecule γ-secretase modulators (GSMs), such as some non-steroidal anti-inflammatory drugs, is a promising therapeutic approach for Alzheimer's disease[1]. To identify the target of these agents we developed biotinylated photoactivatable GSMs. GSM photoprobes did not label the core proteins of the γ-secretase complex, but instead labelled the β-amyloid precursor protein (APP), APP carboxy-terminal fragments and amyloid-β peptide in human neuroglioma H4 cells. Substrate labelling was competed by other GSMs, and labelling of an APP γ-secretase substrate was more efficient than a Notch substrate. GSM interaction was localized to residues 28–36 of amyloid-β, a region critical for aggregation. We also demonstrate that compounds known to interact with this region of amyloid-β act as GSMs, and some GSMs alter the production of cell-derived amyloid-β oligomers. Furthermore, mutation of the GSM binding site in the APP alters the sensitivity of the substrate to GSMs. These findings indicate that substrate targeting by GSMs mechanistically links two therapeutic actions: alteration in Aβ42 production and inhibition of amyloid-β aggregation, which may synergistically reduce amyloid-β deposition in Alzheimer's disease. These data also demonstrate the existence and feasibility of 'substrate targeting' by small-molecule effectors of proteolytic enzymes, which if generally applicable may significantly broaden the current notion of 'druggable' targets[2].

We previously demonstrated that a subset of non-steroidal anti-inflammatory drugs (NSAIDs; that is, ibuprofen and sulindac) are examples of GSMs—agents capable of preferentially lowering Aβ42 *in vitro* and *in vivo*[3,4]. A signature of Aβ42-lowering GSMs is a reciprocal increase in shorter amyloid-β peptides. Further studies have identified tarenflurbil (formerly named *R*-flurbiprofen[5]) as a non-NSAID Aβ42-lowering GSM with potential therapeutic application in Alzheimer's disease[3,6]. In contrast to γ-secretase inhibitors, most GSMs do not inhibit γ-secretase-mediated release of the intracellular cytoplasmic domains of APP or other γ-secretase substrates such as Notch and ErbB4[3,4]. We have also identified other GSMs that selectively increase Aβ42 quantities, similar to mutations in presenilin 1 (PSEN1) and APP which cause familial Alzheimer's disease[7,8]. Although studies have suggested that GSMs target γ-secretase, the mechanism of amyloid-β modulation by GSMs has not been definitively established[3,9,10].

We synthesized two GSM derivatives for photoaffinity labelling studies to determine the molecular targets of the GSMs: fenofibrate-biotin (Fen-B), a derivative of fenofibrate which is an Aβ42-raising GSM, and flurbiprofen-benzophenone-biotin (Flurbi-BpB), a derivative of tarenflurbil which is an Aβ42-lowering GSM (Fig. 1a). Each compound contains benzophenone (a preferred photoactive moiety for protein labelling[11]) and a biotin tag (for detection and affinity purification), and maintains the activity of the parent GSM (Supplementary Figs 2 and 3).

Initial crosslinking of Fen-B (1–100 μM) in lysates from human neuroglioma H4 cells overexpressing APP demonstrated that numerous proteins were labelled but that PSEN1 was not (data not shown). To determine whether this negative result was due to limited sensitivity, we investigated the ability of Fen-B to label a highly purified preparation of active γ-secretase[12]. Photolysis of purified γ-secretase (Fig. 1b) in the presence of Fen-B (300 μM) followed by precipitation with streptavidin did not label the core components of γ-secretase (PSEN1, nicastrin (NCSTN), anterior pharynx-defective 1 (APH1) or presenilin enhancer 2 (PSENEN, also known as PEN2)). Similar studies with biotinylated carprofen[13], an Aβ42-lowering GSM, did not detect labelling of γ-secretase (B.S., unpublished data). Having failed to label the subunits of γ-secretase, we turned our attention to the remaining member of an active γ-secretase 'complex': the substrate itself. We addressed whether a purified, recombinant Flag-tagged γ-secretase substrate derived from APP (APP(C100)–Flag, comprising the 99 C-terminal residues of APP plus a methionine at the N terminus, ref. 12) could be labelled by Fen-B and Flurbi-BpB. Photoactivated crosslinking with concentrations of GSM photoprobes spanning the range over which they modulate Aβ42 led to increasing quantities of biotinylated APP(C100)–Flag (Fig. 1c). Labelling of APP(C100)–Flag by Fen-B was competed by Aβ42-raising and -lowering GSMs (Fig. 1d); however, non-GSM NSAIDs (aspirin, naproxen) did not compete (data not shown). The sulphone derivative of the GSM sulindac, which does not affect Aβ42 levels, did not compete for labelling, showing that small structural features influence the ability to modulate amyloid-β and compete for labelling of APP C-terminal fragments (CTFs; Fig. 1d). These results indicate that binding of GSMs to the APP substrate could mediate their ability to shift the location of γ-secretase cleavage.

Next, we addressed whether the interaction between APP CTF and GSMs could be detected in cells. Direct exposure of H4 cells to ultraviolet and photoprobes was toxic (Supplementary Fig. 3); therefore, crosslinking of Fen-B and Flurbi-BpB was examined in solubilized membrane fractions from human neuroglioma H4 cells enriched for active γ-secretase and substrates (APP CTFs 83 and 99)[7]. Fen-B and Flurbi-BpB labelled an APP-derived protein of ~11 kDa; we assigned

[1]Department of Neuroscience, Mayo Clinic, Mayo Clinic College of Medicine, 4500 San Pablo Road, Jacksonville, Florida 32224, USA. [2]Brain Mind Institute and School of Life Sciences, Swiss Federal Institute of Technology (EPFL), CH-1025 Lausanne, Switzerland. [3]Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Clemens Schöpf-Institute of Chemistry and Biochemistry, Technische Universitaet Darmstadt, Petersenstr. 22, D-64287 Darmstadt, Germany. [5]Laboratory for Neurodegenerative Research, Conway Institute University College Dublin, Dublin 4, Republic of Ireland. [6]Department of Neurosciences, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA.

this protein as α-secretase-derived APP CTF (CTF83) on the basis of molecular weight, increased expression in APP-transfected cells and immunoreactivity with an APP C-terminal antibody (Fig. 1e and Supplementary Fig. 2), but not an antibody against Aβ1–16 which would recognize β-secretase-derived APP CTF (CTF99; not shown). Labelling of APP CTF was photoactivation dependent (Fig. 1e) and no binding to PSEN1 CTF, PSEN1 amino-terminal fragment (NTF) or nicastrin was observed in these experiments (data not shown). These data demonstrate that labelling of APP CTF by GSMs can occur in cellular membrane fractions, providing further evidence that this interaction may be responsible for the modulation of amyloid-β production.

GSMs have been reported to modulate the site of γ-secretase cleavage in other substrates such as Notch[14]. We previously reported that APP is more sensitive than Notch to shifts in the location of γ-cleavage induced by GSMs[15]. Consistent with the differential sensitivity of Notch to GSMs that we observe, we find that whereas Fen-B does label a recombinant substrate derived from mouse Notch (Notch(C100)–Flag), this reaction is less efficient than Fen-B labelling of the APP(C100)–Flag (Fig. 1f). Furthermore, the presence of purified γ-secretase does not prevent labelling of either substrate by Fen-B (Supplementary Fig. 4). These data suggest that a differential affinity of the GSM Fen-B occurs between APP and other γ-secretase substrates such as Notch, and further link substrate targeting to the GSM properties of these compounds.

Initial mapping experiments showed that Fen-B did not label the last 50 amino acids of APP (APP intracellular domain, CTF-γ), raising the possibility that it was binding the amyloid-β region of APP (Supplementary Fig. 5). To define precisely the binding site of GSMs, C-terminal-truncated versions of amyloid-β were irradiated in the presence of Fen-B. Fen-B efficiently labelled Aβ1–40 and Aβ1–36 but did not label Aβ1–28 (Fig. 2a), suggesting that a minimal binding site on APP(C100)–Flag corresponds to residues 29–36 (GAIIGLMV) of amyloid-β (Fig. 2a). Aβ1–36 can also be labelled by Flurbi-BpB (Fig. 2b), and both GSMs can label Flag-tagged Aβ25–36 (Fig. 2c). These residues represent the start of the predicted APP transmembrane domain (625–632 of APP695) that lies within the membrane; however, this region of APP is accessible to small molecules[16]. Because the putative binding region of GSMs is also found in full-length APP we looked for labelling of both APP fragments in cells. Using microsomal membrane fractions from H4 cells expressing wild-type APP, we observed that both Fen-B and Flurbi-BpB labelled full-length APP and APP CTFs (Supplementary Fig. 6). Labelling of APP CTF and APP by GSM photoprobes was competed by both Aβ28–36 and structurally diverse

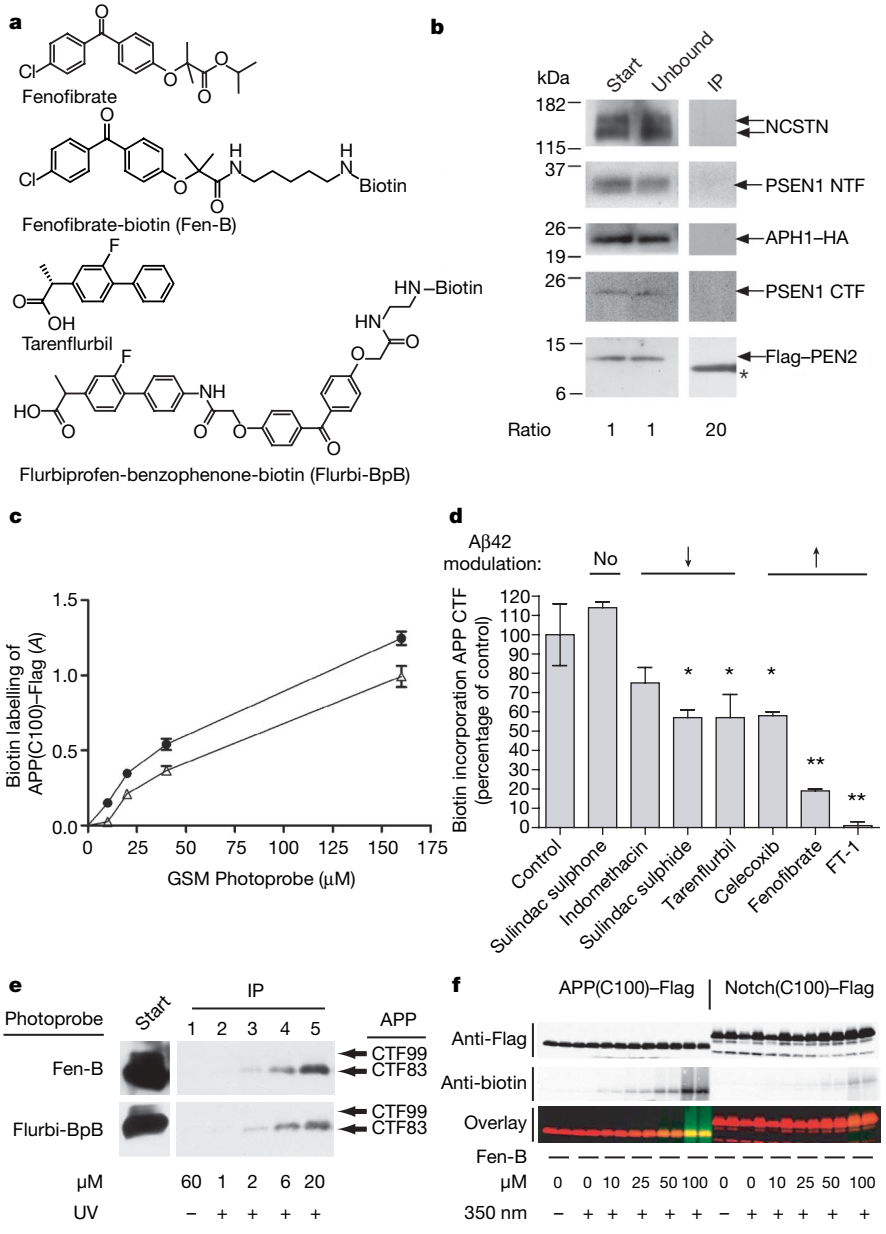


Figure 1 | GSM photoprobes label APP CTF. **a**, Structures of the parent GSMs (fenofibrate and tarenflurbil) and photoprobe derivates (Fen-B and Flurbi-BpB) are shown. **b**, The absence of PSEN1, NCSTN, APH1 and PEN2 labelling by the GSM Fen-B in a purified γ-secretase preparation (from CHO γ-30 cells[12]) and immunoprecipitation with streptavidin. The ratios of sample relative to the starting material are shown. Start and unbound lanes contain 5% of the immunoprecipitated material (lane 3), therefore the ratios are 1, 1 and 20. Asterisk denotes nonspecific reactivity with streptavidin. **c**, GSM photoprobes (Flurbi-BpB, closed circles, and Fen-B, open triangles) label a recombinant APP γ-secretase substrate (APP(C100)–Flag) with similar potency. *A*, absorbance; data are mean ± s.e.m., $n = 2$. **d**, Labelling of APP(C100)–Flag by Fen-B (10 μM) is competed by Aβ42-lowering and -raising GSMs (100 μM) but not by sulindac sulphone, a non-GSM NSAID. Data are presented as percentage control ± s.e.m., $n = 2$. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$; ANOVA with Dunnett's post-hoc analysis. **e**, GSM photoprobes label APP CTF from cells. CHAPSO solubilized membrane fractions from H4 APP-CTF–alkaline phosphatase cells were crosslinked with Fen-B and Flurbi-BpB (50 μM) and analysed by immunoprecipitation with streptavidin and immunblotting for APP (antibody CT20). Both GSMs label a fragment of APP that co-migrates with APP(C83). UV, ultraviolet. **f**, A GSM photoprobe preferentially labels a recombinant APP substrate (APP(C100)–Flag; left panel) relative to Notch (Notch(C100)–Flag; right panel). Samples were analysed by western blotting for incorporation of Fen-B. Green, biotin; red, Flag; yellow, dual reactivity; LiCor Odyssey.

GSMs, further supporting the specificity of this interaction (Supplementary Fig. 7). Taken together, these data provide evidence that the minimal binding site for raising and lowering GSM is Aβ28–36. Notably, this region of Aβ and APP is important for amyloid-β aggregation[17] and has been implicated as a binding site for amyloid-β aggregation inhibitors[18]. Furthermore, we[19] and others[20] have shown that mutations in this region of APP can drastically change the spectrum of amyloid-β species, supporting the notion that binding of GSMs to substrate in this region could alter γ-secretase cleavage. Indeed, we find that a peptide (I1, NH_2-FEGKF-$CONH_2$) known to bind to this region of amyloid-β acts as a GSM[17]. I1 elevated Aβ42 production without lowering total amyloid-β amounts, similar to fenofibrate (Fig. 2d).

If GSMs target the amyloid-β region of APP CTF, then any compound that binds to Aβ is a potential GSM. To test this possibility, we examined 15 compounds for GSM activity that had previously been reported as amyloid-β binding compounds, amyloid-β aggregation inhibitors or amyloid-β binding agents (Supplementary Fig. 8). Six of these compounds did not act as GSMs (for example, melatonin, BTA-1), but the remainder all showed GSM activity. These included two Aβ42-raising (the kinase inhibitor DAPH (4,5-dianilinophthalimide) and the calmodulin inhibitor, calmidazolium) and seven Aβ42-lowering GSMs (for example, the amyloid dye X-34 (1,4-*bis*(3-carboxy-4-hydroxyphenylethenyl)-benzene) and the hydrophobic probe Bis-ANS (4,4′-dianilino-1,1′-binaphthyl-5,5′-disulphonic acid); Fig. 3a). Immunoprecipitation and mass spectrometry analysis showed that the lowering of Aβ42 by X-34 and chrysamine G was accompanied by increased amounts of shorter amyloid-β, a characteristic signature of Aβ42-lowering GSMs (Supplementary Fig. 9). In cell-free γ-secretase assays, Congo red and chrysamine G lowered Aβ42 selectively, demonstrating that they are bona fide GSMs (Supplementary Fig. 10).

We focused on X-34 to characterize its GSM properties further. Cellular dose-response studies showed that X-34 lowered Aβ42 (effector concentration for half-maximum response, EC_{50}, 13.7 μM) and Aβ40 at higher concentrations (EC_{50} 60.7 μM), but did not decrease total amyloid-β amounts (Fig. 3b). X-34 competes for binding of APP(C100)–Flag by Fen-B and Flurbi-BpB (Fig. 3c). We also found that X-34 fluorescence increased when incubated with monomeric Aβ42. This allowed saturation binding experiments to determine that X-34 bound to Aβ42 with a dissociation constant (K_d) of 12.1 ± 4.7 μM (Supplementary Fig. 11). Thus, the affinity of X-34 for a peptide containing the putative GSM binding site is similar to concentrations at which it acts as a GSM in cells. Finally, we addressed whether X-34 modulated Aβ42 amounts in an APP transgenic mouse (Tg2576). We found that X-34 (100 mg kg^{-1}) decreased soluble Aβ42 (35%, $P < 0.01$) in the brains of Tg2576 mice with no effects on Aβ40 (Fig. 3d). Tarenflurbil (50 mg kg^{-1}) in the same paradigm also lowered Aβ42 (25%, $P < 0.05$) without a decrease in Aβ40 (123%, Fig. 3d). Collectively, these observations support the hypothesis that GSMs interact with substrate. Furthermore, the data show that selecting compounds on the basis of their ability to bind amyloid-β and APP is an efficient strategy to identify GSMs and suggest that screening compounds for binding to amyloid-β should be a simple and useful method to discover previously unknown GSMs.

On the basis of our previous findings which showed that GSMs bind a region in APP and amyloid-β that is involved in fibrilization[17], we reasoned that GSMs may also affect the aggregation of amyloid-β. Previous studies have shown that compounds with GSM activity, including certain NSAIDs[21], Congo red[22] and Bis-ANS[23], can inhibit the *in vitro* aggregation of synthetic amyloid-β. However, it is unknown if GSMs influence the concentration of secreted amyloid-β oligomers, such as those released by Chinese hamster ovary (CHO) cells expressing the APP V717F mutation (referred to as 7PA2 cells), which alter long-term potentiation and perturb the memory of learned behaviour when injected into rat brain[24,25]. To address this question we treated 7PA2 cells with two Aβ42-raising GSMs and a novel Aβ42-lowering GSM (Supplementary Fig. 12). Fenofibrate and FT-1 (ref. 7) increased amounts of Aβ42 whereas FT-9 lowered Aβ42, but all three GSMs decreased amounts of amyloid-β dimers and trimers. Extended dose experiments with fenofibrate and FT-9 show a consistent decrease in amyloid-β oligomers at doses where total amyloid-β amounts are not altered (Supplementary Fig. 12). In contrast to increased amyloid-β oligomer formation, when CHO cells are genetically manipulated to increase Aβ42 production by co-expression of APP and FAD-linked mutant PSEN1 proteins, we find that both Aβ42-lowering and -raising GSMs decrease oligomer formation[26]. Notably, amyloid-β oligomers in 7PA2 and neuronal cells are generated intracellularly before secretion[27].

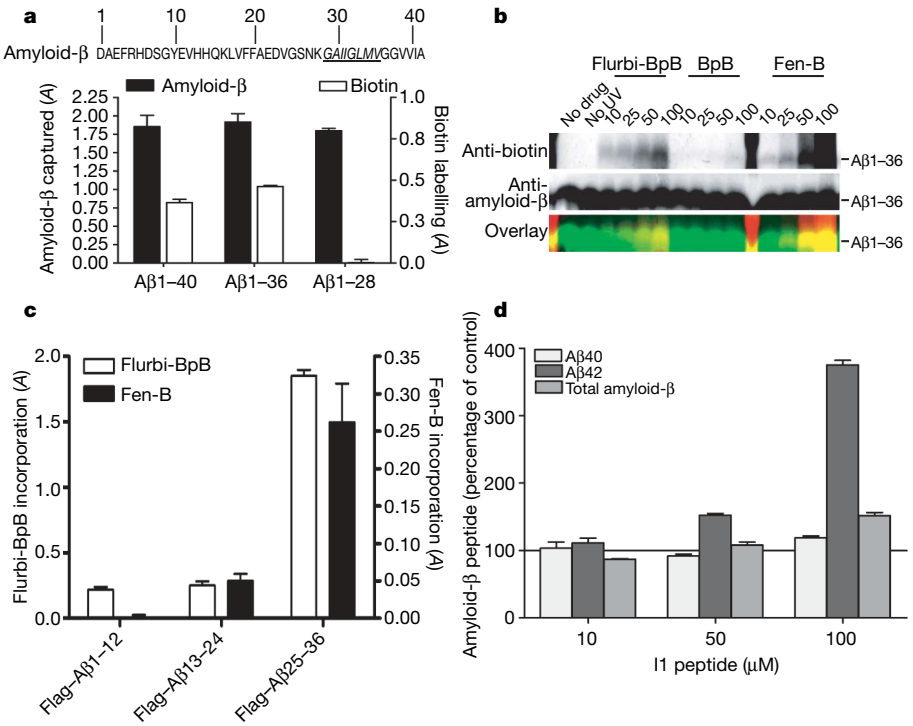


Figure 2 | GSM photoprobes bind to the amyloid-β region of APP. **a**, Fen-B labels Aβ1–40 and Aβ1–36 but not Aβ1–28, suggesting that the binding site for Fen-B is located between residues 28 to 36 of amyloid-β, which are highlighted in italics. **b**, Flurbi-BpB and Fen-B label Aβ1–36 (biotin incorporation), whereas the photoaffinity tag alone (BpB) shows minor labelling. **c**, Flurbi-BpB and Fen-B preferentially label Flag-tagged Aβ25–36. Data are presented as biotin incorporation (absorbance, *A*) ± s.e.m., $n = 3$. **d**, The peptide fragment I1 (NH_2-FEGKF-$CONH_2$) increases Aβ42 in H4 cells expressing APP similar to the GSM fenofibrate.

Therefore, the finding that GSMs shift amyloid-β cleavage by targeting a region of substrate present in the amyloid-β cleavage product provides a mechanistic link between GSM activity and general anti-amyloid-β aggregation effects and also suggests that the binding of GSM to substrate and inhibition of amyloid-β oligomerization can occur in the same cellular compartment.

Our data suggest that mutation of the GSM binding site in APP should alter sensitivity to GSMs. To test this hypothesis we exchanged a portion of the GSM binding site in APP with the analogous region of human NOTCH (Fig. 4a, b and Supplementary Fig. 13). We used this chimaeric construct because in our hands the NOTCH transmembrane domain (TMD) is resistant to GSM effects[15]. The APP-NOTCH TMD construct produces a spectrum of chimaeric amyloid-β species in conditioned cell media (Supplementary Fig. 13). Although the molecular weight and abundance of each chimaeric peptide differs from amyloid-β peptides from APP, the main species produced from the APP-NOTCH TMD construct have identical C termini to Aβ1–40 and Aβ1–42, and can be quantified by enzyme-linked immunosorbent assays (ELISAs; Supplementary Fig. 13). We therefore examined the relative sensitivity to GSMs of APP-NOTCH TMD versus APP in stably transfected cells. Cleavage of APP-NOTCH TMD was not significantly affected by either an Aβ42-lowering GSM or an Aβ42-raising GSM but remained sensitive to inhibition by treatment with the γ-secretase inhibitor L-685458 (Fig. 4a, b and Supplementary Fig. 13). Thus, exchange of residues in the putative binding site render the substrate much less sensitive to GSMs.

Several explanations for the GSM activity of NSAIDs have been proposed. It has recently been hypothesized that NSAIDs alter dimerization of APP[20]. Although plausible, there is no conclusive evidence that an APP dimer is the substrate for γ-secretase, and it is difficult to account for both Aβ42-raising and -lowering properties of structurally related GSMs with this model. Other studies propose that GSMs lower Aβ42 by shifting presenilin conformation through allosteric binding[28,29]. Our data implicating substrate targeting by GSMs is compatible with these previous findings (Supplementary Discussion). It is also possible that GSMs shift cleavage by changing the position of APP CTF in the plane of the membrane leading to shifts in the amyloid-β fragments produced[19]. This model is attractive as it accounts for the difference between the classes of GSMs: Aβ42-raising GSMs would 'pull' the substrate out of the membrane, whereas Aβ42-lowering GSMs allow the substrate to 'sink' further into the membrane.

Given the likelihood that GSMs exist that directly target the enzyme, we believe it is appropriate to refer to the GSMs we have identified as substrate-targeting GSMs. Substrate-targeting GSMs can, in theory, have two therapeutic consequences—alteration in Aβ42 production and inhibition of amyloid-β aggregation—that might synergistically benefit the Alzheimer's disease phenotype (Supplementary Fig. 1). Future substrate-targeting GSMs could be developed that optimize the selective Aβ42-lowering and anti-aggregation properties. These data may also explain why substrate-targeting GSMs that raise Aβ42 may not be as risky as our data initially implied[7]: although they raise Aβ42, they may also inhibit amyloid-β aggregation. Finally, these data demonstrate a new means of modulating intramembrane cleaving proteases through substrate-binding small molecules, which may have broad therapeutic applications[2,30].

METHODS SUMMARY

Photoaffinity labelling. All experiments used the following general protocol. Samples (recombinant or synthetic peptides, cell lysates and membrane preparations) were exposed in borosilicate test tubes to ultraviolet light (350 nm) in a Rayonet Photoreactor R100 (RPR-3500 lamps, Southern New England Ultra Violet Company) in a cold room (4 °C) for 30 min or as otherwise noted. Crosslinked samples were analysed by ELISA for biotin incorporation, or analysed directly by SDS–PAGE (Criterion-XT, Bio-Rad), or after precipitation with streptavidin ultralink plus beads (Pierce). After immunblotting, proteins were detected using chemiluminesence (ECL Plus, GE Healthcare) or near-infrared fluorescence (LiCor Odyssey). Full-length APP and APP CTFs (CTF83, CTF99) were detected with CT20, a rabbit polyclonal antibody against the C terminus of APP. Biotin was detected with an affinity-purified rabbit polyclonal antibody (Bethyl).

Cell-based screens for amyloid-β modulation. Human H4 neuroglioma cells (American Type Culture Collection, ATCC) expressing wild-type APP695 protein or CTF105 fused to secreted alkaline phosphatase which is efficiently processed to APP CTFs (CTF83, CTF99) and produce high levels of amyloid-β, were used for the cell-based screens as previously described[7]. Cells were incubated for 5–6 h in the presence of the various compounds in Opti-Mem culture medium

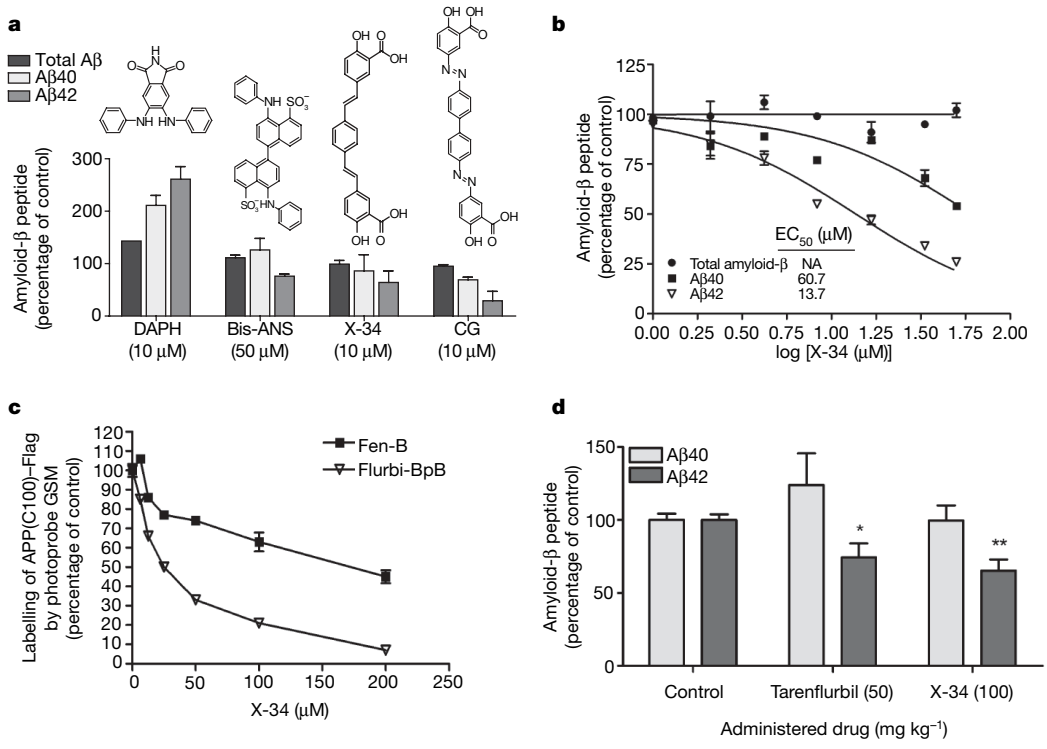


Figure 3 | Compounds that bind Aβ are GSMs *in vitro* and *in vivo*. **a**, A cell-based screen of Aβ-binders identified molecules that increase Aβ42 (DAPH) or decrease Aβ42 (Bis-ANS, X-34 or chrysamine G (CG)). Data are mean ± s.e.m., $n = 3$. **b**, X-34 is an Aβ42-lowering GSM. Changes in amyloid-β peptide amounts after X-34 treatment are shown. Data are presented as percentage control ± s.e.m., $n = 3$. EC_{50} values were calculated as described in Methods. Total amyloid-β did not decrease. NA, not applicable. **c**, X-34 binds to APP(C100)–Flag and decreases labelling by GSM photoprobes. Biotin incorporation into APP by Fen-B and Flurbi-BpB is presented as percentage of control (peptide without X-34) ± s.e.m., $n = 2$. **d**, X-34 lowers Aβ42 in Tg2576 mice after 4 h. X-34 ($n = 7$) and tarenflurbil ($n = 5$) reduce Aβ42 selectively; control ($n = 7$). Data are presented as amyloid-β percentage of control relative to vehicle ± s.e.m. Animals per group (n). Asterisk, $P < 0.05$; two asterisks, $P < 0.01$; ANOVA with Dunnett's post-hoc analysis.

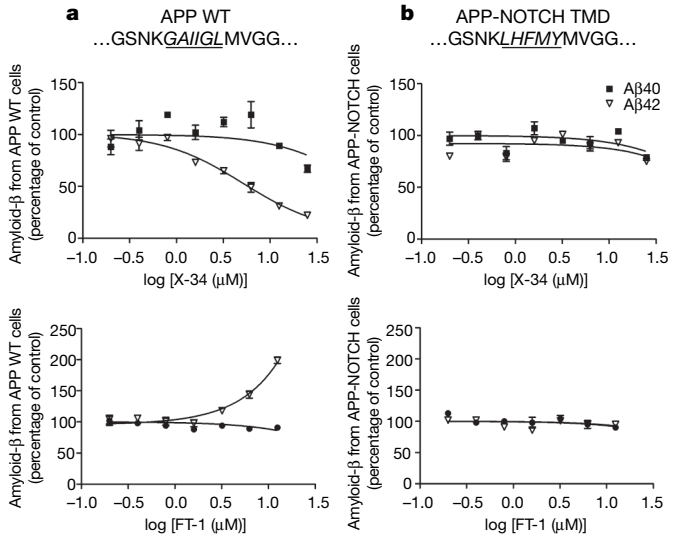


Figure 4 | The ability of GSMs to shift Aβ42 amounts is sensitive to the amino acid sequence of the binding site on APP. **a**, **b**, The sequences of wild-type (WT; **a**) APP and the mutated substrate (**b**) containing the homologous region (italic, underlined) of the NOTCH transmembrane domain (TMD) in APP where GSMs are hypothesized to bind. Top row, X-34 lowers Aβ42 (EC_{50} 5.9 μM) from APP wild-type cells (**a**) but did not change either Aβ40 or Aβ42 concentrations in the APP-NOTCH TMD line (**b**). Bottom row, FT-1 (12.5 μM) raised Aβ42 200% in APP wild-type cells (**a**); however, in APP-NOTCH TMD cells (**b**), FT-1 caused minimal changes in the Aβ42 (95%) or Aβ40 (90%) signal. Data are presented as amyloid-β percentage of control ± s.e.m., $n = 3$.

containing 1% fetal bovine serum. Compounds were dissolved in dimethylsulphoxide (DMSO; 0.5% final concentration) and diluted 200-fold. Media was analysed for various amyloid-β species (40, 42 and total) using ELISAs as described below and previously[7]. The EC_{50} values for changes in amyloid-β species were calculated by fitting sigmoid dose-response curves using nonlinear regression in Prism (GraphPad) and are shown as values ± s.e.m.

Statistical analysis. Data are presented as either percentage control or mean ± s.e.m. Results were analysed using Prism (Graph Pad) with *t*-tests or one-way analysis of variance analyses (ANOVAs) with Dunnett's post-hoc correction for comparison of multiple samples to a control. Statistical significance is shown as $P < 0.05$ (one asterisk), $P < 0.01$ (two asterisks) or $P < 0.001$ (three asterisks).

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

Received 20 December 2007; accepted 1 May 2008.

1. Kukar, T. & Golde, T. E. Possible mechanisms of action of NSAIDs and related compounds that modulate γ-secretase cleavage. *Curr. Top. Med. Chem.* **8,** 47–53 (2008).
2. Kodadek, T. Inhibition of proteolysis and other posttranslational modifications with substrate-targeted inhibitors. *Biopolymers* **66,** 134–140 (2002).
3. Eriksen, J. L. *et al.* NSAIDs and enantiomers of flurbiprofen target γ-secretase and lower Aβ 42 *in vivo*. *J. Clin. Invest.* **112,** 440–449 (2003).
4. Weggen, S. *et al.* A subset of NSAIDs lower amyloidogenic Aβ42 independently of cyclooxygenase activity. *Nature* **414,** 212–216 (2001).
5. Christensen, D. D. Changing the course of Alzheimer's disease: anti-amyloid disease-modifying treatments on the horizon. *Prim. Care Companion J. Clin. Psychiatry* **9,** 32–41 (2007).
6. Kukar, T. *et al.* Chronic administration of R-flurbiprofen attenuates learning impairments in transgenic amyloid precursor protein mice. *BMC Neurosci.* **8,** 54 (2007).
7. Kukar, T. *et al.* Diverse compounds mimic Alzheimer disease-causing mutations by augmenting Aβ42 production. *Nature Med.* **11,** 545–550 (2005).
8. Narlawar, R., Baumann, K., Czech, C. & Schmidt, B. Conversion of the LXR-agonist TO-901317–From inverse to normal modulation of γ-secretase by addition of a carboxylic acid and a lipophilic anchor. *Bioorg. Med. Chem. Lett.* **17,** 5428–5431 (2007).
9. Leuchtenberger, S. *et al.* Inhibitors of Rho-kinase modulate amyloid-β (Aβ) secretion but lack selectivity for Aβ42. *J. Neurochem.* **96,** 355–365 (2006).
10. Weggen, S. *et al.* Evidence that nonsteroidal anti-inflammatory drugs decrease amyloid β42 production by direct modulation of γ-secretase activity. *J. Biol. Chem.* **278,** 31831–31837 (2003).
11. Dorman, G. & Prestwich, G. D. Benzophenone photophores in biochemistry. *Biochemistry* **33,** 5661–5673 (1994).
12. Fraering, P. C. *et al.* Purification and characterization of the human γ-secretase complex. *Biochemistry* **43,** 9774–9789 (2004).
13. Narlawar, R. *et al.* Scaffold of the cyclooxygenase-2 (COX-2) inhibitor carprofen provides Alzheimer γ-secretase modulators. *J. Med. Chem.* **49,** 7588–7591 (2006).
14. Okochi, M. *et al.* Secretion of the Notch-1 Aβ-like peptide during Notch signaling. *J. Biol. Chem.* **281,** 7890–7898 (2006).
15. Sagi, S. A. *et al. NSAIDs Show Specificity for APP γ-Secretase Cleavage over Notch and CD44* (Society for Neuroscience, Program no. 264.10, Abstract Viewer/Itinerary Planner, Washington DC, 2004).
16. Grziwa, B. *et al.* The transmembrane domain of the amyloid precursor protein in microsomal membranes is on both sides shorter than predicted. *J. Biol. Chem.* **278,** 6803–6808 (2003).
17. Sato, T. *et al.* Inhibitors of amyloid toxicity based on beta-sheet packing of Aβ40 and Aβ42. *Biochemistry* **45,** 5503–5516 (2006).
18. Chen, Y.-R. & Glabe, C. G. Distinct early folding and aggregation properties of Alzheimer amyloid-β peptides Aβ40 and Aβ42. *J. Biol. Chem.* **281,** 24414–24422 (2006).
19. Murphy, M. P. *et al.* γ-secretase, evidence for multiple proteolytic activities and influence of membrane positioning of substrate on generation of amyloid β peptides of varying length. *J. Biol. Chem.* **274,** 11914–11923 (1999).
20. Munter, L. M. *et al.* GxxxG motifs within the amyloid precursor protein transmembrane sequence are critical for the etiology of Aβ42. *EMBO J.* **26,** 1702–1712 (2007).
21. Hirohata, M., Ono, K., Naiki, H. & Yamada, M. Non-steroidal anti-inflammatory drugs have anti-amyloidogenic effects for Alzheimer's β-amyloid fibrils *in vitro*. *Neuropharmacology* **49,** 1088–1099 (2005).
22. Podlisny, M. B. *et al.* Oligomerization of endogenous and synthetic amyloid β-protein at nanomolar levels in cell culture and stabilization of monomer by Congo red. *Biochemistry* **37,** 3602–3611 (1998).
23. Ferrao-Gonzales, A. D. *et al.* Controlling β-amyloid oligomerization by the use of naphthalene sulfonates: trapping low molecular weight oligomeric species. *J. Biol. Chem.* **280,** 34747–34754 (2005).
24. Walsh, D. M. *et al.* Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* **416,** 535–539 (2002).
25. Calabrese, B. *et al.* Rapid, concurrent alterations in pre- and postsynaptic structure induced by naturally-secreted amyloid-β protein. *Mol. Cell. Neurosci.* **35,** 183–193 (2007).
26. Xia, W. *et al.* Enhanced production and oligomerization of the 42-residue amyloid β-protein by Chinese hamster ovary cells stably expressing mutant presenilins. *J. Biol. Chem.* **272,** 7977–7982 (1997).
27. Walsh, D. M., Tseng, B. P., Rydel, R. E., Podlisny, M. B. & Selkoe, D. J. The oligomeriation of amyloid β protein begins intracellularly in cells derived from human brain. *Biochemistry* **39,** 10831–10839 (2000).
28. Lleo, A. *et al.* Nonsteroidal anti-inflammatory drugs lower Aβ42 and change presenilin 1 conformation. *Nature Med.* **10,** 1065–1066 (2004).
29. Beher, D. *et al.* Selected non-steroidal anti-inflammatory drugs and their derivatives target γ-secretase at a novel site: evidence for an allosteric mechanism. *J. Biol. Chem.* **279,** 43419–43426 (2004).
30. Dong, D. L., Liu, R., Sherlock, R., Wigler, M. H. & Nestler, H. P. Molecular forceps from combinatorial libraries prevent the farnesylation of Ras by binding to its carboxyl terminus. *Chem. Biol.* **6,** 133–141 (1999).

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

Acknowledgements This work was supported by grants from the US National Institutes of Health National Institute on Aging (to D.M.W., M.S.W., T.E.G., A.F. and E.H.K.), the AFAR CART award (T.E.G.), and the Mayo Foundation. T.L.K. was supported by the American Health Assistance Foundation ADR program, the Mayo ADRC, and a Robert and Clarice Smith Fellowship. A.T.W. was funded by a European Union 6th Framework Marie Curie Early Stage Training fellowship. P.C.F was supported by the Swiss National Science Foundation and by the NCCR 'Neural Plasticity and Repair'. 21F12 was a gift from P. Seubert and D. Schenk.

Author Contributions T.L.K., D.M.W., E.H.K. and T.E.G. conceived the project. T.L.K., T.B.L. and M.A.B. planned, performed and analysed the photoaffinity experiments and amyloid-β assays. P.C.F., W.Y., and M.S.W. performed photoaffinity experiments with purified γ-secretase and APP(C100)–Flag and Notch(C100)–Flag; T.L.K. wrote most of the paper. T.L.K., M.A.B., C.V. and R.W.P. performed and analysed mouse experiments. K.J.-W. cloned constructs and made stable cell lines. S.S., B.C., J.T. and E.H.K. made the APP-NOTCH chimaera construct and cell line and performed mass spectrometry analysis of amyloid-β. A.T.W. and D.M.W. analysed amyloid-β oligomers. D.Y. and C.E. screened and provided compound libraries. G.M.M., B.H., R.C., and A.F. synthesized fenofibrate and various photoprobes. J.E. characterized initial photoprobes. R.N. and B.S. synthesized flurbiprofen and related photoprobes. B.M., V.R., B.C. and T.L.R. performed mass spectrometry analysis of amyloid-β, APP-NOTCH amyloid-β and saturation binding experiments with X-34.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to T.L.K. (kukar.thomas@mayo.edu) and T.E.G. (golde.todd@mayo.edu).

METHODS

Compounds and peptides. The following commercially available chemicals and peptides used in this work were purchased: BSB ((trans,trans)-1-bromo-2,5-bis-(3-hydroxycarbonyl-4-hydroxy)styrylbenzene), chrysamine G, FSB ((E,E)-1-fluoro-2,5-bis-(3-hydroxycarbonyl-4-hydroxy)styrylbenzene), half chrysamine G and amyloid-β peptides 1–40, 1–36 and 1–28 (Anaspec); APP-CTF50 (rPeptide); aspirin, BTA-1, Bis-ANS, clioquinol, Congo red, curcumin, DAPH, fenofibrate, melatonin, naproxen, sulindac sulphide and sulphone (Sigma); *R*-flurbiprofen (Cayman); NIAD4 (Nomadics). The compounds Fen-B, Flurbi-BpB, BpB, FT-1 and FT-9 were synthesized as described in Supplementary Information. AOI987, LY-411575 and X-34 were synthesized by the Mayo Clinic Laboratory of Synthetic Organic Chemistry following published references as described in Supplementary Information. The I1 peptide fragment (NH_2-FEGKF-$CONH_2$) and N-terminal Flag-tagged truncated amyloid-β fragments (Flag-Aβ1–12, 13–24 and 25–36) were synthesized by the Mayo Clinic MPRC Peptide Synthesis Facility. Aβ28–36 was synthesized by Elim Biopharmaceuticls; Celecoxib was a gift from Myriad Pharmaceuticals.

Peptide crosslinking. Amyloid-β peptides were diluted from DMSO stocks into PBS (5 μM final concentration), incubated with photoprobes and crosslinked. Flag–amyloid-β fragments (10 μM) were incubated with a GSM photoprobe (Fen-B, 100 μM; Flurbi-BpB, 40 μM) in PBS and crosslinked for 30 min. Recombinant peptides (APP(C100)–Flag and Notch(C100)–Flag; 1 μM final concentration) were diluted from stock into buffer (50 mM HEPES, pH 7.4, 0.25% CHAPSO) incubated with competitor (if required) at 37 °C for 15 min, spiked with photoprobe and crosslinked.

Membrane isolation. H4 cells overexpressing wild-type APP695 were lysed by nitrogen bomb cavitation at 700 p.s.i. for 1 h in PBS (4639, Parr instrument company), and then spun at 1000*g* for 10 min to remove unbroken cells and nuclei. Membranes were pelleted from the supernatant after centrifugation at 100,000*g*. Membranes were homogenized in sodium carbonate buffer (100 mM, pH 11.5) by brief sonication or resuspension using a glass–teflon homogenizer to remove peripherally-associated membrane proteins and spun again. Membranes were then extracted with 0.25% CHAPSO in 50 mM HEPES, pH 7.4, and spun to remove insoluble proteins.

Mass spectrometry of amyloid-β. For matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry analyses of amyloid-β peptides, the same H4 cell lines used in screens were treated with the indicated compounds as described[4]. Secreted amyloid-β peptides were analysed by minor modifications of the technique previously described[7]. Aβ1–x was immunoprecipitated from conditioned medium with antibody 9 covalently coupled to seize beads (Pierce), with synthetic Aβ40 with methionine 35 substituted by norleucine ([Nle35]-β-Amyloid(1-40); Anaspec) added as an immunoprecipitation control and mass standard. Proteins were eluted into 0.1% trifluoroacetic acid, 75% acetonitrile. Samples were mixed 1:1 with alpha-cyano-4-hydroxycinnamic acid matrix (6.2 mg ml^{-1}) in methanol:acetonitrile:water (36%:56%:8%; Agilent). Samples were analysed on a 4800 MALDI–TOF–TOF (Applied Biosystems).

***In vitro* γ-secretase assays.** These assays were performed as previously described[3,7]. Briefly, buoyant cholesterol-rich fractions showing enriched γ-secretase activity were isolated from H4 cells stably overexpressing secreted alkaline phosphatase fused to the N terminus of APP(C105) that had been treated with IL-CHO, a reversible γ-secretase inhibitor, to accumulate APP CTF. Fractions 3–5 (1 ml each) were collected and activity was measured by incubation at 37 °C for 2 h and comparing amyloid-β and CTF-γ (AICD) amounts to a frozen sample. Amyloid-β concentrations were determined by sandwich ELISA and CTF-γ by western blotting with antibody CT-20.

ELISA. Amyloid-β species were detected with sandwich ELISAs using monoclonal antibodies as previously described[7]. Briefly, amyloid-β peptides were captured by C-terminal-specific antibodies for Aβ40 (antibody 40.1) or Aβ42 (antibody 42.2) that were coated on Immulon 4 HBX ELISA plates (Thermo Scientific) at 25 μg ml^{-1} in PBS. Captured amyloid-β was then detected by an HRP-conjugated antibody reactive to the N-terminal epitope 1–16 of amyloid-β (antibody 9). To measure incorporation of the photoprobes and corresponding biotin tag into peptides two strategies were used: (1) recombinant γ-secretase substrates (APP(C100)–Flag and Notch(C100)–Flag) or Flag-epitope tagged fragments of amyloid-β (Flag–Aβ1–12, 13–24 and 25–36) were captured on anti-Flag ELISA plates (M2 coated; Sigma) and biotin incorporation was detected with streptavidin- or NeutrAvidin-HRP (Pierce); and (2) C-terminal truncations of amyloid-β (1–28, 1–36 and 1–40) were captured at the N terminus by antibody 9, and again, biotin incorporation was measured by detection with streptavidin- or NeutrAvidin-HRP.

Production and purification of γ-secretase and recombinant substrates. The CHO γ-30 cell line expresses human PSEN1, Flag–PEN2 and APH1α2–HA. γ-secretase was purified from γ-30 using a multi-step purification as previously described[12]. For western blotting analysis of γ-secretase, samples were run on 4–20% Tris-glycine PAGE gels, transferred to polyvinylidene difluoride filter membranes and probed with antibody 14 (for PSEN1–NTF, 1:2,000, a gift from S. Gandy), antibody N19 (for PSEN1–NTF, 1:200, Santa Cruz), antibody 13A11 (for PSEN1–CTF, a gift from Elan Pharmaceuticals), antibody 3F10 (for APH1α2–HA, Roche), anti-Flag M2 antibody (for Flag–PEN2, 1:1,000, Sigma), R302 antibody (for NCSTN, 1:4,000, a gift from D. Miller and P. Savam), and guinea-pig anti-NCSTN antibody (1:2,000, Chemicon). Recombinant γ-secretase substrates, APP(C100)–Flag and Notch(C100)–Flag (N100-Flag), were produced in *Escherichia coli* and purified over an anti-Flag M2 affinity column (Sigma) as previously described[12]. APP(C100)–Flag is derived from the C-terminal fragment of APP (CTF99), the endogenous precursor to amyloid-β, with the addition of an N-terminal methinone and a Flag epitope tag to facilitate detection and purification[12]. The Notch(C100)–Flag is derived from the analogous region of mouse Notch[12].

Animal studies. Female Tg2576 mice were treated with test compounds at 3 months of age before amyloid-β deposition. Compounds were mixed with vehicle (PEG-400 90%, DMSO 10%) and delivered by oral gavage (200 μl) and animals were killed 4 h later. Plasma, cerebellum and two hemi-brains were saved from each animal. One hemi-brain was extracted using RIPA buffer (1 ml per 150 mg) with complete protease inhibitors (Roche) and amyloid-β concentrations were assessed with Aβ40 and Aβ42 ELISAs. All animal studies were approved by the Mayo Clinic Institutional Animal Care and Use Committee.

Determination of X-34 affinity for Aβ1–42 by fluorescence titration. When X-34 was mixed with monomeric Aβ1–42 in 10 mM Tris-HCl (pH 8.0) and 10% DMSO, enhancement of X-34 fluorescence was noted at excitation wavelengths above 366 nm, the wavelength which gives maximum X-34 absorbance. Fluorescence was monitored on a Cary Eclipse Fluorescence Spectrophotometer (Varian Instruments) at 25 °C with excitation at 412 nm, emission at 502 nm, and excitation and emission slits at 2.5 and 10 nm, respectively. Stock solution A (10 μM Aβ1–42 in 10 mM Tris pH 8.0 and 10% DMSO) was titrated by sequential addition of solution B (10 μM Aβ1–42 plus varying amounts of X-34 (μM) in the same buffer). A parallel series of blank fluorescence values was obtained by omitting Aβ1–42 from solutions A and B. Data were fitted to the following equation by unweighted nonlinear regression analysis (SigmaPlot), with $[L]_{tot}$ as the independent variable and K_d and F_C as the fitted parameters[31]:

$$\Delta F = 0.5F_{C-L}[D-(D^2-4[A\beta]_{tot}[L]_{tot})^{1/2}]$$

where F_C is the fluorescence of X-34 in the presence of Aβ1–42 and F_L is the fluorescence in its absence; $D = [A\beta]_{tot} + [L]_{tot} + K_d$, where $[A\beta]_{tot}$ and $[L]_{tot}$ are the total Aβ1–42 and X-34 concentrations, respectively, and K_d is the equilibrium dissociation constant; and F_{C-L} is the difference in fluorescence intensity coefficients for bound and free X-34. Because high concentrations of X-34 were used in the titration, the absorbance of X-34 was significant even at the 412 nm excitation wavelength ($\varepsilon_{412\,nm} = 850\,M^{-1}cm^{-1}$), observed fluorescence values were corrected for inner filter effects as described previously[32].

Construction of APP-NOTCH chimaera cDNA and selection of stable cell lines. APP695 was subcloned into the retroviral vector pLHCX (BD Biosciences/Clontech) to facilitate generation of stable cell lines. The retroviral expression vector for the APP-NOTCH TMD chimaera was generated by site-directed mutagenesis with the Stratagene Quick-Change kit according to the manufacturer's instructions. Primers were designed to insert the analogous sequence of human NOTCH1 and excise the corresponding sequence of APP (shown in Fig. 4). All new constructs were sequenced to assure accuracy. To generate a stable cell line, CHO cells were infected with retroviruses encoding APP-NOTCH TMD and selected with hygromycin B. GSMs did not consistently change the amounts of chimaeric Aβ X-40 or X-42 from the APP-NOTCH TMD line and thus an EC_{50} could not be calculated.

31. De Ferrari, G. V., Mallender, W. D., Inestrosa, N. C. & Rosenberry, T. L. Thioflavin T is a fluorescent probe of the acetylcholinesterase peripheral site that reveals conformational interactions between the peripheral and acylation Sites. *J. Biol. Chem.* **276,** 23282–23287 (2001).

32. Johnson, J. L., Cusack, B., Davies, M. P., Fauq, A. & Rosenberry, T. L. Unmasking tandem site interaction in human acetylcholinesterase. Substrate activation with a cationic acetanilide substrate. *Biochemistry* **42,** 5438–5452 (2003).

LETTERS

Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer

Ryohei Sekido[1] & Robin Lovell-Badge[1]

The mammalian Y chromosome acts as a dominant male determinant as a result of the action of a single gene, *Sry*, whose role in sex determination is to initiate testis rather than ovary development from early bipotential gonads[1–3]. It does so by triggering the differentiation of Sertoli cells from supporting cell precursors, which would otherwise give follicle cells. The related autosomal gene *Sox9* is also known from loss-of-function mutations in mice and humans to be essential for Sertoli cell differentiation[4,5]; moreover, its abnormal expression in an XX gonad can lead to male development in the absence of *Sry*[6,7]. These genetic data, together with the finding that *Sox9* is upregulated in Sertoli cell precursors just after SRY expression begins[8,9], has led to the proposal that *Sox9* could be directly regulated by SRY. However, the mechanism by which SRY action might affect *Sox9* expression was not understood. Here we show that SRY binds to multiple elements within a *Sox9* gonad-specific enhancer in mice, and that it does so along with steroidogenic factor 1 (SF1, encoded by the gene *Nr5a1* (*Sf1*)), an orphan nuclear receptor. Mutation, co-transfection and sex-reversal studies all point to a feedforward, self-reinforcing pathway in which SF1 and SRY cooperatively upregulate *Sox9* and then, together with SF1, SOX9 also binds to the enhancer to help maintain its own expression after that of SRY has ceased. Our results open up the field, permitting further characterization of the molecular mechanisms regulating sex determination and how they have evolved, as well as how they fail in cases of sex reversal.

SRY contains a high-mobility group (HMG)-box DNA-binding domain characteristic of the SOX family of transcription factors. It is transiently expressed for a few hours in each Sertoli cell precursor between embryonic day (E)10.5 and E12.5. Since its discovery in 1990, a variety of mechanisms have been proposed by which SRY might act: first, as a repressor (or antagonist) of a repressor of male development[10]; second, through effects on local chromatin structure[11]; third, through a role in pre-mRNA splicing[12]; or fourth, as a specific transcriptional activator of one or more critical male-specific target genes, perhaps through partner proteins[13]. Moreover, several genes have been proposed as SRY targets, but so far with no evidence of a direct interaction.

Sf1 is essential in mice for the development of the gonads of both sexes from about E11.5 (ref. 14). However, SF1 seems particularly important for testis differentiation because heterozygotes can show XY female sex reversal in humans[15], and decreased expression of Sertoli cell-specific genes in mice[16]. SF1 is first expressed during mouse gonadal development at E9.5 by cells within the coelomic epithelium[17]. These give rise to daughter cells that enter the genital ridge, a proportion retaining SF1 and becoming supporting cell precursors[9]. We previously showed, using mice transgenic for a Myc-epitope-tagged *Sry* gene (Sry^{Myc}), that SRY is expressed exclusively in these precursor cells, which then show rapid upregulation of *Sox9* (ref. 9). Once SOX9 has reached a critical threshold, SRY is repressed by means of a SOX9-dependent negative feedback loop[5,9]. SF1 and SOX9 expression are subsequently maintained at high levels in Sertoli cells[17].

Sox9 is expressed in various tissues during embryogenesis[18] by means of a complex regulatory region thought to be spread over at least 1 megabase (Mb). XX sex reversal in *Odsex* mice was originally ascribed to a deletion associated with a transgene insertion about 1 Mb upstream of *Sox9*, but this is now considered to be a long-range effect of the transgene on an unidentified gonad-specific enhancer[19]. Transgenic mouse studies, with human *SOX9* YAC clones (−350 to +250 kilobases (kb))[20] or elements highly conserved between human and *Fugu SOX9* within −290 to +450 kb (ref. 21), revealed enhancers for several tissues, but not for the gonad.

We initiated our search by using a mouse *Sox9* bacterial artificial chromosome (BAC) clone (−70 to +50 kb) carrying a *lacZ* reporter (Fig. 1a). Mice transgenic for this construct (*120-lacZ*) recapitulated a significant subset of the endogenous *Sox9* expression pattern (Fig. 1c–g). Gonadal LacZ expression started shortly after E10.5 (data not shown) and increased up to E11.5 in both sexes, before becoming upregulated in the testis and downregulated in the ovary by E12.5 (Fig. 1h). Because LacZ co-localized with endogenous SOX9 in the testis, the transgene was expressed exclusively in the Sertoli cell lineage (Fig. 1h).

The 70-kb upstream sequence (*70-lacZ*) gave a similar expression pattern to that of *120-lacZ* (Supplementary Fig. 1), implying the lack of crucial tissue-specific elements within 3′ sequences. Deletions of the 70 kb revealed that 16 kb immediately 5′ of this (*16-lacZ*) still retained gonadal expression, but 7.0 kb (*7.0-lacZ*) did not (Fig. 1a). The 16-kb region was divided into four fragments (16a, 16b, 16c and 16d) and each was assayed in conjunction with the *hsp68* minimal promoter linked to *lacZ* or *CFP* (encoding cyan fluorescent protein) reporter genes. Fragments 16a and 16b, but not 16c or 16d, gave expression at E11.5 in both XX and XY genital ridges (Fig. 1b). We therefore tested the 3.2-kb region of overlap between 16a and 16b and found it sufficient for gonad-specific expression in an orientation-independent manner (Fig. 2a). It gave little, if any, expression in XX genital ridges; we therefore named it TES (for testis-specific enhancer of *Sox9*).

Because the normal level of *Sox9* expression in Sertoli cells depends on *Sry* and *Sf1* (ref. 9; data not shown), we investigated the epistatic relationship between TES and these two genes. The enhancer was active in XXSry^{Myc}, but completely inactive in XY$Sf1^{\Delta/\Delta}$ gonads (Fig. 2b, c), demonstrating that TES activity is downstream of both SRY and SF1. However, because SRY is present only transiently within Sertoli cell precursors, *Sox9* expression must be maintained by other factors, such as fibroblast growth factor 9/fibroblast growth factor receptor 2 (FGF2/FGFR2)[22,23]. This raises the possibility that the enhancer activity is a consequence of maintenance mechanisms rather than initial upregulation by SRY. We therefore backcrossed *tes-lacZ* onto mice carrying a conditional null allele of *Sox9*. TES was

[1]Division of Developmental Genetics, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

still active in homozygous XY null mutants, in cells located within the central portion of the genital ridge where SRY expression starts (Fig. 2d). At this stage it was absent from XX genital ridges, whether from $Sox9^{\Delta/\Delta}$ or wild-type animals (data not shown). As SF1 and other factors will be present in both sexes, but SRY will be XY-specific even in the absence of SOX9 (ref. 5), this proves that the enhancer is able to respond to SRY.

The question still remains whether SRY and SF1 physically interact with TES to regulate *Sox9* directly in the gonad. To address this, we performed chromatin immunoprecipitation (ChIP) assays. Twelve sets of primers spanning TES were designed to amplify any enriched fragments by PCR. In the SRY ChIP assays, we took advantage of the specific antibody against the Myc tag in Sry^{Myc} transgenic mice. There was a significant enrichment of seven fragments (IP fragments (IPFs) 2, 4, 8, 9, 10, 11 and 12) (Fig. 3a). We searched for putative SRY-binding sites (SRY-BSs) in each. Four possessed a single (IPF 12) or multiple (IPFs 2, 4 and 8) SRY-BSs, whereas three (IPFs 9, 10 and 11) contained none (Fig. 3d). In SF1 ChIP, seven fragments (IPFs 5, 6, 7, 8, 10, 11 and 12) were enriched (Fig. 3b). Each contained single (IPFs 8, 10, 11 and 12) or multiple (IPFs 5 and 7) SF1-binding sites (SF1-BSs), except for IPF 6 (Fig. 3d). Because SRY-BSs might be occupied by SOX9 after SRY expression has ceased, SOX9 ChIP was performed with E13.5 gonads. Six fragments (IPFs 5, 7, 8, 10, 11 and 12) were enriched (Fig. 3c), but it is notable that not all the fragments enriched in SRY ChIP were detected. In particular IPFs 2 and 4 were missing, suggesting that these sites are dispensable for autoregulation. IPFs 5, 7, 10 and 11, which were detected in SF1 ChIP, were also enriched even though they lack SRY-BSs (Fig. 3d), implying that SF1 bound to its BSs recruits SOX9 through protein–protein interactions[24]. The enrichment of sequences lacking either consensus site (for example IPFs 6 and 9) could be due to interactions with unidentified partner proteins, to cryptic *in vivo* binding sites, or to inefficient DNA fragmentation such that they were amplified by PCR even though the protein was bound to sites in adjacent DNA. To validate our assays further, we tested *Amh* (anti-Müllerian hormone) regulatory sequences known to be a direct target of SF1 and SOX9 *in vivo*[25] and detected specific enrichment in the testis (Fig. 3b, c).

To examine whether all the SF1- and SRY-BSs are required, we divided TES into three subfragments (EC1, EC2 and EC3) and found that gonadal expression was governed by EC2, the 1.9-kb central region (Fig. 4a). We then compared upstream sequences of *Sox9* between several vertebrate species and identified a 1.4-kb region within EC2, highly conserved in rat, dog and human (Supplementary Fig. 3). This core sequence was found to be sufficient for the enhancer activity (Fig. 4a) and named TESCO (for TES core).

To verify the significance of SRY-BSs and SF1-BSs for TESCO activity, we mutated the three consensus SRY-BSs (SRYmut), the six SF1-BSs (SF1mut) and all these sites together (BOTHmut) (Fig. 4b). Transgenic mice with SRYmut or SF1mut showed enhancer activities similar to that of the wild-type sequence, whereas those with BOTHmut showed none at all, except for one (that shown) with very weak expression (Fig. 4c). This indicates that SRY and SF1 bind TESCO directly and activate it synergistically. This synergism was quantified by co-transfection assays in COS7 cells, in which SF1 activated TESCO fivefold on its own, SRY by itself was ineffective, but a tenfold activation was observed with both factors together. This synergism was also seen in CHO and HEK-293 cells (Supplementary Fig. 4). The activity disappeared with the BOTHmut reporter construct, and almost with SF1mut, although fivefold activation by SF1 remained with SRYmut as expected (Fig. 4d). Because SF1 and SOX9 interact physically, we next tested their synergism and observed a 45-fold synergistic activation in COS7 cells (Fig. 4e). SRY and SOX9 by themselves failed to activate TESCO, despite their respective very weak and strong transactivation domains[18,26]. This was expected, because SOX protein function usually depends on partner factors. Significant levels of activation remained when either SRY-BSs or

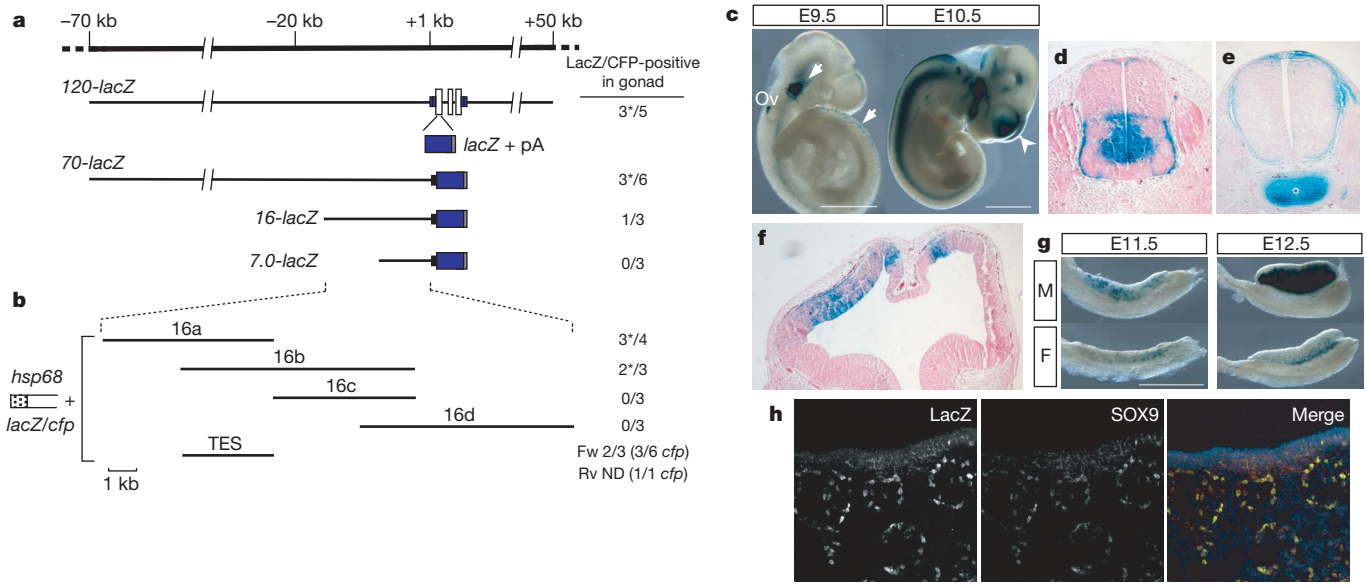


Figure 1 | Characterization of the *Sox9* locus to identify a testis-specific enhancer. a, **b**, The 120-kb BAC clone containing the *Sox9* gene, which consists of three exons (black boxes, non-coding; white boxes, coding). **a**, The reporter construct *120-lacZ* was made by replacing part of the coding region in exon 1 with *lacZ* (blue) followed by an SV40 polyadenylation signal (pA, grey). Relevant deletion constructs are shown (*70-lacZ*, *16-lacZ* and *7.0-lacZ*). **b**, Five subfragments (16a–16d and TES) assayed for their enhancer activities using *lacZ* or *cfp* reporter genes driven by the *hsp68* minimal promoter. The TES element provides testis-specific expression in both the forward (Fw) and reverse (Rv) orientations. Numbers indicate ratios of testes positive for LacZ or CFP to total transgenic lines at E12.5. ND, not determined. Asterisks: two also showed *lacZ* expression in the ovary. **c**, Expression of the *120-lacZ* transgene in whole embryos at E9.5 and E10.5. Prominent β-galactosidase activity was seen in the otic vesicle (Ov), cranial and trunk neural crest cells (arrows) and the telencephalon (arrowhead). **d**–**f**, Transverse sections through the trunk region at E10.5 (**d**) and E12.5 (**e**), counterstained with eosin, and a coronal section through the telencephalon at E11.5 (**f**) show *120-lacZ* expression in the ventral ventricular/subventricular zone of the spinal cord, sclerotomal vertebral condensation and the dorsal telencephalon, respectively. **g**, The transgene is also expressed in developing male (M) and female (F) gonads from E11.5 to E12.5. **h**, Male gonad immunostained for LacZ (left) and endogenous SOX9 (middle) shows their co-localization in Sertoli cells at E12.5 (right, yellow signals). Scale bars, 1.5 mm (**c**); 1 mm (**g**).

SF1-BSs were mutated, whereas very little activation was detected with both sets of sites mutated (Fig. 4h).

Sox9 has been found to be activated in all cases of XX male sex reversal so far examined. To examine whether the enhancer can also be activated *in vivo* in the absence of SRY, E13.5 ovaries and testes (as positive controls) from *Tesco-cfp* transgenic mice were grafted to kidney capsules of three-month-old wild-type females, where the ovaries transdifferentiate into testis-like structures[8]. As expected, CFP expression was maintained in testis grafts. Although initially negative for CFP, the transgene became active in the ovary grafts within ten days, with expression particularly evident at 30 days. Immunohistochemical analysis showed that all CFP-positive cells in the latter were positive for SF1 and SOX9, and were therefore Sertoli-like, whereas SF1 single-positive cells presumably correspond to steroidogenic cells (Supplementary Fig. 5).

These data, together with results obtained by ChIP, co-transfection assays, mutation analysis in transgenics and epistatic tests *in vivo*, support a model for the regulation of *Sox9* expression in the gonad, consisting of three phases: initiation, upregulation and maintenance (Supplementary Fig. 6). First, the initiation, which depends on SF1 but could involve other factors, need only be weak and sex-independent: it has to be sufficient to sensitize the gonad to SRY, but it must be below the threshold required to establish any additional self-reinforcing regulatory pathway. TES may not reliably reproduce the low level of *Sox9* expression seen in XX (and XY) gonads at E10.5. Nevertheless, this is seen with slightly longer fragments, which have additional SF1-BSs. WT1 is another candidate for an initiating factor because certain mutations lead to XY female sex reversal[27] and it is expressed in the supporting cell lineage, where it acts upstream of *Sf1* (ref. 28). However, WT1 and its splice variants had no effect on TESCO (data not shown). Second, SRY then acts synergistically with SF1 to upregulate *Sox9*. There may be a physical interaction between SRY and SF1, as shown previously for the HMG box of SOX9 and the carboxy-terminal domain of SF1 with respect to *Amh* regulation[24]. This could explain the enrichment of fragments lacking SRY-BS (Fig. 3d) and also previous results[29] in which mis-expression of *Sry* in transgenic mice failed to induce ectopic *Sox9* expression in SF1-negative cells. The presence of multicopy transgenes, recruitment of the other factor by protein–protein interactions and maintenance mechanisms probably account for the expression

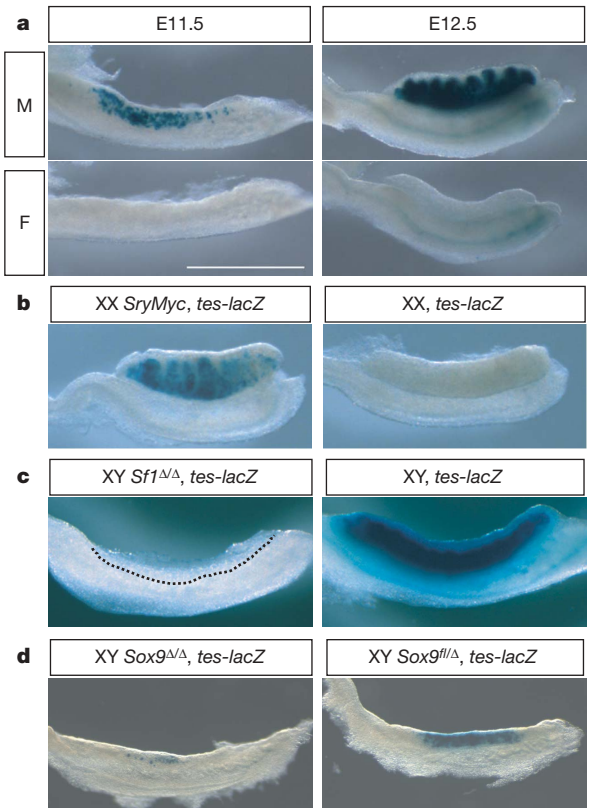


Figure 2 | Gonadal expression of *tes-lacZ* and genetic interactions with *Sry*, *Sf1* and *Sox9*. **a**, In the wild type, TES enhancer activity is detected exclusively in male gonads. **b**, **c**, The activity is induced in the gonads of XXSry^{Myc} sex-reversed mice at E12.5 (**b**, left) and is totally abolished in those of XY$Sf1^{\Delta/\Delta}$ null mutant mice at E11.5 (**c**, left). The right panels are controls. The size of the XY$Sf1^{\Delta/\Delta}$ gonad (demarcated by the dotted line) is reduced because apoptosis has started. **d**, At E11.5 in XY$Sox9^{\Delta/\Delta}$ null mutant mice, LacZ is detectable only in the central region of the gonads (left), whereas in littermates heterozygous for the flox allele (XY$Sox9^{fl/\Delta}$) it is seen throughout the gonads (right). Scale bar, 1 mm.

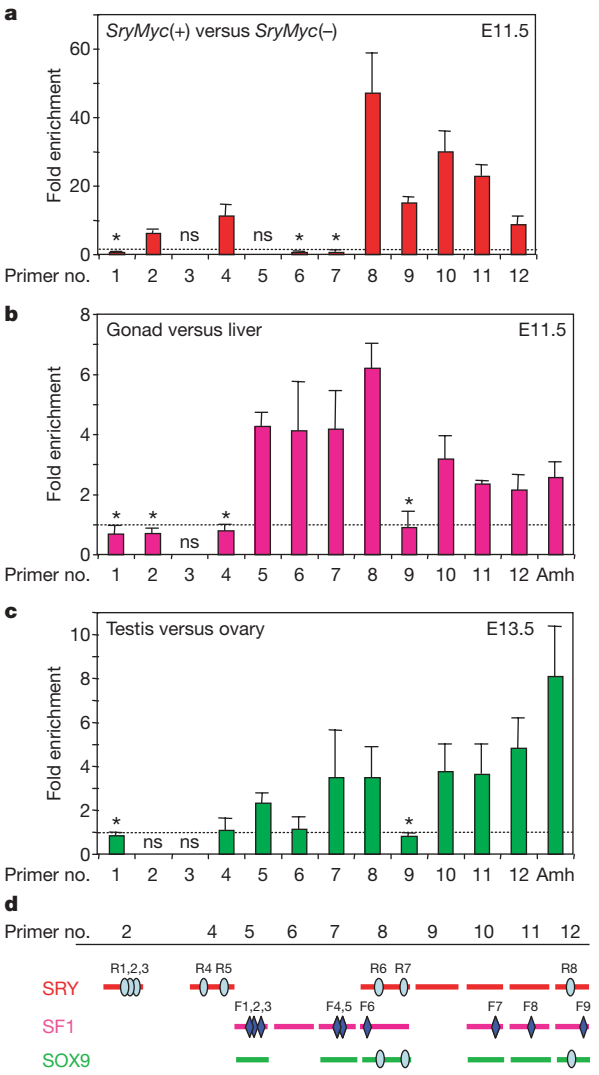


Figure 3 | Chromatin immunoprecipitation (ChIP) assays for SF1, SRY and SOX9. **a**–**c**, Quantification of fragments enriched by SRY (**a**), SF1 (**b**) and SOX9 (**c**) ChIP. An anti-Myc antibody against Myc-tagged SRY was used in SRY ChIP, whereas specific antibodies against the endogenous proteins were used in other ChIPs. Fold enrichment is expressed relative to negative control tissue, meaning that values greater than 1 represent specific enrichment. Means and s.d. from a minimum of three independent duplicate assays ($n = 6$ or more) are shown, except for Amh ($n = 4$). ns, no signal detected in either positive or negative control tissue. Asterisks, values below one (0.68–0.91) reflect site-specific backgrounds and do not imply specific enrichment in the negative control tissues. **d**, Summary of the ChIP assays. Enriched fragments are indicated by colour bars, namely red (SRY), pink (SF1) and green (SOX9). The locations of consensus SF1-BSs (F1–F9, blue diamonds) and SRY-BSs (R1–R8, light green circles) in each enriched fragment are shown. Their sequences are shown in Supplementary Fig. 2.

still seen *in vivo* when only one type of site was mutated. Third, *Sox9* requires other factors for its maintenance after SRY has ceased. An *Fgf9*/*Fgfr2* positive feedback loop has been described[22,23]. This may go through TES or TESCO (B. Capel, personal communication). There is also evidence that SOX9 may regulate SF1 expression[30], which would ensure sufficient levels of its own partner factor. In addition, we show here the likely presence of an autoregulatory loop in which SOX9 activates itself.

Although our transgenic analysis revealed an enhancer for *Sox9* that specifically recapitulates the endogenous expression pattern in the normal testis and during sex reversal, we cannot rule out the presence of additional gonad-specific enhancers, or regulatory elements interacting with TES, outside the 120-kb BAC. However, by itself TES has all the properties expected of such an enhancer. Furthermore, even though binding sites for SF1 and SRY/SOX9 are spread throughout the 3.2-kb TES, the 1.4-kb TESCO is sufficient for enhancer activity. It is possible that the additional sites are non-functional even if they are occupied by their cognate factors.

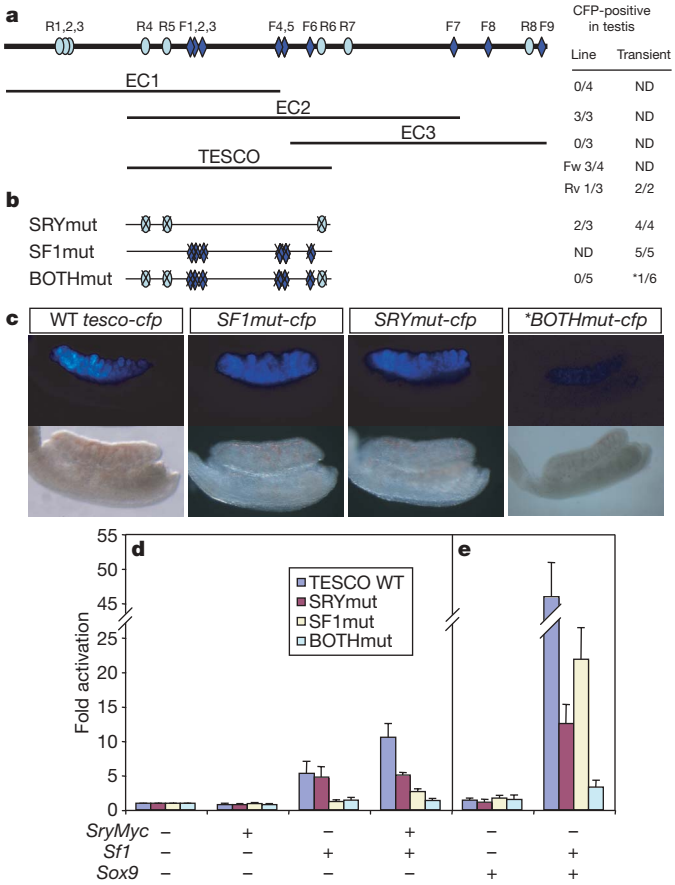


Figure 4 | Identification of the 1.4-kb enhancer core and its activation by SF1, SRY and SOX9 in a synergistic manner. a, Further deletion of TES (EC1–EC3) and comparative genome-based approaches (Supplementary Fig. 3) reveal the 1.4-kb enhancer core TESCO. **b**, Mutational analysis in transgenic mice shows no enhancer activity only when both SF1-BSs and SRY-BSs are mutated, except for one very weak sample (asterisk). Numbers indicate ratios of testes positive for CFP to total transgenics of either established lines (line) or founder embryos (transient) at E12.5. ND, not determined. **c**, CFP expression in E12.5 testes driven by wild-type and mutant TESCO sequences shown in **b**. Top, CFP fluorescence; bottom, bright field. Strong CFP fluorescence was observed with wild-type (WT) TESCO, SF1mut and SRYmut. No CFP fluorescence was detected with BOTHmut, except in one showing faint fluorescence (asterisk in **b**). **d**, **e**, *In vitro* co-transfection assays with Sry^{Myc} (**d**) or *Sox9* (**e**) in COS7 cells for quantifying mutational effects. Fold activation was calculated relative to the base level (no effectors transfected for each reporter construct). Means and s.d. from at least three independent experiments are shown.

Alternatively, this may reflect redundancy of binding sites as observed in other cases of transcriptional regulation. Finally, TES seems to integrate not only all the inputs into the regulation of *Sox9* required to follow a Sertoli cell fate, but also those that prevent this in ovaries, and will therefore help to define their identity and mode of action.

METHODS SUMMARY

The 120-kb *Sox9* BAC clone was isolated from a 129^{SV} mouse library (Research Genetics). ChIP assays were performed with the DNA–protein complex fragmented to an average size of about 300 base pairs. In the grafting experiments, gonads dissected from *Tesco-cfp* transgenic embryos at E13.5 were grafted to the kidney capsules of three-month-old wild-type females.

Received 29 January; accepted 20 March 2008.
Published online 4 May 2008.

1. Gubbay, J. *et al.* A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a new family of embryonically expressed genes. *Nature* **346,** 245–250 (1990).
2. Sinclair, A. H. *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346,** 240–244 (1990).
3. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. Male development of chromosomally female mice transgenic for *Sry*. *Nature* **351,** 117–121 (1991).
4. Foster, J. W. *et al.* Campomelic dysplasia and autosomal sex reversal caused by mutation in an *SRY*-related gene. *Nature* **372,** 525–530 (1994).
5. Chaboissier, M.-C. *et al.* Functional analysis of *Sox8* and *Sox9* during sex determination in the mouse. *Development* **131,** 1891–1901 (2004).
6. Huang, B., Wang, S., Ning, Y., Lamb, A. N. & Bartley, J. Autosomal XX sex reversal caused by duplication of *SOX9*. *Am. J. Med. Genet.* **87,** 349–353 (1999).
7. Vidal, V. P. I., Chaboissier, M.-C., Rooji, D. G. D. & Schedl, A. *Sox9* induces testis development in XX transgenic mice. *Nature Genet.* **28,** 216–217 (2001).
8. Morais da Silva, S. *et al. Sox9* expression during gonadal development implies a conserved role for the gene in testis differentiation in mammals and birds. *Nature Genet.* **14,** 62–68 (1996).
9. Sekido, R., Bar, I., Narváez, V., Penny, G. & Lovell-Badge, R. SOX9 is up-regulated by the transient expression of SRY specifically in Sertoli cell precursors. *Dev. Biol.* **274,** 271–279 (2004).
10. McElreavey, K., Vilain, E., Abbas, N., Herskowitz, I. & Fellous, M. A regulatory cascade hypothesis for mammalian sex determination: SRY represses a negative regulator of male development. *Proc. Natl Acad. Sci. USA* **90,** 3368–3372 (1993).
11. Pontiggia, A. *et al.* Sex-reversing mutations affect the architecture of SRY-DNA complexes. *EMBO J.* **13,** 6115–6124 (1994).
12. Ohe, K., Lalli, E. & Sassone-Corsi, P. A direct role of SRY and SOX9 proteins in pre-mRNA splicing. *Proc. Natl Acad. Sci. USA* **99,** 1146–1151 (2002).
13. Thevenet, L. *et al.* NHERF2/SIP-1 interacts with mouse SRY via a different mechanism than human SRY. *J. Biol. Chem.* **280,** 38625–38630 (2005).
14. Luo, X., Ikeda, Y. & Parker, K. L. A cell-specific nuclear receptor is essential for adrenal and gonadal development and sexual differentiation. *Cell* **77,** 663–673 (1994).
15. Lin, L. *et al.* Heterozygous missense mutations in steroidogenic factor 1 (SF1/Ad4BP, NR5A1) are associated with 46,XY disorders of sex development with normal adrenal function. *J. Clin. Endocrinol. Metab.* **92,** 991–999 (2007).
16. Park, S. Y. *et al.* Nuclear receptors Sf1 and Dax1 function cooperatively to mediate somatic cell differentiation during testis development. *Development* **132,** 2415–2423 (2005).
17. Ikeda, Y. *et al.* Comparative localization of Dax-1 and Ad4BP/SF-1 during development of the hypothalamic–pituitary–gonadal axis suggests their closely related and distinct function. *Dev. Dyn.* **220,** 363–376 (2001).
18. Ng, L.-J. *et al.* SOX9 binds DNA, activates transcription, and coexpresses with type II collagen during chondrogenesis in the mouse. *Dev. Biol.* **183,** 108–121 (1997).
19. Qin, Y. *et al.* Long-range activation of *Sox9* in *Odd Sex* (*Ods*) mice. *Hum. Mol. Genet.* **13,** 1213–1218 (2004).
20. Wunderle, V. M., Critcher, R., Hastie, N., Goodfellow, P. N. & Schedl, A. Deletion of long-range regulatory elements upstream of *SOX9* causes campomelic dysplasia. *Proc. Natl Acad. Sci. USA* **95,** 10649–10654 (1998).
21. Bagheri-Fam, S. *et al. Long-range* upstream and downstream enhancers control distinct subsets of the complex spatiotemporal *Sox9* expression pattern. *Dev. Biol.* **291,** 382–397 (2006).
22. Kim, Y. *et al. Fgf9* and *Wnt4* act as antagonistic signals to regulate mammalian sex determination. *PLoS Biol.* **4,** 1000–1009 (2006).
23. Kim, Y. *et al.* Fibriblast growth factor receptor 2 regulates proliferation and Sertoli differentiation during male sex determination. *Proc. Natl Acad. Sci. USA* **104,** 16558–16563 (2007).

24. de Santa Barbara, P. *et al.* Direct interaction of SRY-related protein SOX9 and steroidgenic factor 1 regulates transcription of the human anti-Mullerian hormone gene. *Mol. Cell. Biol.* **18,** 6653–6665 (1998).
25. Arango, N. A., Lovell-Badge, R. & Behringer, R. R. Targeted mutagenesis of the endogenous mouse *Mis* gene promoter: *in vivo* definition of genetic pathways of vertebrate sexual development. *Cell* **99,** 409–419 (1999).
26. Dubin, R. A. & Ostrer, H. Sry is a transcriptional activator. *Mol. Endocrinol.* **12,** 1182–1192 (1994).
27. Hammes, A. *et al.* Two splice variants of Wilm's tumor 1 gene have distinct functions during sex determination and nephron formation. *Cell* **106,** 319–329 (2001).
28. Wilhelm, D. & Englert, C. The Wilms tumor suppressor WT1 regulates early gonad development by activation of *Sf1*. *Genes Dev.* **16,** 1839–1851 (2002).
29. Kidokoro, T. *et al.* Influence on spatiotemporal patterns of a male-specific *Sox9* activation by ectopic *Sry* expression during early phases of testis differentiation in mice. *Dev. Biol.* **278,** 511–525 (2005).
30. Shen, J. H. & Ingraham, H. A. Regulation of the orphan nuclear receptor steroidogenic factor 1 by Sox proteins. *Mol. Endocrinol.* **16,** 529–540 (2002).

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

Acknowledgements We thank J. Carvajal and T. Kondo for teaching us BAC engineering and ChIP assays, respectively; D. Bell and S. Wood for help in generating transgenic mice; M. Cheung for analysing neural crest cells in transgenic embryos; K. Parker for providing SF1 complementary DNA and SF1 null mutant mice; A. Schedl for providing conditional *Sox9* mutant mice; K. Morohashi for providing the Ad4BP/SF1 antibody; H. Sasaki for providing the Asshsp68lacZ vector; Y. Kamachi for providing the pCMV/SV2-Sox9 and pδ51LucII vectors; and V. Harley and the members of the Division of Developmental Genetics in the National Institute for Medical Research, especially C. Wise, F. W. Buaas, J. Turner and P. Burgoyne, for suggestions and discussions. This work was supported by the Medical Research Council and the Louis Jeantet Foundation. R.S. was a recipient of long-term fellowships from the European Molecular Biology Organization and the Human Frontier Science Program.

Author Contributions R.S. was responsible for most of the experiments, and R.L.B. for a few experiments as well as for directing the laboratory. Both planned the project and 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 R.S. (rsekido@nimr.mrc.ac.uk) or R.L.B (rlovell@nimr.mrc.ac.uk).

ERRATUM

doi:10.1038/nature07622

Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer

Ryohei Sekido & Robin Lovell-Badge

Nature 453, 930–934 (2008)

The Supplementary Methods for this Letter should have been uploaded at the time of publication. This oversight has now been rectified.

RETRACTION

doi:10.1038/nature07645

The RNA-binding protein FCA is an abscisic acid receptor

Fawzi A. Razem, Ashraf El-Kereamy, Suzanne R. Abrams & Robert D. Hill

Nature 439, 290–294 (2006)

Portions of the work repeated with respect to abscisic acid (ABA) binding have revealed errors in the calculations associated with Fig. 1, with the result that the molar ratio of ABA bound to FCA is substantially lower than claimed. There are also difficulties with the data in Fig. 2a, b that arose from the preparation of FY. We conclude that there is no effect of ABA on the FCA–FY interaction, and therefore requested to retract this paper on 14 July 2008. See the Brief Communication Arising in this issue[1].

1. Risk, J. M., Macknight, R. C. & Day, C. L. FCA does not bind abscisic acid. *Nature* doi:10.1038/nature07646 (this issue).

LETTERS

Draper-dependent glial phagocytic activity is mediated by Src and Syk family kinase signalling

Jennifer S. Ziegenfuss[1]*, Romi Biswas[2]*, Michelle A. Avery[1], Kyoungja Hong[2], Amy E. Sheehan[1], Yee-Guide Yeung[2], E. Richard Stanley[2] & Marc R. Freeman[1]

The cellular machinery promoting phagocytosis of corpses of apoptotic cells is well conserved from worms to mammals. An important component is the *Caenorhabditis elegans* engulfment receptor CED-1 (ref. 1) and its *Drosophila* orthologue, Draper[2]. The CED-1/Draper signalling pathway is also essential for the phagocytosis of other types of 'modified self' including necrotic cells[3], developmentally pruned axons[4,5] and dendrites[6], and axons undergoing Wallerian degeneration[7]. Here we show that *Drosophila* Shark, a non-receptor tyrosine kinase similar to mammalian Syk and Zap-70, binds Draper through an immunoreceptor tyrosine-based activation motif (ITAM) in the Draper intracellular domain. We show that Shark activity is essential for Draper-mediated signalling events *in vivo*, including the recruitment of glial membranes to severed axons and the phagocytosis of axonal debris and neuronal cell corpses by glia. We also show that the Src family kinase (SFK) Src42A can markedly increase Draper phosphorylation and is essential for glial phagocytic activity. We propose that ligand-dependent Draper receptor activation initiates the Src42A-dependent tyrosine phosphorylation of Draper, the association of Shark and the activation of the Draper pathway. These Draper–Src42A–Shark interactions are strikingly similar to mammalian immunoreceptor–SFK–Syk signalling events in mammalian myeloid and lymphoid cells[8,9]. Thus, Draper seems to be an ancient immunoreceptor with an extracellular domain tuned to modified self, and an intracellular domain promoting phagocytosis through an ITAM-domain–SFK–Syk-mediated signalling cascade.

Developing tissues produce excessive numbers of cells and selectively destroy a subpopulation through programmed cell death to regulate growth. Rapid clearance of cell corpses is essential for maintaining tissue homeostasis and preventing the release of potentially cytotoxic or antigenic molecules from dying cells, and defects in cell corpse clearance are closely associated with autoimmune and inflammatory diseases[10–13]. In *C. elegans* the CED-1 receptor is expressed in engulfing cells, where it acts to recognize cell corpses and drive their phagocytosis[1]. CED-1 promotes engulfment through an intracellular NPXY motif, a binding site for proteins containing a phosphotyrosine-binding (PTB) domain, and a YXXL motif, a potential interaction site for proteins containing SH2 domains[1]. The PTB domain adaptor protein CED-6 can bind the NPXY motif of CED-1 (ref. 14), is required for cell corpse engulfment[15] and acts in the same genetic pathway as CED-1 (ref. 16). CED-1 ultimately mediates actin-dependent cytoskeletal reorganization through the Rac1 GTPase[17], and Dynamin modulates vesicle dynamics downstream of CED-1 during engulfment[18], but the molecular signalling cascade that allows CED-1 to execute phagocytic events remains poorly defined.

Glia are the primary phagocytic cell type in the developing and mature brain. Glia rapidly engulf neuronal cell corpses produced during development, as well as neuronal debris generated during axon pruning[19,20] or during Wallerian degeneration in the adult brain[21]. In *Drosophila*, glial phagocytosis of these engulfment targets requires Draper, the fly orthologue of CED-1 (refs 2, 4–6). Draper, like CED-1, contains 15 extracellular atypical epidermal growth factor (EGF) repeats, a single transmembrane domain, and NPXY and YXXL motifs in its intracellular domain (Fig. 1a)[2]. *Drosophila* Ced-6 is also required for the clearance of pruned axons[4], indicating possible conservation of the interaction between CED-1 and CED-6 in flies, but additional signalling molecules acting downstream of Draper have not been identified.

In a yeast two-hybrid screen for molecules interacting with the regulatory region of Shark[22,23], we identified Draper. We found that when LexA–Shark, constitutively active Src kinase and AD–Draper are present, Shark and Draper interact physically (Fig. 1a). In the absence of Src kinase, Shark and Draper fail to interact, indicating that phosphorylation of Draper by Src may be essential for Shark–Draper interactions. We found that the Draper intracellular domain contains an ITAM (YXXI/L-X_{6-12}-YXXL), a key domain found in many mammalian immunoreceptors including Fc, T-cell and B-cell receptors. SFKs phosphorylate the tyrosines in ITAM domains, thereby allowing ITAM association with SH2-domain-containing signal transduction proteins including Syk and Zap-70 (refs 9, 24). We therefore generated Y→F substitutions of the tyrosine residues within or near the Draper ITAM, and found that Tyr 949 and Tyr 934 were critical for robust Draper–Shark binding (Fig. 1a, b). These correspond to the consensus tyrosine residues in the predicted Draper ITAM (Fig. 1a). We next transfected plasmids with carboxy-terminally haemagglutinin-tagged Draper (Draper–HA) or with Draper–HA and Shark with an amino-terminal Myc tag (Myc–Shark) into *Drosophila* S2 cells, immunoprecipitated with anti-HA antibodies, and performed western blots with anti-phosphotyrosine, anti-Myc and anti-HA antibodies (Fig. 1c). We found that Myc–Shark co-immunoprecipitated with Draper–HA, and that anti-phosphotyrosine antibodies labelled a band corresponding to the position of Draper–HA that was absent in empty vector controls. Further, we found that a Y949F substitution markedly reduced Draper–Shark association (Fig. 1c). Taken together, these data indicate Draper and Shark can associate physically through the Draper ITAM domain.

We next sought to determine whether Shark is required for glial phagocytic activity *in vivo*. Severing adult *Drosophila* olfactory receptor neurons (ORNs) initiates Wallerian degeneration of ORN axons. Antennal lobe glia surrounding these severed axons respond to this

[1]Department of Neurobiology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605-2324, USA. [2]Department of Developmental and Molecular Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA.
*These authors contributed equally to this work.

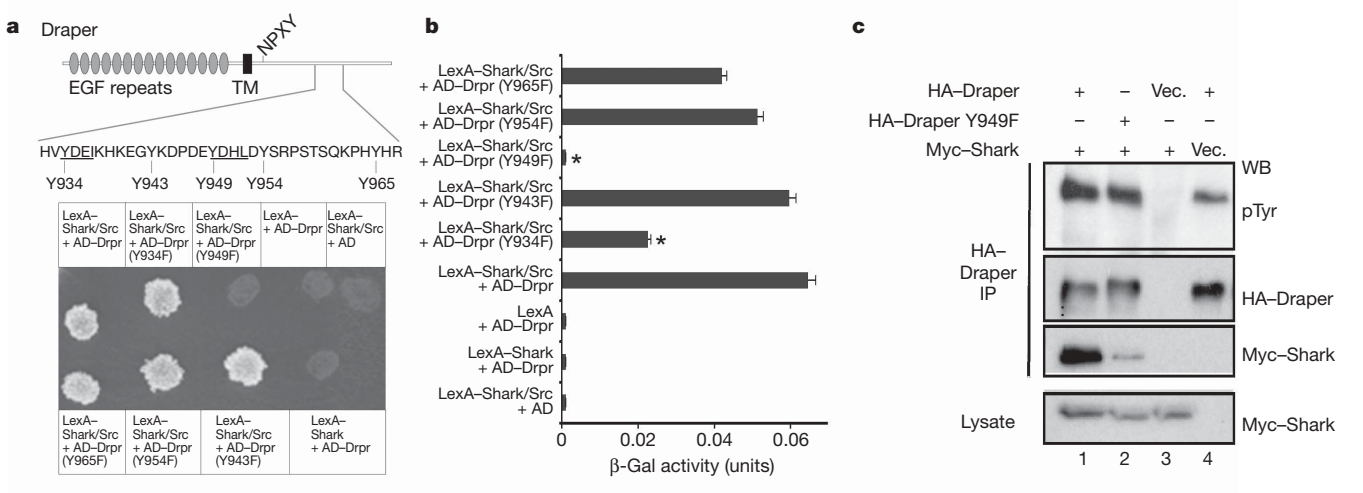


Figure 1 | Shark binds an ITAM in the Draper intracellular domain. **a**, Draper contains an ITAM domain from Y934-L952 (YXXI-X_{11}-YXXL). The requirements for the five tyrosine residues within and adjacent to this domain (shown) and Src were assayed in the yeast two-hybrid system[23]. **b**, Lysates from yeast cultures in **a** were tested in quantitative β-galactosidase (β-Gal) assays. Error bars represent s.e.m.; $n = 3$; *, $P < 0.05$. **c**, S2 cells were transfected with *pMT-Myc–Shark* and *pMT-Drpr-I* constructs. Draper immunoprecipitates (IP) were analysed by western blotting (WB) with anti-phosphotyrosine (pTyr), anti-Myc and anti-HA antibodies. TM, transmembrane; Vec., vector.

injury by extending membranes towards severed axons and engulfing degenerating axonal debris[7]. These glia express high levels of Draper, and in *draper*$^{\Delta 5}$ null mutants, glia fail to respond morphologically to axon injury, and severed axons are not cleared from the central nervous system (CNS)[7]. Thus, both the extension of glial membranes to severed axons and the phagocytosis of degenerating axonal debris require Draper signalling.

We explored whether Shark function in glia is essential for glial responses to axon injury by driving an upstream activating sequence (UAS)-regulated double-stranded RNA interference construct designed to target *shark* (*shark*^{RNAi}) with the glial-specific *repo-Gal4* driver, severing ORN axons, and assaying the recruitment of Draper and green fluorescent protein (GFP)-labelled glial membranes to severed axons. Maxillary palp-derived ORN axons project to 6 of the roughly 50 glomeruli in the antennal lobe. Within hours after maxillary palps have been ablated in control animals, Draper immunoreactivity decorates severed axons projecting to (Fig. 2a, arrow) and within maxillary palp ORN-innervated glomeruli (Fig. 2a, b), and GFP-labelled glial membranes are recruited to these severed axons (Fig. 2c, d). Strikingly, knocking down Shark in glia completely suppressed these events (Fig. 2). We next severed antennal ORN axons; these axons project to about 44 of the 50 antennal lobe glomeruli. Antennal ablation therefore injures nearly all glomeruli in the antennal lobe and results in the majority of antennal lobe glia in control animals upregulating Draper (Fig. 2a, b, open arrowhead) and undergoing hypertrophy (Fig. 2c, d). We found that knocking down

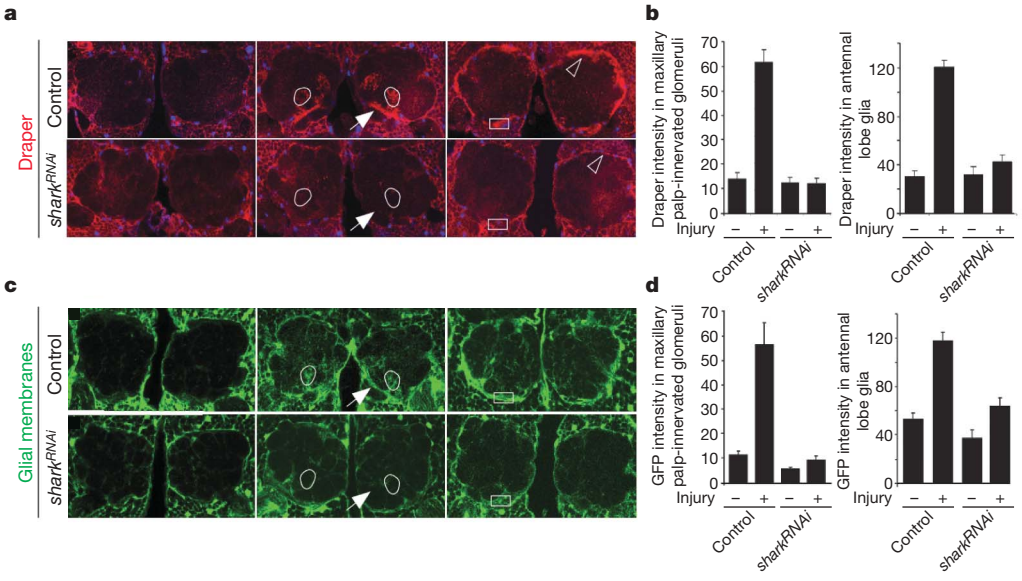


Figure 2 | Shark is required for recruitment of Draper and glial membranes to severed axons. **a**, Control animals (*y w;+/UAS-shark*^{RNAi}) and those with glia-specific knockdown of *shark* (*y w;repo-Gal4/UAS-shark*^{RNAi}) were assayed for expression of Draper (red). Glial nuclei were stained with Repo (blue). Left, uninjured; centre, maxillary palp ablation (day 1); right, antennal ablation (day 1). Outlined, example of a maxillary palp-innervated glomerulus; arrow, nerve containing severed maxillary palp ORN axons; open arrowhead, antennal lobe glial cell; boxes, areas quantified in **b**. **b**, Quantification of data from **a**. Error bars represent s.e.m.; $n \geq 10$. **c**, Glial membranes were labelled in control (*y w;UAS-GFPS65T/+;repo-Gal4/+*) or glial *shark*^{RNAi} animals (*y w;UAS-GFPS65T/+;repo-Gal4/UAS-shark*^{RNAi}) and assayed for morphology before or after injury (panel order as in **a**). Outlined, maxillary palp-innervated glomerulus; arrow, nerve containing severed maxillary palp ORN axons; boxes, areas used to quantify glial hypertrophy in **d**. **d**, Quantification data from **c**. Error bars represent s.e.m.; $n \geq 10$.

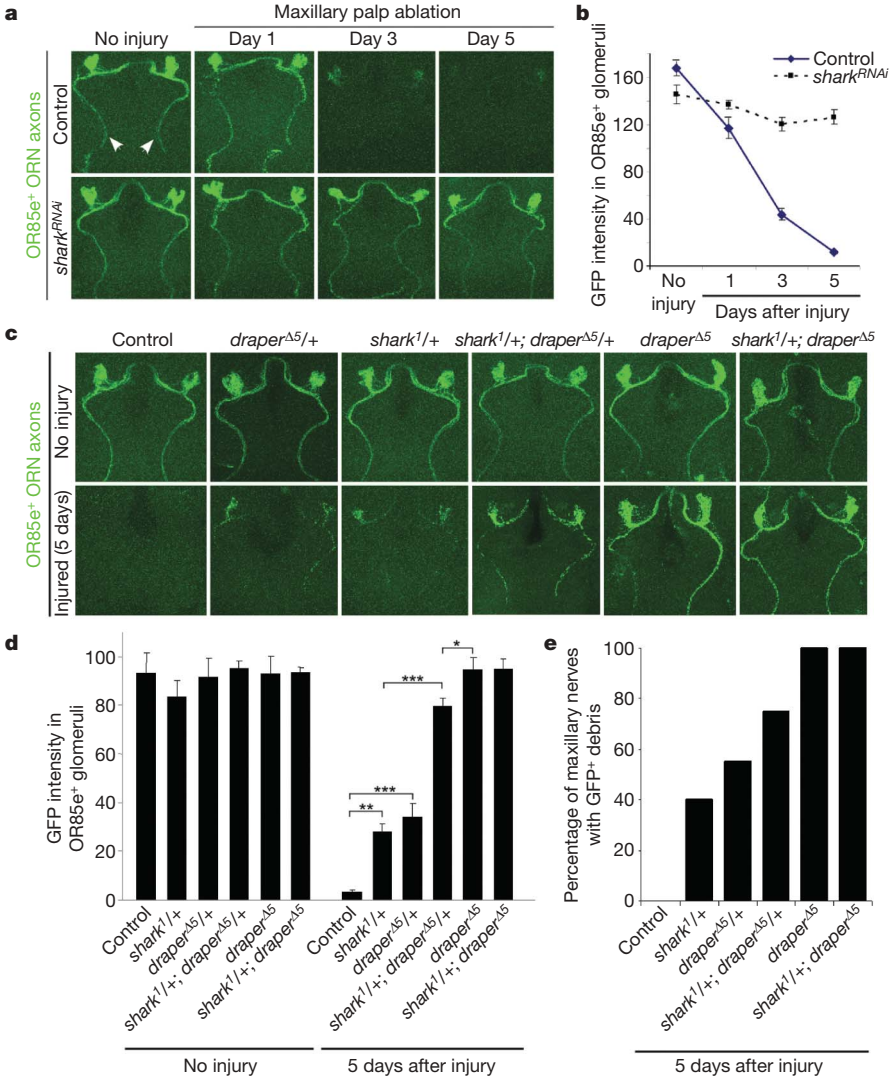


Figure 3 | Shark is required for glial clearance of severed axons from the CNS. a, The axons of OR85e-expressing ORNs were labelled with mCD8::GFP in control (*y w;OR85e-mCD8::GFP/+;repo-Gal4/+*) and glial $shark^{RNAi}$ (*y w;OR85e-mCD8::GFP/+;repo-Gal4/UAS-*$shark^{RNAi}$) animals, maxillary palps were ablated, and the clearance of severed ORN axons from the CNS was assayed with anti-GFP antibody stains (green). Maxillary nerves are indicated (arrowheads). **b**, Quantification of data from **a**. Error bars represent s.e.m.; $n \geq 10$. **c**, OR85e-expressing ORN axons were labelled in control (*y w;OR85e-mCD8::GFP/+*) animals and in $shark^{1}$ or $draper^{\Delta 5}$ null mutant backgrounds, maxillary palps were ablated, and clearance was assayed at 5 days. **d**, Quantification of data from **c**. Error bars represent s.e.m.; $n \geq 10$; *, $P < 0.05$; **, $P < 0.001$; ***, $P < 0.0001$. **e**, Quantification of brain hemispheres containing GFP-labelled ORN axonal debris along the maxillary nerve for genotypes described in **c**.

Shark in glia also blocked this glial response to axon injury (Fig. 2). Thus, Shark is essential for all axon-injury-induced changes in glial morphology and Draper expression.

To determine whether Shark is required for glial phagocytosis of severed axons we labelled a subset of maxillary palp ORN axons with mCD8::GFP, knocked down Shark function in glia, and assayed the clearance of severed axons. In control animals severed GFP-labelled ORN axonal debris was cleared from the CNS within 5 days (Fig. 3a, b). In contrast, glial-specific $shark^{RNAi}$ potently suppressed the clearance of degenerating axons, with severed axons lingering in the CNS for at least 5 days (Fig. 3a, b). We then examined whether mutations in the *shark* gene affected the glial clearance of degenerating axons. The null allele of *shark*, $shark^{1}$, is pupal lethal[25]. We therefore assayed glial responses to axon injury in $shark^{1}$ heterozygous mutants, and tested for dominant genetic interactions between $draper^{\Delta 5}$ and $shark^{1}$. We found that both $draper^{\Delta 5}/+$ and $shark^{1}/+$ animals showed defects in glial phagocytic function: 5 days after injury, significant amounts of axonal debris remained within OR85e-innervated glomeruli (Fig. 3c, d) and in the maxillary nerve (Fig. 3c, e). Moreover, $shark^{1}/+; draper^{\Delta 5}/+$ animals showed a striking suppression of glial clearance of severed axons almost equivalent to that of $draper^{\Delta 5}$ mutants (Fig. 3c–e). Thus, *shark* mutations dominantly suppress the glial clearance of degenerating ORN axons, and this phenotype is strongly enhanced by removing one copy of *draper*. These data, taken together with our $shark^{RNAi}$ data, show that Shark is essential for the clearance of degenerating axons by glia.

Is Shark required for the glial clearance of neuronal cell corpses? In embryonic stage 14–15 control animals we found 24.4 cell corpses per hemisegment (Table 1). In contrast, we found that $shark^{1}$ null

Table 1 | Quantification of cell corpse engulfment defects in *shark* and *draper* mutants

Genotype	Cell corpses per hemisegment*	Hemisegments, *n*
y w;+/+	24.4 ± 0.91	52
y w;$shark^{1}$	43.3 ± 1.88	50
y w;$shark^{1}/+$	25.8 ± 0.80	54
y w;Df(2R)6063	39.0 ± 2.07	51
y w;Df(2R)6063/+	24.6 ± 0.68	51
y w;Df(2R)6063/$shark^{1}$	46.5 ± 1.52	41
y w;$draper^{\Delta 5}$	38.5 ± 1.68	52

* Values are mean±s.e.m

mutants showed a marked increase in CNS cell corpses, with null mutants containing almost twice as many corpses per hemisegment (43.3 cell corpses per hemisegment; Table 1). *shark*^{1}/*Df(2R)6063* mutants accumulated cell corpses at levels similar to those in *shark*^{1}, indicating that this phenotype maps to *shark*. These cell corpse engulfment phenotypes are indistinguishable from that of *draper*$^{\Delta 5}$ mutants (38.5 ± 1.68 cell corpses per hemisegment; Table 1). We conclude that Shark, like Draper, is also essential for the efficient clearance of embryonic neuronal cell corpses by glia.

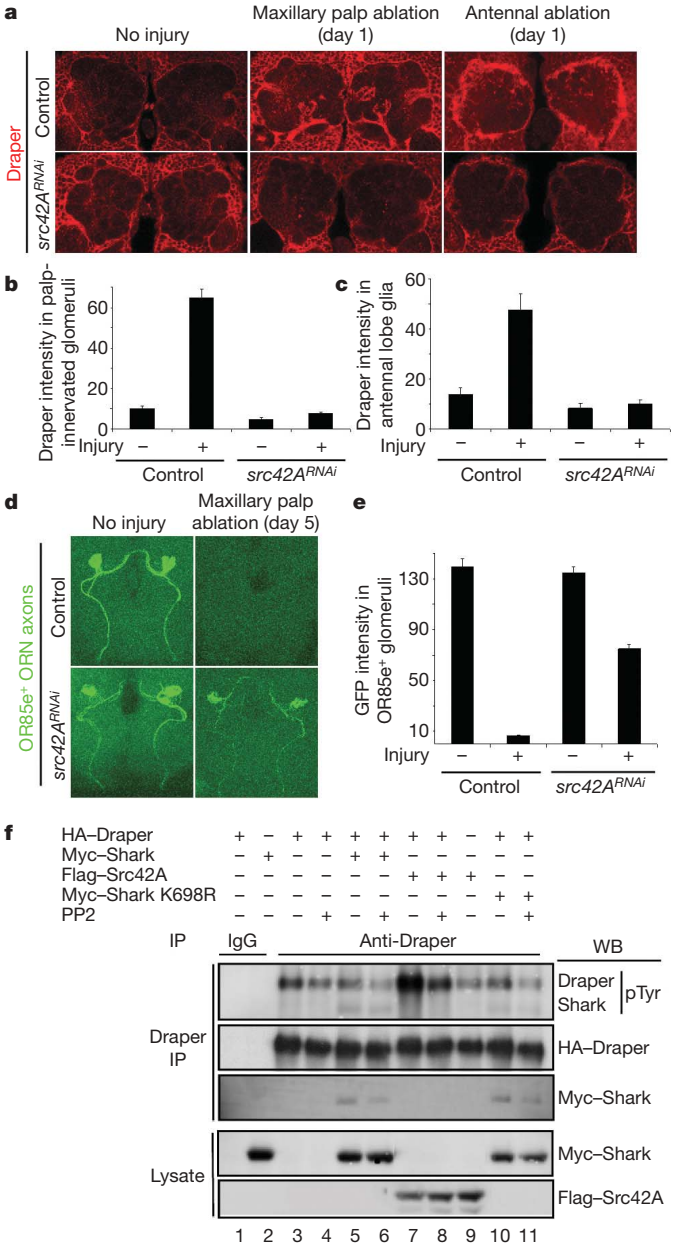


Figure 4 | Src42Aa is required for glial responses to axon injury and modulates Draper phosphorylation status. a, Control animals (*y w;UAS-src42A*^{RNAi}/+, no driver) and those with glia-specific knockdown of *src42A* (*y w;UAS-src42A*^{RNAi}/+;*repo-Gal4*/+) were assayed for injury-induced changes in glial Draper expression and for recruitment of Draper to severed axons (red). **b**, **c**, Quantification of data from **a** for palp-innervated glomeruli (**b**) and antennal lobe glia (**c**). Error bars represent s.e.m.; $n \geq 10$. **d**, The axons of OR85e-expressing ORNs were labelled with mCD8::GFP in control (*y w;OR85e-mCD8::GFP*/+;*repo-Gal4*/+) and glial *src42A*^{RNAi} (*y w;OR85e-mCD8::GFP*/ *UAS-src42A*^{RNAi};*repo-Gal4*/+) animals, maxillary palps were ablated, and the clearance of severed ORN axons from the CNS was assayed with anti-GFP antibody stains (green) 5 days after injury. **e**, Quantification of data from **d**. Error bars represent s.e.m.; $n \geq 10$. **f**, S2 cells were co-transfected with *pMT-HA–Draper*, *pMT-Myc–Shark*, *pMT-Flag–Src42A*, *pMT-Myc–Shark K698R* (kinase-dead) and pMT vector. After transfection and expression, some cells were incubated for 60 min with the SFK inhibitor PP2 (10 μM) before cell lysis. Anti-Draper and IgG control immunoprecipitates (IP) from cells were analysed by SDS–PAGE and western blotted (WB) with antibodies against pTyr, Draper, Myc and Flag.

Because we found that Shark binds Draper only in the presence of an active Src kinase in our two-hybrid assays, we screened *Drosophila* Src kinases for roles in glial phagocytic activity. Interestingly, we found that glia-specific knockdown of Src42A (*src42A*^{RNAi}) potently suppressed glial phagocytic activity: in *src42A*^{RNAi} animals, Draper was not recruited to severed maxillary palp axons (Fig. 4a, b); glial hypertrophy and upregulation of Draper after antennal ablation was blocked (Fig. 4a, c); and GFP-labelled severed axons lingered in the CNS for 5 days (Fig. 4d, e). Knockdown of two other *Drosophila* Src kinases, Btk29A and Src64B, had no effect on the glial phagocytosis of severed axons (Supplementary Fig. 1). Thus, Src42A seems to be essential for all morphological responses of glia to axon injury and for the efficient clearance of degenerating axonal debris from the CNS.

We predicted that Draper phosphorylation status should be sensitive to the SFK inhibitor PP2. Indeed, addition of PP2 to S2 cultures led to a decrease in the phosphorylation of Draper (Fig. 4f, lanes 3 versus 4, 5 versus 6, 7 versus 8, and 10 versus 11) and Draper–Shark association (Fig. 4f, lanes 5 versus 6, and 10 versus 11). Strikingly, co-transfection of Draper and Src42A led to a marked increase in Draper phosphorylation (Fig. 4f, lane 7), which was PP2-sensitive (Fig. 4f, lane 8) and Draper-specific (Fig. 4f, lane 9). Draper–Shark interactions are not dependent on Shark kinase activity because kinase-dead Shark (Shark K698R) associates with Draper (Fig. 4f, lane 10). These data indicate that Src42A may phosphorylate the Draper intracellular domain, thereby increasing the association of Shark with Draper and the activation of downstream glial phagocytic signalling.

We have identified Shark and Src42A as novel components of the Draper pathway. One potential model for Draper–Shark–Src42A interactions is that Shark and Src42A drive the recruitment of Draper to engulfment targets. However, CED-1 has been shown to cluster around cell corpses even in the absence of its intracellular domain[1]. Moreover, Zap-70 and Syk bind phosphorylated ITAM domains in mammalian immunoreceptors when ITAM domains are phosphorylated by Src after ligand-dependent receptor clustering[26–28]. We therefore favour a model in which the engagement of Draper with its ligand (presumably presented by engulfment targets) promotes receptor clustering, tyrosine phosphorylation of Draper by Src42A, association of Shark, and activation of downstream phagocytic signalling events. Our work suggests that Draper is an ancient immunoreceptor in which the extracellular domain is tuned to recognize modified self and the intracellular domain signals through ITAM–Src–Syk-mediated mechanisms. This is the first identification of ITAM–Src–Syk signalling in invertebrates, and it suggests that a pathway similar to Draper–Ced-1 may ultimately have given rise to ITAM-based signalling cascades in mammalian myeloid and lymphoid cells, including those regulated by Fc, B-cell and T-cell receptors.

METHODS SUMMARY

Molecular biology and transgenics. To generate the *shark*^{RNAi} construct we amplified a 522-base-pair *shark* cDNA fragment (nucleotides 577–1,098) from the BDGP complementary DNA *LD41606*, placing a 5′ *Nhe*I site and 3′ *Xba*I site, and cloned this into the *Nhe*I and *Xba*I sites in the *pWiz* UAS-RNAi vector to generate *pshark*^{RNAi}*-3*. We next amplified the same 522-base-pair *shark* fragment but placed a 5′ *Xho*I site and 3′ *Bgl*II site, and cloned this fragment into *pshark*^{RNAi}*-3*, to generate *pUAS-shark*^{RNAi}. The *OR85e-mCD8::GFP* line[29] was a gift from B. Dickson. Transgenic *Drosophila* strains were generated by standard methods. The *UAS-src42A*, *UAS-src64B* and *UAS-btk29A* RNAi lines were obtained from the Vienna *Drosophila* Resource Center. The two-hybrid system used to identify Draper and the methods used to study its interaction in S2 cells have been described previously[23]. The Shark SH2-Ank-SH2-PRB region was

fused to the DNA-binding domain (LexA) in the *pGilda* plasmid without Src (LexA–Shark) or with Src (LexA–Shark/Src). The activation domain (AD) in *pB42AD* was fused to the Draper intracellular domain (AD–Drpr) or to different Y→F Draper mutations (AD–DrprY*nnn*F). Transformed yeast were plated on Leu$^-$ plates for screening Leu-positive colonies. The anti-Draper antiserum used for immunoprecipitations was prepared by immunization of rabbits with GST fused to amino-acid residues 385–594 of the intracellular domain of Draper II.

Nerve injury assays and cell corpse quantification. Wallerian degeneration of ORN axons and quantification of clearance of axonal debris were performed as described previously[7]. Quantification of cell corpses was performed as described previously[2].

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

Received 19 January; accepted 10 March 2008.
Published online 23 April 2008.

1. Zhou, Z., Hartwieg, E. & Horvitz, H. R. CED-1 is a transmembrane receptor that mediates cell corpse engulfment in *C. elegans. Cell* **104,** 43–56 (2001).
2. Freeman, M. R., Delrow, J., Kim, J., Johnson, E. & Doe, C. Q. Unwrapping glial biology. Gcm target genes regulating glial development, diversification, and function. *Neuron* **38,** 567–580 (2003).
3. Chung, S., Gumienny, T. L., Hengartner, M. O. & Driscoll, M. A common set of engulfment genes mediates removal of both apoptotic and necrotic cell corpses in *C. elegans. Nature Cell Biol.* **2,** 931–937 (2000).
4. Awasaki, T. *et al.* Essential role of the apoptotic cell engulfment genes *draper* and *ced-6* in programmed axon pruning during *Drosophila* metamorphosis. *Neuron* **50,** 855–867 (2006).
5. Hoopfer, E. D. *et al.* Wld^s protection distinguishes axon degeneration following injury from naturally occurring developmental pruning. *Neuron* **50,** 883–895 (2006).
6. Williams, D. W., Kondo, S., Krzyzanowska, A., Hiromi, Y. & Truman, J. W. Local caspase activity directs engulfment of dendrites during pruning. *Nature Neurosci.* **9,** 1234–1236 (2006).
7. MacDonald, J. M. *et al.* The *Drosophila* cell corpse engulfment receptor Draper mediates glial clearance of severed axons. *Neuron* **50,** 869–881 (2006).
8. Fodor, S., Jakus, Z. & Mocsai, A. ITAM-based signaling beyond the adaptive immune response. *Immunol. Lett.* **104,** 29–37 (2006).
9. Underhill, D. M. & Goodridge, H. S. The many faces of ITAMs. *Trends Immunol.* **28,** 66–73 (2007).
10. Hengartner, M. O. Programmed cell death in the nematode *C. elegans*. *Recent Prog. Horm. Res.* **54,** 213–222 (1999).
11. Fadok, V. A., Bratton, D. L. & Henson, P. M. Phagocyte receptors for apoptotic cells: recognition, uptake, and consequences. *J. Clin. Invest.* **108,** 957–962 (2001).
12. Henson, P. M., Bratton, D. L. & Fadok, V. A. Apoptotic cell removal. *Curr. Biol.* **11,** R795–R805 (2001).
13. Savill, J., Dransfield, I., Gregory, C. & Haslett, C. A blast from the past: clearance of apoptotic cells regulates immune responses. *Nature Rev. Immunol.* **2,** 965–975 (2002).
14. Su, H. P. *et al.* Interaction of CED-6/GULP, an adapter protein involved in engulfment of apoptotic cells with CED-1 and CD91/low density lipoprotein receptor-related protein (LRP). *J. Biol. Chem.* **277,** 11772–11779 (2002).
15. Liu, Q. A. & Hengartner, M. O. Candidate adaptor protein CED-6 promotes the engulfment of apoptotic cells in *C. elegans. Cell* **93,** 961–972 (1998).
16. Ellis, R. E., Jacobson, D. M. & Horvitz, H. R. Genes required for the engulfment of cell corpses during programmed cell death in *Caenorhabditis elegans. Genetics* **129,** 79–94 (1991).
17. Kinchen, J. M. *et al.* Two pathways converge at CED-10 to mediate actin rearrangement and corpse removal in *C. elegans. Nature* **434,** 93–99 (2005).
18. Yu, X., Odera, S., Chuang, C. H., Lu, N. & Zhou, Z. C. *elegans* Dynamin mediates the signaling of phagocytic receptor CED-1 for the engulfment and degradation of apoptotic cells. *Dev. Cell* **10,** 743–757 (2006).
19. Watts, R. J., Schuldiner, O., Perrino, J., Larsen, C. & Luo, L. Glia engulf degenerating axons during developmental axon pruning. *Curr. Biol.* **14,** 678–684 (2004).
20. Awasaki, T. & Ito, K. Engulfing action of glial cells is required for programmed axon pruning during *Drosophila* metamorphosis. *Curr. Biol.* **14,** 668–677 (2004).
21. Aldskogius, H. & Kozlova, E. N. Central neuron-glial and glial-glial interactions following axon injury. *Prog. Neurobiol.* **55,** 1–26 (1998).
22. Ferrante, A. W. Jr, Reinke, R. & Stanley, E. R. Shark, a Src homology 2, ankyrin repeat, tyrosine kinase, is expressed on the apical surfaces of ectodermal epithelia. *Proc. Natl Acad. Sci. USA* **92,** 1911–1915 (1995).
23. Biswas, R., Stein, D. & Stanley, E. R. *Drosophila* Dok is required for embryonic dorsal closure. *Development* **133,** 217–227 (2006).
24. Berton, G., Mocsai, A. & Lowell, C. A. Src and Syk kinases: key regulators of phagocytic cell activation. *Trends Immunol.* **26,** 208–214 (2005).
25. Fernandez, R. *et al.* The *Drosophila* shark tyrosine kinase is required for embryonic dorsal closure. *Genes Dev.* **14,** 604–614 (2000).
26. Monroe, J. G. ITAM-mediated tonic signalling through pre-BCR and BCR complexes. *Nature Rev. Immunol.* **6,** 283–294 (2006).
27. Pitcher, L. A. & van Oers, N. S. T-cell receptor signal transmission: who gives an ITAM? *Trends Immunol.* **24,** 554–560 (2003).
28. Cox, D. & Greenberg, S. Phagocytic signaling strategies: Fcγ receptor-mediated phagocytosis as a model system. *Semin. Immunol.* **13,** 339–345 (2001).
29. Couto, A., Alenius, M. & Dickson, B. J. Molecular, anatomical, and functional organization of the *Drosophila* olfactory system. *Curr. Biol.* **15,** 1535–1547 (2005).

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

Acknowledgements This work was supported by National Institutes of Health grants (1R01NS053538 to M.R.F., and 1RO1GM55293 and 1RO1CA26504 to E.R.S.), by an Albert Einstein Cancer Center Grant (PO3-13330 to E.R.S.), a Smith Family New Investigator Award (to M.R.F.) from the Smith Family Foundation, and a grant from the Christopher and Dana Reeves Foundation (to M.R.F.). M.R.F. is an Alfred P. Sloan Research Fellow.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.R.F. (marc.freeman@umassmed.edu) or E.R.S. (rstanley@aecom.yu.edu).

METHODS

Yeast two-hybrid screen. The yeast two-hybrid screen was performed in accordance with the manufacturer's instructions (Matchmaker Kit; Clontech). The cDNA encoding residues 10–468 of Shark was subcloned into the pGilda vector at the *Sma*I and *Bam*HI sites to create the pLexA–Shark bait. To create the pLexA–Shark/Src bait encoding activated c-Src kinase driven by the *ADH1* promoter, the cDNA encoding an activated c-Src-kinase domain was excised from the BTM116-Src vector[30,31] as a *Bam*HI fragment and subcloned into pLexA–Shark at the *Sac*I site. A cDNA library from *Drosophila* embryos 0–21 h old cloned into pB42AD (Clontech) was used as the prey. The bait and prey plasmids were co-transformed into the *Saccharomyces cerevisiae* EGY48 (p8op-LacZ) host strain that carries *LEU2* and *lacZ* under the control of LexA operators. Approximately 5×10^6 co-transformants were screened on *Ura*$^-$, *His*$^-$, *Trp*$^-$, *Leu*$^-$, 5-bromo-4-chloro-3-indolyl-β-D-galactoside plates. Seventeen independent transformants activated both the *lacZ* and *LEU2* reporter genes. The pB42AD plasmids were individually isolated from each of the yeast transformants and sequenced. Three different overlapping sequences represented in 15 clones encompassed the Drpr I intracellular domain.

Schneider cell experiments. *Drosophila* Schneider (S2) cells were grown at room temperature (21 °C) in Schneider medium (Invitrogen) containing 10% fetal bovine serum (Invitrogen). cDNAs encoding Shark and Src42A, amino-terminally tagged with Myc and Flag, respectively, and Draper-1 and Draper-1 Y949F, carboxy-terminally tagged with HA, were cloned into the metallothionein promoter pMT/V5-His A vector (Invitrogen). For transfection, cells were plated in six-well plates (Falcon 3046; 2×10^6 cells per well; 2 ml) and cultured for 24 h. Cells were transfected with 2 μg of DNA per construct per well with the calcium phosphate procedure. The total DNA per transfection was kept constant by the addition of pMT vector DNA. After 15 h, induction of expression was initiated by the addition of 10 μl of 100 mM $CuSO_4$ to each well and an additional 24–48 h was allowed for expression. Cells were harvested, lysed by vortex-mixing in 1% Nonidet P40, 10 mM Tris-HCl, 50 mM NaCl, 30 mM $Na_4P_2O_7$, 50 mM NaF, 100 μM Na_3VO_4, 5 μM $ZnCl_2$, 1 mM benzamidine, 10 μg ml^{-1} leupeptin and 10 μg ml^{-1} aprotinin pH 7.2 at 4 °C, and centrifuged at 13,000*g* for 20 min; the supernatants were used for immunoprecipitation[32]. Immunoprecipitates were washed six times with lysis buffer without leupeptin and aprotinin and the immunoprecipitated proteins were eluted with SDS sample buffer and subjected to 7.5% SDS–PAGE, transfer and western blotting.

Induction of axon injury. ORN axon injury was induced by bilateral surgical ablation of third antennal segments and/or maxillary palps. After injury, animals were transferred to fresh food vials and aged at 25 °C.

Immunolabelling and confocal microscopy. Heads were severed from adult animals and fixed (1 × PBS/4% formaldehyde/0.1% Triton X-100) at room temperature for 20 min. Samples were then washed five times in PTx (1 × PBS/0.1% Triton X-100), brains were dissected at room temperature, and tissues were post-fixed (1 × PBS/4% formaldehyde/0.1% Triton X-100) for 15 min. Samples were then washed five times in PTx, blocked in PBTx (1 × PBS/0.1% Triton X-100/1% BSA) for 30 min and incubated overnight with primary antibodies in PBTx. The next day, samples were washed five times in PBTx, incubated with secondary antibodies (in PBTx) for 2 h at room temperature, washed five times with PBTx, and finally mounted in Bio-Rad antifade reagent.

Primary antibodies were used at the following dilutions: mouse anti-Repo, 1:10; rabbit anti-Draper, 1:500; mouse anti-GFP (Invitrogen), 1:500. Secondary antibodies (Jackson Immunoresearch) used for immunofluorescence were used at a dilution of 1:250; samples were mounted in Bio-Rad antifade reagent and viewed on a Zeiss LSM5 Pascal confocal microscope. For all experiments the entire antennal lobe was imaged in 1–2-μm steps and scored for relevant phenotypes. In experiments in which GFP or Draper intensity in different genetic backgrounds was quantified, samples were fixed and stained side by side, and we imaged them at the same time with standardized confocal settings for each series of experiments. Quantification of fluorescence intensity was performed on single *Z* sections in either the centre of the relevant glomerulus, at the same position of the maxillary nerve, or at the same dorsal position at the edge of the antennal lobe, with ImageJ software (http://rsb.info.nih.gov/ij/).

30. Keegan, K. & Cooper, J. A. Use of the two hybrid system to detect the association of the protein-tyrosine-phosphatase, SHPTP2, with another SH2-containing protein, Grb7. *Oncogene* **12**, 1537–1544 (1996).
31. Lioubin, M. N. *et al.* p150Ship, a signal transduction molecule with inositol polyphosphate-5-phosphatase activity. *Genes Dev.* **10**, 1084–1095 (1996).
32. Li, W., Yeung, Y. G. & Stanley, E. R. Tyrosine phosphorylation of a common 57-kDa protein in growth factor- stimulated and -transformed cells. *J. Biol. Chem.* **266**, 6808–6814 (1991).

Vol 453|12 June 2008|doi:10.1038/nature07017

LETTERS

Magnetic resonance imaging of pH *in vivo* using hyperpolarized ^{13}C-labelled bicarbonate

Ferdia A. Gallagher[1,2,3]*, Mikko I. Kettunen[1,2]*, Sam E. Day[1,2]†, De-En Hu[1,2], Jan Henrik Ardenkjær-Larsen[4], René in 't Zandt[5], Pernille R. Jensen[5], Magnus Karlsson[5], Klaes Golman[5], Mathilde H. Lerche[5] & Kevin M. Brindle[1,2]

As alterations in tissue pH underlie many pathological processes, the capability to image tissue pH in the clinic could offer new ways of detecting disease and response to treatment[1]. Dynamic nuclear polarization is an emerging technique for substantially increasing the sensitivity of magnetic resonance imaging experiments[2,3]. Here we show that tissue pH can be imaged *in vivo* from the ratio of the signal intensities of hyperpolarized bicarbonate ($H^{13}CO_3^-$) and $^{13}CO_2$ following intravenous injection of hyperpolarized $H^{13}CO_3^-$. The technique was demonstrated in a mouse tumour model, which showed that the average tumour interstitial pH was significantly lower than the surrounding tissue. Given that bicarbonate is an endogenous molecule that can be infused in relatively high concentrations into patients[4], we propose that this technique could be used clinically to image pathological processes that are associated with alterations in tissue pH, such as cancer, ischaemia and inflammation.

Acid–base balance is tightly controlled in mammalian tissue[5]. The main extracellular buffer is HCO_3^-, which resists changes in pH through interconversion with carbon dioxide (CO_2) in the reaction catalysed by the enzyme carbonic anhydrase[6]. Many pathological states are associated with changes in tissue acid–base balance, including inflammation, ischaemia, renal disease, chronic lung disorders and intrauterine abnormalities[4,7,8]. In tumours, extracellular pH is lower than in normal tissue and can be correlated with prognosis and response to treatment[1,9,10]. The lower pH is due to increased lactic acid production, reduced interstitial fluid buffering and reduced perfusion[1,11]. pH measurements of body fluids can give an indication of systemic acid–base imbalance[4], but there is currently no clinical tool available to provide an image of the spatial distribution of pH in humans.

Probe molecules for measuring pH have been developed for magnetic resonance spectroscopy (MRS) and magnetic resonance imaging (MRI)[12–18] and for positron emission tomography (PET)[19]; for a comprehensive review of the field, see ref. 1. Spectroscopic probe molecules have been developed for ^{1}H, ^{31}P and ^{19}F MRS[16–18,20], but the inherent low sensitivity of spectroscopy makes it difficult to obtain a pH map at high resolution and the pH-dependent chemical shift of these agents is often small, which is problematic for accurate pH estimation at clinical magnetic field strengths. Other ^{1}H MRI methods for measuring pH have exploited pH-dependent magnetization transfer between bulk phase water and either endogenous protein amide groups or exogenous probe molecules[14,15]; another approach uses gadolinium (Gd^{3+}) chelate relaxation agents, which show a pH-dependent hydrogen exchange to the Gd^{3+}-bound water[12,13]; all of these require an accurate determination of the probe concentration, which in practice is difficult to achieve *in vivo*. A ^{11}C PET probe has been developed, which is based on the distribution of neutral and ionized species of the agent, but this makes a number of assumptions that may not be justified[19]. A major concern underlying injection of exogenous probe molecules is that they may be toxic, which could limit their use in the clinic. Furthermore, several of these probes only measure intracellular pH, which is more resistant to pH changes than the extracellular space and is therefore less sensitive to metabolic alterations. In conclusion, although there are several probes suitable for pH imaging *in vivo* in animal models, none of these could be translated to routine clinical imaging. By contrast, we describe here a non-toxic pH probe that exploits the endogenous mammalian physiological buffering system with a technique that is independent of probe concentration and most sensitive in the pathological pH range of interest. We propose that this could be used for high resolution human pH imaging.

Tissue pH could be determined from ^{13}C MRS measurements of endogenous $H^{13}CO_3^-$ and $^{13}CO_2$, as their concentration ratio can be used to calculate pH from the Henderson-Hasselbalch equation, assuming that the pK_a is known *in vivo*[5,21]:

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}([\mathrm{HCO_3^-}]/[\mathrm{CO_2}])$$

Although the plasma bicarbonate concentration in humans is ~24 mM (ref. 4), the low natural abundance of ^{13}C (1.1%) combined with its very low nuclear spin polarization (2.5×10^{-6} polarization at 3 T and 37 °C) means that it is not feasible to measure the relative concentrations of $H^{13}CO_3^-$ and $^{13}CO_2$ without a large increase in sensitivity.

Dynamic nuclear polarization (DNP) has emerged recently as a technique for radically increasing the sensitivity of solution-state ^{13}C MRS (ref. 2). The technique involves cooling a ^{13}C-labelled molecule, doped with a free radical and glass former, to ~1 K and then transferring polarization from electron spins on the radical to the ^{13}C nuclei with microwave irradiation. The sample is then warmed rapidly to body temperature while retaining a high level of nuclear spin polarization; see ref. 2 for further details. Following intravenous injection, the spatial distribution of the hyperpolarized molecule and the metabolites formed from it can be imaged using ^{13}C spectroscopic imaging techniques. To date, these studies have focused on monitoring the metabolism of [1-^{13}C]pyruvate in cardiac[22,23] and tumour tissue[3,24]. Here we demonstrate a new application for DNP for measuring extracellular tissue pH using hyperpolarized ^{13}C-labelled bicarbonate, which may potentially have many applications in fundamental studies of biology and in the clinic.

[1]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [2]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, UK. [3]Department of Radiology, University of Cambridge, Level 5, Box 219, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [4]GE Healthcare, The Grove Centre GC/18, White Lion Road, Amersham HP7 9LL, UK. [5]Imagnia AB, Box 8225, SE-200 41 Malmö, Sweden. †Present address: Laboratory of Functional and Molecular Imaging, National Institute for Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892-1065, USA.
*These authors contributed equally to this work.

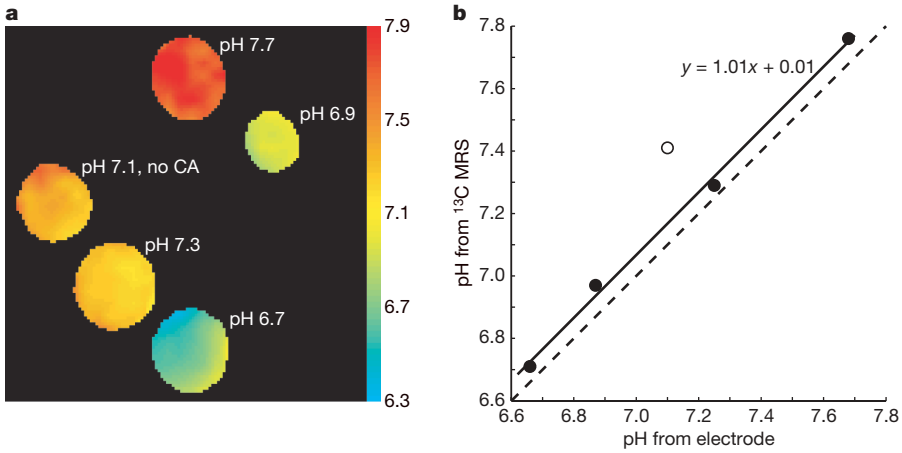


Figure 1 | Imaging of pH *in vitro*. **a**, pH map of tubes (cross-sections) containing buffers at the pH values shown following injection of hyperpolarized bicarbonate. Four contained carbonic anhydrase (CA) and one did not ('no CA'). The pH map, calculated from the relative concentrations of $H^{13}CO_3^-$ and $^{13}CO_2$, is displayed using a colour scale (right). **b**, Correlation of the calculated pHs determined from **a** with the values measured using a pH electrode, for tubes containing CA (filled circles) and the tube without (open circle). A solid linear regression line with its equation is shown for the tubes with CA. The dashed line represents direct proportionality between the two variables.

We have polarized ^{13}C-labelled bicarbonate to levels of up to 16% in the solution state, which represents a 20,000-fold increase above thermal equilibrium (at 9.4 T and 37 °C). Hyperpolarized bicarbonate was injected into tubes containing buffers at different pHs, and ^{13}C chemical shift images were rapidly acquired (Fig. 1a). In the presence of carbonic anhydrase, the estimated pH closely correlated with that measured using a pH meter (average pH difference ± s.d. = 0.07 ± 0.03; $n = 4$; Fig. 1b) but was slightly overestimated in the absence of the enzyme (pH difference: 0.32 ± 0.13; $n = 6$; $P = 0.01$; part of the data is shown in Fig. 1b). The close correlation reflects the rapid establishment of chemical equilibrium and equilibration of polarization between $H^{13}CO_3^-$ and $^{13}CO_2$ catalysed by the enzyme.

Injection of hyperpolarized bicarbonate into mice bearing subcutaneous lymphomas showed that it could also be used to measure pH *in vivo*. Spectra of $^{13}CO_2$ and $H^{13}CO_3^-$ (Fig. 2a) from 5-mm-thick tumour slices gave an average calculated pH of 6.71 ± 0.14 (±s.d.; $n = 12$) compared with a non-slice selective pH, which included a weighted mean of the tumour pH and that of surrounding tissue, of 7.09 ± 0.10 ($n = 8$; $P < 0.01$). Administration of sodium bicarbonate in the drinking water[10] increased the measured tumour pH to 7.02 ± 0.04 ($n = 5$; $P < 0.01$). Similarly, gavage with ammonium chloride (NH_4Cl) over 4 h before the imaging experiment[9] decreased the pH to 6.47 ± 0.11 ($n = 6$; $P < 0.01$). These tumour pHs were compared with the pHs determined using an extracellular ^{31}P MRS pH-probe, namely 3-aminopropylphosphonate (3-APP; Fig. 2b)[1,18,25]. The agreement between the pHs determined using these two methods suggests that the bicarbonate is measuring predominantly the extracellular pH. This probably reflects the higher concentration of extracellular bicarbonate compared to intracellular bicarbonate[26] and the rapid acquisition of spectra immediately after injection of hyperpolarized bicarbonate. The larger variation in pH values following NH_4Cl treatment is presumably due to relatively rapid changes in pH experienced by these animals, rather than the more chronic alteration in bicarbonate-treated animals. This is supported by the larger spread of pH values observed in the 3-APP measurements, which may be due to pH drift over the longer time course of the ^{31}P experiment (26 min), as compared to the rapid determination of pH using hyperpolarized bicarbonate.

Bicarbonate injection could potentially alter the measured pH. However, dilution will result in only a small increase in the high background tissue concentration and ^{31}P MRS measurements with 3-APP showed that this did not significantly change tumour extracellular pH (pH change = 0.02 ± 0.03 units; $n = 4$).

The intracellular pH (pH_i), estimated from the chemical shift of the intracellular inorganic phosphate resonance[11], was 7.44 ± 0.16 in untreated animals ($n = 7$) and was unchanged in mice treated with bicarbonate ($pH_i = 7.47 \pm 0.15$; $n = 4$; $P = 0.75$) or NH_4Cl ($pH_i = 7.36 \pm 0.08$; $n = 6$; $P = 0.30$).

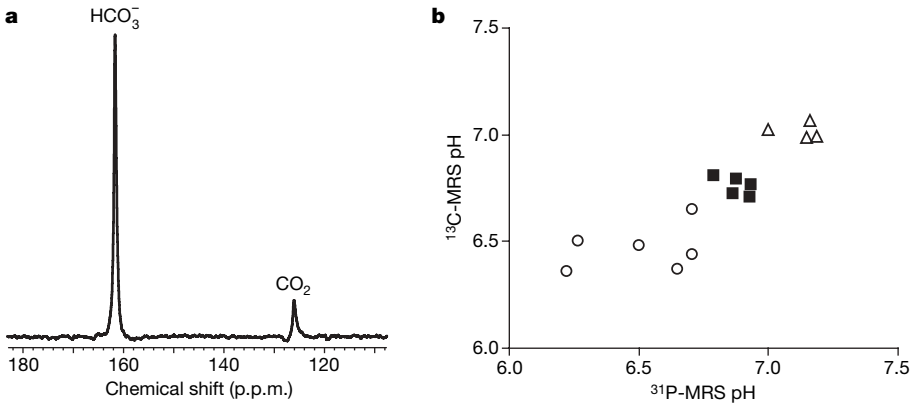


Figure 2 | Measurements of tumour pH *in vivo*. **a**, ^{13}C spectrum from a murine lymphoma (EL4) tumour *in vivo*, acquired ~10 s after the intravenous injection of hyperpolarized $H^{13}CO_3^-$, showing resonances from $H^{13}CO_3^-$ (at 161 p.p.m.) and $^{13}CO_2$ (at 125 p.p.m.). **b**, Correlation of tumour pHs measured using ^{13}C MRS with those determined using ^{31}P MRS and 3-APP in the same animal: filled squares, untreated animals; open triangles, animals treated with oral bicarbonate for 5–7 d; open circles, animals treated by ammonium chloride gavage.

A limitation of DNP is that the signal from the hyperpolarized nuclear spin decreases rapidly owing to decay of the polarization, which is characterized by the spin-lattice relaxation time (T_1), and therefore imaging must be completed within a few T_1 times following injection. Small differences in the rate at which $H^{13}CO_3^-$ and $^{13}CO_2$ lose polarization would lead to the calculated pH varying with time. However, this was not observed over a period of 20 s *in vivo* (data not shown) and can be explained by the fact that $^{13}CO_2$ and $H^{13}CO_3^-$ are interconverted so rapidly[27] that their apparent T_1s become equal. The T_1s measured *in vivo* were 10.1 ± 2.9 s ($n = 9$) for $H^{13}CO_3^-$ and 9.8 ± 2.5 s for $^{13}CO_2$ ($n = 7$; $P = 0.83$). Even in the absence of carbonic anhydrase the exchange is rapid; the rate constant describing flux between bicarbonate and CO_2 was measured *in vitro* using magnetization transfer methods as $\sim 0.1\ s^{-1}$ (data not shown), and this exchange is further increased by the enzyme. Exchange between the two molecules and its dependence on carbonic anhydrase activity was demonstrated *in vivo* by the loss of polarization of $H^{13}CO_3^-$ following selective saturation of the $^{13}CO_2$ resonance (Fig. 3a). In the absence of saturation, the decay of the $H^{13}CO_3^-$ resonance will be dominated by its T_1. However, in the presence of saturation of the $^{13}CO_2$ resonance, the $H^{13}CO_3^-$ signal will decay with a time constant given by $1/(1/T_1 + k)$, where k is the rate constant for the conversion of $H^{13}CO_3^-$ to $^{13}CO_2$ (ref. 28). Inhibition of carbonic anhydrase with acetazolamide[29] abolished this decrease in the $H^{13}CO_3^-$ signal, demonstrating the importance of the enzyme for the exchange of polarization between the two molecules (Fig. 3b). Ignoring inflow of hyperpolarized $H^{13}CO_3^-$ into the tumour and neglecting the fact that there was no saturation pulse during the acquisition periods (103 ms), the decrease in the $H^{13}CO_3^-$ resonance intensity following saturation of the $^{13}CO_2$ resonance (Fig. 3c) indicates that the apparent exchange rate constant is ~ 8 times greater than the apparent rate constant ($1/T_1$) describing loss of polarization.

The spatial distribution of $^{13}CO_2$ and $H^{13}CO_3^-$ in a transverse slice through the animal was imaged using a gradient echo pulse sequence. The image resolution was 16×16 voxels, each of which measured $2 \times 2 \times 6$ mm (Fig. 4). Superimposition of these maps over the corresponding proton image of tissue water showed that the highest concentration of $^{13}CO_2$ was in the tumour. The $H^{13}CO_3^-$ signal was highest in the aorta and there was little difference in signal intensities between muscle and tumour. The pH map, calculated from the ratio of the two signal intensities, demonstrated a low intratumoral pH, as has been observed previously[1].

We have demonstrated a novel non-invasive method for imaging tissue pH *in vivo*. As the technique exploits an endogenous buffer that is already safely administered to humans, we believe that this technique could be transferred to the clinic. Given the range of pathological conditions in which acid–base balance is altered, it may prove to be of diagnostic value not only in oncology but also in the imaging of ischaemia and inflammation.

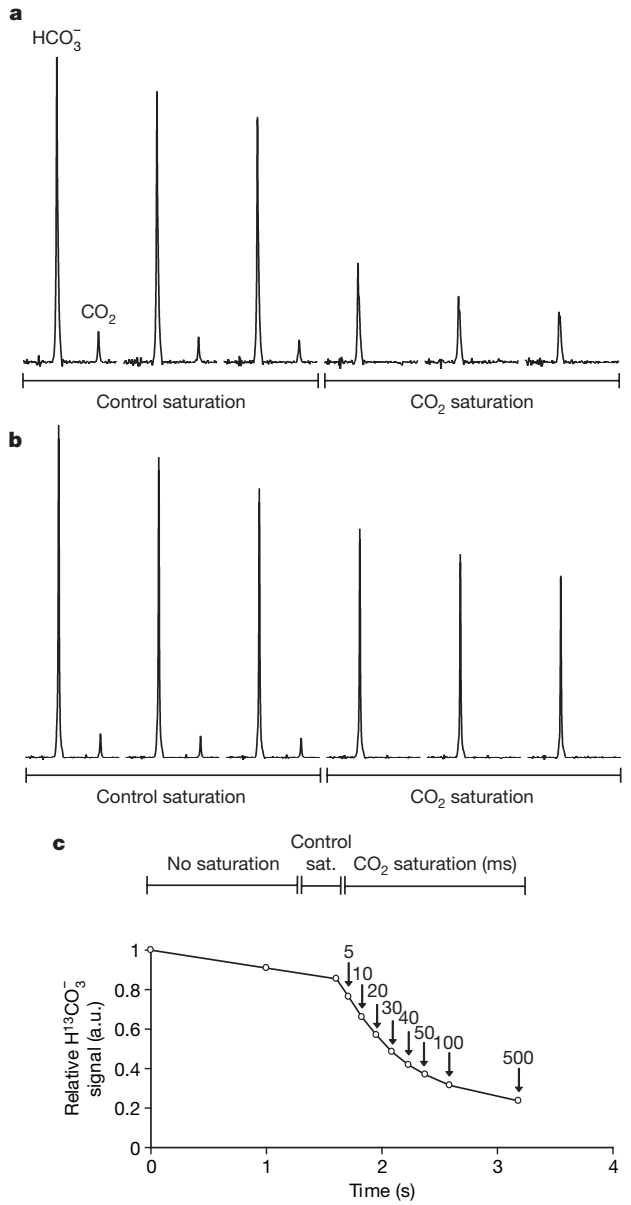


Figure 3 | Demonstration that $H^{13}CO_3$ and $^{13}CO_2$ are in rapid chemical exchange in the reaction catalysed by carbonic anhydrase. **a**, Series of non-slice selective spectra acquired from the tumour using a surface coil with a control saturation (first three spectra from left) and with saturation of the $^{13}CO_2$ resonance (second three spectra). **b**, The same experiment as in **a**, except that the mouse was treated with the carbonic anhydrase inhibitor, acetazolamide. **c**, ^{13}C bicarbonate signal intensity over time with varying duration of the $^{13}CO_2$ resonance saturation pulse. The first three spectra (from left) were acquired without saturation of the $^{13}CO_2$ resonance (a.u., arbitrary units).

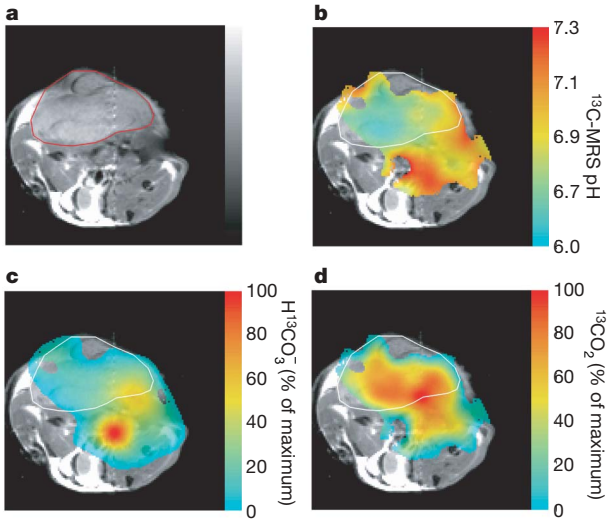


Figure 4 | Imaging of tumour pH *in vivo*. **a**, Transverse proton magnetic resonance image of a mouse with a subcutaneously implanted EL4 tumour (outlined in red). **b**, pH map of the same animal calculated from the ratio of the $H^{13}CO_3^-$ (**c**) and $^{13}CO_2$ (**d**) voxel intensities in ^{13}C chemical shift images acquired ~ 10 s after intravenous injection of ~ 100 mM hyperpolarized $H^{13}CO_3^-$ and assuming a pK_a of 6.17. The spatial distribution of $H^{13}CO_3^-$ (**c**) and $^{13}CO_2$ (**d**) are displayed as voxel intensities relative to their respective maxima. The tumour margin in **b**, **c** and **d** is outlined in white.

METHODS SUMMARY

$CsH^{13}CO_3$ was polarized using a method similar to that described previously for pyruvate[2]. Caesium was removed using an ion-exchange column and the final concentration of bicarbonate was ~100 mM. For the *in vivo* experiments, 200 µl was injected intravenously into mice bearing murine lymphoma tumours; ^{13}C spectra and spectroscopic images[3] were acquired using a surface coil placed over the tumour. Ammonium chloride-treated animals were given two doses of 400 µl of a 1 M solution by gavage 2 h and 4 h before imaging and bicarbonate-treated animals were given 200 mM sodium bicarbonate solution *ad libitum* instead of water for 5–7 days. pH was calculated assuming a pK_a for bicarbonate of 6.17. ^{31}P spectra were acquired from a volume enclosing the whole of the tumour following the intra-peritoneal administration of 3-APP (0.3 ml of 64 mg ml^{-1}) and extracellular tumour pH was calculated as described previously[18].

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

Received 9 January; accepted 18 April 2008.
Published online 28 May 2008.

1. Gillies, R. J., Raghunand, N., Garcia-Martin, M. L. & Gatenby, R. A. pH imaging. A review of pH measurement methods and applications in cancers. *IEEE Eng. Med. Biol. Mag.* **23,** 57–64 (2004).
2. Ardenkjaer-Larsen, J. H. *et al.* Increase in signal-to-noise ratio of > 10,000 times in liquid-state NMR. *Proc. Natl Acad. Sci. USA* **100,** 10158–10163 (2003).
3. Day, S. E. *et al.* Detecting tumor response to treatment using hyperpolarized ^{13}C magnetic resonance imaging and spectroscopy. *Nature Med.* **13,** 1382–1387 (2007).
4. Adrogue, H. J. & Madias, N. E. Management of life-threatening acid-base disorders. First of two parts. *N. Engl. J. Med.* **338,** 26–34 (1998).
5. White, A., Handler, P. & Smith, E. L. *Principles of Biochemistry* (McGraw-Hill Kogakusha, New York, 1968).
6. Meldrum, N. U. & Roughton, F. J. Carbonic anhydrase. Its preparation and properties. *J. Physiol. (Lond.)* **80,** 113–142 (1933).
7. Grinstein, S., Swallow, C. J. & Rotstein, O. D. Regulation of cytoplasmic pH in phagocytic cell function and dysfunction. *Clin. Biochem.* **24,** 241–247 (1991).
8. Hohn-Berlage, M., Okada, Y., Kloiber, O. & Hossmann, K. A. Imaging of brain tissue pH and metabolites. A new approach for the validation of volume-selective NMR spectroscopy. *NMR Biomed.* **2,** 240–245 (1989).
9. Raghunand, N., Mahoney, B., van Sluis, R., Baggett, B. & Gillies, R. J. Acute metabolic alkalosis enhances response of C3H mouse mammary tumors to the weak base mitoxantrone. *Neoplasia* **3,** 227–235 (2001).
10. Raghunand, N. *et al.* Enhancement of chemotherapy by manipulation of tumour pH. *Br. J. Cancer* **80,** 1005–1011 (1999).
11. Stubbs, M. *et al.* An assessment of ^{31}P MRS as a method of measuring pH in rat tumours. *NMR Biomed.* **5,** 351–359 (1992).
12. Beauregard, D. A., Parker, D. & Brindle, K. M. Relaxation-based mapping of tumor pH. *Proc. Int. Soc. Magn. Reson. Med.* 653 (1998).
13. Raghunand, N., Zhang, S., Sherry, A. D. & Gillies, R. J. *In vivo* magnetic resonance imaging of tissue pH using a novel pH-sensitive contrast agent, GdDOTA-4AmP. *Acad. Radiol.* **9** (Suppl 2), S481–S483 (2002).
14. Ward, K. M. & Balaban, R. S. Determination of pH using water protons and chemical exchange dependent saturation transfer (CEST). *Magn. Reson. Med.* **44,** 799–802 (2000).
15. Zhou, J., Payen, J. F., Wilson, D. A., Traystman, R. J. & van Zijl, P. C. Using the amide proton signals of intracellular proteins and peptides to detect pH effects in MRI. *Nature Med.* **9,** 1085–1090 (2003).
16. Mason, R. P. Transmembrane pH gradients *in vivo*: Measurements using fluorinated vitamin B6 derivatives. *Curr. Med. Chem.* **6,** 481–499 (1999).
17. van Sluis, R. *et al. In vivo* imaging of extracellular pH using ^{1}H MRSI. *Magn. Reson. Med.* **41,** 743–750 (1999).
18. Gillies, R. J., Liu, Z. & Bhujwalla, Z. ^{31}P-MRS measurements of extracellular pH of tumors using 3-aminopropylphosphonate. *Am. J. Physiol.* **267,** C195–C203 (1994).
19. Rottenberg, D. A. *et al. In vivo* measurement of brain tumor pH using [^{11}C]DMO and positron emission tomography. *Ann. Neurol.* **17,** 70–79 (1985).
20. Vermathen, P., Capizzano, A. A. & Maudsley, A. A. Administration and ^{1}H MRS detection of histidine in human brain: Application to *in vivo* pH measurement. *Magn. Reson. Med.* **43,** 665–675 (2000).
21. Stabenau, E. K. & Heming, T. A. Determination of the constants of the Henderson-Hasselbalch equation, αCO_2 and pKa, in sea turtle plasma. *J. Exp. Biol.* **180,** 311–314 (1993).
22. Golman, K. & Petersson, J. S. Metabolic imaging and other applications of hyperpolarized ^{13}C. *Acad. Radiol.* **13,** 932–942 (2006).
23. Merritt, M. E. *et al.* Hyperpolarized ^{13}C allows a direct measure of flux through a single enzyme-catalyzed step by NMR. *Proc. Natl Acad. Sci. USA* **104,** 19773–19777 (2007).
24. Golman, K., Zandt, R. I., Lerche, M., Pehrson, R. & Ardenkjaer-Larsen, J. H. Metabolic imaging by hyperpolarized ^{13}C magnetic resonance imaging for *in vivo* tumor diagnosis. *Cancer Res.* **66,** 10855–10860 (2006).
25. McCoy, C. L. *et al.* The effect of blood flow modification on intra- and extracellular pH measured by ^{31}P magnetic resonance spectroscopy in murine tumours. *Br. J. Cancer* **72,** 905–911 (1995).
26. Swietach, P., Vaughan-Jones, R. D. & Harris, A. L. Regulation of tumor pH and the role of carbonic anhydrase 9. *Cancer Metastasis Rev.* **26,** 299–310 (2007).
27. Hoffman, D. W. & Henkens, R. W. The rates of fast reactions of carbon dioxide and bicarbonate in human erythrocytes measured by carbon-13 NMR. *Biochem. Biophys. Res. Commun.* **143,** 67–73 (1987).
28. Brindle, K. M. NMR methods for measuring enzyme kinetics *in vivo*. *Prog. NMR Spectrosc.* **20,** 257–293 (1988).
29. Maren, T. H. Use of inhibitors in physiological studies of carbonic anhydrase. *Am. J. Physiol.* **232,** F291–F297 (1977).

Acknowledgements We thank R. McLaughlin and T. Witney for comments on the manuscript and S. Vowler for help with statistics. F.A.G. is in receipt of a Cancer Research UK and Royal College of Radiologists (UK) clinical research training fellowship, and S.E.D. a National Institutes of Health-Cambridge studentship. The work was supported by a Cancer Research UK Programme grant (to K.M.B.; C197/A3514). The polarizer and related materials were provided by GE Healthcare.

Author Contributions F.A.G. and S.E.D. operated the polarizer and J.H.A.-L. and K.G. provided advice on its use. M.I.K. conducted the MRI experiments and D.-E.H. prepared the animals. M.I.K. and F.A.G. analysed the data. R.I.Z., P.R.J., M.K. and M.H.L. devised the bicarbonate formulation. F.A.G. and K.M.B. wrote the paper and K.M.B. devised and organized the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.M.B. (kmb@mole.bio.cam.ac.uk).

METHODS

^{13}C-labelled bicarbonate polarization. $CsH^{13}CO_3$ was made by slowly adding $^{13}CO_2$ (Sigma-Aldrich) to an evacuated flask containing 0.36 M CsOH hydrate (Sigma-Aldrich) until the pH reached ~7.3, when the sample was lyophilized. Up to 98% of the bicarbonate was ^{13}C-labelled as determined by ^{13}C MRS using an equimolar concentration of ^{13}C-labelled sodium acetate as an intensity standard.

This $CsH^{13}CO_3$ (0.70 mmol) was dissolved in 63.3 µl of water and 0.54 mmol glycerol (Sigma-Aldrich), 2.0 µmol of a free radical (OX063, GE Healthcare) and 47 nmol of a gadolinium chelate (Gd-3, GE Healthcare) were added. Aliquots (10 µl) were dropped into liquid nitrogen to form pellets, which were placed in the DNP polarizer. Polarization was performed using a microwave source at 93.982 GHz and 100 mW for 2 h. The frozen sample was dissolved using 6 ml of an 80 mM phosphate buffer at pH 7.5 containing 100 mg l^{-1} diaminoethanetetraacetic acid (EDTA) heated to 180 °C and pressurized to 10 bar. For the animal experiments, the dissolved solution was rapidly passed through 3 g of an ion-exchange resin in the sodium form (Chelex 100 Resin, Bio-Rad Laboratories). This exchanged the majority of the caesium for sodium, as was confirmed by ^{133}Cs MRS, without significant loss of polarization. The concentration of $H^{13}CO_3^-$ in the final solution was approximately 100 mM, which was less than the expected 116 mM because 10–20% of the label was lost as $^{13}CO_2$ on dissolution. The final temperature of the solution was ~37 °C.

Phantom experiment. Five 1 ml syringes were filled with 400 µl of 500 mM phosphate buffer adjusted to pHs between 6-8. Each syringe contained 100 mg EDTA l^{-1}. To some of the tubes, 5 µg of carbonic anhydrase was added (2,532 units mg^{-1}; Sigma Aldrich). 300 µl of hyperpolarized ^{13}C-labelled bicarbonate was injected simultaneously into each syringe. Spectroscopic ^{13}C images were acquired as described below. The pH of each syringe was measured using a pH meter immediately after image acquisition.

Tumour implantation. Tumours were grown as described previously for 10 days: volume ~2 cm^3; maximum diameter 1.5 cm (ref. 30). Procedures were carried out under the authority of project and personal licenses issued by the Home Office, UK.

Magnetic resonance spectroscopy of tumours. Mice were anaesthetized by intraperitoneal administration of 10 ml per kg body weight of a 5:4:31 mixture of Hypnorm (VetaPharma), Hypnovel (Roche) and saline. A catheter was inserted into a tail vein and a 1 cm diameter surface coil tuned to ^{13}C (100 MHz) was positioned over the tumour. The entire assembly was placed in a quadrature ^{1}H-tuned volume coil (Varian), in a 9.4 T vertical wide-bore magnet. Transverse ^{1}H images were acquired using a spin-echo pulse sequence (TR 1.5 s; TE 30 ms; field of view 32 mm × 32 mm; data matrix 256 × 256; slice thickness 2 mm; 11 slices). After injection of 0.2 ml hyperpolarized bicarbonate, which took approximately 3 s, a single transient spectrum from a 5 mm thick tumour slice and a single transient spectrum from the entire sensitive volume of the surface coil were collected. We used a pulse-acquire sequence with a slice-selective 600 µs sinc pulse, with a nominal flip angle of 5°.

Apparent spin-lattice relaxation times (T_1) were obtained from a series of spectra acquired every second using non-slice selective pulses. The relaxation time was calculated from the slope of the natural logarithm of signal intensity versus time.

^{31}P spectra (TR 3 s; 412 transients; spectral width 12 kHz; acquisition time 26 min) were acquired from a volume enclosing the whole of the tumour following the intra-peritoneal administration of 3-APP (Sigma-Aldrich) in phosphate buffered saline. The pH of the solution had been adjusted to 7.4 with 5 M NaOH.

Magnetic resonance imaging. A ^{13}C chemical shift image (field-of-view 24 × 24 mm; data matrix 32 × 32) from a 6 mm slice through the phantom was acquired immediately after injection of hyperpolarized ^{13}C-labelled bicarbonate. The imaging sequence was the same as that used for the spectroscopic measurements, with additional phase-encoding gradients preceding signal acquisition. The pH in each voxel was calculated from the relative concentrations of $H^{13}CO_3^-$ and $^{13}CO_2$, using the Henderson-Hasselbalch equation and assuming a pK_a of 6.17.

For the animal experiments, transverse ^{13}C chemical shift images were acquired from a 6 mm thick slice through the tumour mass between 12 and 15 s after injection of hyperpolarized ^{13}C-labelled bicarbonate (TR 20 ms; TE 1.5 ms; field of view 32 × 32 mm; data matrix 16 × 16; spectral width 8 kHz; total acquisition time, 5 s).

Processing of ^{13}C chemical shift images was performed in Matlab (Mathworks). The data were multiplied by a cosine function and zero-filled to 128 points in both spatial directions, line-broadened to 30 Hz and zero-filled to 512 points in the spectral direction before Fourier transformation. Bicarbonate and CO_2 peaks were fitted in the frequency domain and only pixels with a frequency separation between the two peaks of 36 ± 1 p.p.m. were included.

To reduce noise in the phantom experiment, only those pixels where the integrals of both peaks were higher than the mean +2 s.d. of the background noise were included for analysis. Additionally, in the final pH map, only those pixels where the bicarbonate peak intensity was greater than 25% of the maximum signal intensity were displayed. For the animal imaging, only those pixels whose integrals were higher than the mean −1 s.d. of all the integrals over the whole animal were used in calculation of the pH maps.

***In vivo* saturation experiments.** Non-slice selective ^{13}C spectra were acquired as described above. The first three spectra were acquired with a control irradiation at 197 p.p.m. and the second three with the saturating irradiation centred on the $^{13}CO_2$ resonance at 125 p.p.m. (Fig. 3a, b). Saturation was performed using a spectrally selective 1 s pulse with a nominal B1 field of 100 Hz. Acetazolamide-treated animals were injected intraperitoneally with 200 µl of a 4 g l^{-1} solution of the drug (40 mg kg^{-1}; in phosphate buffered saline and 10% dimethyl sulphoxide) 15 min before the imaging experiment. In a separate experiment, designed to demonstrate the rapidity of the carbonic anhydrase-catalysed exchange, the saturation pulse was varied between 5 and 500 ms and the decrease in the $H^{13}CO_3^-$ resonance intensity was measured (Fig. 3c). Initially, two spectra were acquired after a 900 ms delay with no saturation pulse. This was followed by a 500 ms control saturation pulse. Finally, a series of $^{13}CO_2$ resonance-selective saturation pulses of varying lengths (5 ms to 500 ms) were applied.

Statistics. Where performed, Student's *t*-tests and linear regression were calculated using Excel (Microsoft). All errors quoted are standard deviations from the mean.

30. Zhao, M., Beauregard, D. A., Loizou, L., Davletov, B. & Brindle, K. M. Non-invasive detection of apoptosis using magnetic resonance imaging and a targeted contrast agent. *Nature Med.* **7**, 1241–1244 (2001).

LETTERS

Global control of cell-cycle transcription by coupled CDK and network oscillators

David A. Orlando[1,2], Charles Y. Lin[1], Allister Bernard[3], Jean Y. Wang[1], Joshua E. S. Socolar[4], Edwin S. Iversen[5], Alexander J. Hartemink[3] & Steven B. Haase[1]

A significant fraction of the *Saccharomyces cerevisiae* genome is transcribed periodically during the cell division cycle[1,2], indicating that properly timed gene expression is important for regulating cell-cycle events. Genomic analyses of the localization and expression dynamics of transcription factors suggest that a network of sequentially expressed transcription factors could control the temporal programme of transcription during the cell cycle[3]. However, directed studies interrogating small numbers of genes indicate that their periodic transcription is governed by the activity of cyclin-dependent kinases (CDKs)[4]. To determine the extent to which the global cell-cycle transcription programme is controlled by cyclin–CDK complexes, we examined genome-wide transcription dynamics in budding yeast mutant cells that do not express S-phase and mitotic cyclins. Here we show that a significant fraction of periodic genes are aberrantly expressed in the cyclin mutant. Although cells lacking cyclins are blocked at the G1/S border, nearly 70% of periodic genes continued to be expressed periodically and on schedule. Our findings reveal that although CDKs have a function in the regulation of cell-cycle transcription, they are not solely responsible for establishing the global periodic transcription programme. We propose that periodic transcription is an emergent property of a transcription factor network that can function as a cell-cycle oscillator independently of, and in tandem with, the CDK oscillator.

The biochemical oscillator controlling periodic events during the cell cycle is centred on the activity of CDKs (reviewed in ref. 5). The cyclin–CDK oscillator governs the major events of the cell cycle, and in embryonic systems this oscillator functions in the absence of transcription, relying only on maternal stockpiles of messenger RNAs and proteins. CDKs are also thought to act as the central oscillator in somatic cells and yeast, and directed studies suggest that they are important for controlling the temporally ordered programme of transcription (reviewed in refs 4, 6). However, systems-level analyses using high-throughput technologies[1,2,7,8] have suggested alternative models for the regulation of periodic transcription during the yeast cell cycle[1,3,9]. By correlating genome-wide transcription data with global transcription factor binding data, models have been constructed in which periodic transcription is an emergent property of a transcription factor network[1,3,9]. In these networks, transcription factors expressed in one cell-cycle phase bind to the promoters of genes encoding transcription factors that function in a subsequent phase. Thus, the temporal programme of transcription could be controlled by sequential waves of transcription factor expression, even in the absence of extrinsic control by cyclin–CDK complexes.

The validity and relevance of the hypotheses regarding intrinsically oscillatory networks of transcription factors remain uncertain, because for the limited number of periodic genes that have been dissected in detail, periodic transcription was found to be governed by CDKs (reviewed in ref. 4). We therefore sought to determine to what extent CDKs and transcription factor networks contribute to global regulation of the cell-cycle transcription programme. To this end, we investigated the dynamics of genome-wide transcription in budding yeast cells disrupted for all S-phase and mitotic cyclins (*clb1,2,3,4,5,6*). These cyclin-mutant cells are unable to replicate DNA, to separate SPBs, to undergo isotropic bud growth or to complete nuclear division, indicating that they are devoid of functional Clb–CDK complexes[10–12]. So, by conventional cell-cycle measures, *clb1,2,3,4,5,6* cells arrest at the G1/S border. We have shown previously that *clb1,2,3,4,5,6* cells trigger G1 events cyclically[10], including the activation of G1-specific transcription and bud emergence. Nevertheless, if Clb–CDK activities are essential for triggering the transcriptional programme, then periodic expression of S-phase-specific and G2/M-specific genes should not be observed.

We examined global transcription dynamics in synchronized populations of both wild-type cells and cyclin-mutant cells. Synchronous populations of early G1 cells were collected by centrifugal elutriation. Cell aliquots were then harvested at 16-min intervals for 270 min (equivalent to about two cell cycles in the wild type

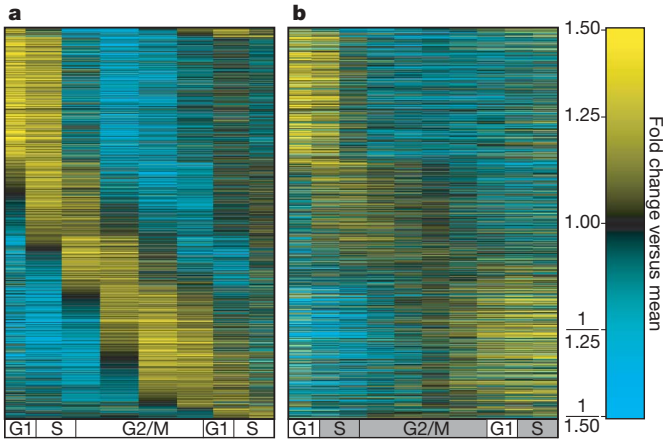


Figure 1 | Dynamics of periodic transcripts in wild-type and cyclin-mutant cells. Heat maps depicting mRNA levels of periodic genes for wild-type (**a**) and cyclin-mutant (**b**) cells. Each row in **a** and **b** represents data for the same gene (Supplementary Table 1). Transcript levels are expressed as a $\log_2$-fold change relative to mean expression. Transcript levels at each point in the time series were mapped onto a cell-cycle timeline (see Methods). The S and G2/M phases of the cyclin-mutant timeline are shaded, indicating that, by conventional definitions, cyclin-mutant cells arrest at the G1/S-phase border.

[1]Department of Biology, [2]Program in Computational Biology and Bioinformatics, [3]Department of Computer Science, [4]Department of Physics, and [5]Department of Statistical Science, Duke University, Durham, North Carolina 27708, USA.

and about 1.5 cell cycles in the cyclin mutant). Transcript levels were measured genome-wide for each time point with the use of Yeast 2.0 oligonucleotide arrays. Results from two independent experiments each for both wild-type and cyclin-mutant cells were highly reproducible, with adjusted r^2 values of 0.995 and 0.989, respectively (Supplementary Fig. 1). All statistical analyses were performed with replicate data sets; however, to facilitate illustration, single data sets were used for all graphical representations.

To identify periodically transcribed genes, we applied a modification of a method developed previously[13] to data acquired from our wild-type cells. We established a set of 1,271 genes that were transcribed periodically (Fig. 1a and Supplementary Table 1). This set of periodic genes shares 510 and 577 genes with those sets previously identified as periodic by Spellman *et al.*[2] and Pramila *et al.*[1], respectively (Supplementary Fig. 2), with 440 consensus periodic genes identified by all three studies (Supplementary Table 2). We then examined the transcriptional dynamics of our set of 1,271 periodic genes in the cyclin mutant (Fig. 1b). The behaviour of many genes changed significantly in the cyclin mutant, supporting previous findings. However, despite the fact that cyclin-mutant cells arrest at the G1/S border, a large fraction of periodic genes in all cell-cycle phases continued to be expressed on schedule (Fig. 1b). Similar cyclin-dependent and cyclin-independent behaviours are also observed in the set of 440 consensus periodic genes (Supplementary Fig. 3).

Using absolute change and Pearson correlation analyses (see Supplementary Information), we determined that 833 of the periodic genes showed changes in expression behaviour in the cyclin mutant and are therefore likely to be directly or indirectly regulated by B-cyclin–CDK.

Our genome-level experiments accurately reproduced previous findings on several well-studied B-cyclin–CDK-regulated genes (Fig. 2). We observed that a subset of late G1 transcripts (SBF-regulated genes such as *CLN2* but not MBF-regulated genes such as *RNR1*) were not fully repressed (Fig. 2a, b) as expected in mitotic cyclin-mutant cells[14,15]. A subset of M/G1 transcripts (including *SIC1* and *NIS1)* are targets of the transcription factors Swi5 and Ace2, which are normally excluded from the nucleus by CDK phosphorylation until late mitosis[16–19]. *SIC1* and *NIS1* were expressed earlier in the cyclin mutant (Fig. 2c, d), presumably because nuclear exclusion of Swi5 and Ace2 is lost in cyclin-mutant cells. The modest degree of shift in the timing of *SIC1* and *NIS1* transcription probably reflects the fact that *SWI5* and *ACE2* transcripts do not accumulate to maximal levels in cyclin-mutant cells as expected for Clb2-cluster genes (including *CDC20*) (Fig. 2e, f)[14,20,21]. Although a significant fraction of periodic genes showed changes in the amplitude of expression (increased or decreased), a statistical analysis of the dynamic range of expression across all periodic genes revealed that most genes in cyclin-mutant cells show only modest changes, if any, in comparison with wild-type cells (Supplementary Fig. 4).

To identify new subsets of co-regulated genes on the basis of transcriptional behaviours observed in both wild-type and cyclin-mutant cells, we employed the affinity propagation algorithm[22], first to cluster genes based on expression in wild-type cells, and then to subcluster genes on the basis of their behaviour in cyclin-mutant cells (Fig. 3 and Supplementary Fig. 5). Of the 833 cyclin-regulated genes, 513 were assigned to 30 discrete clusters showing similar behaviours in wild-type cells (Fig. 3a and Supplementary Fig. 6), and these were then subclustered into 56 clusters on the basis of their transcription profiles in cyclin-mutant cells (Fig. 3b and Supplementary Table 3). Using data from global transcription factor localization studies[23], we identified subsets of transcription factors that may regulate these subclusters with the use of over-representation analyses (Fig. 3 and Supplementary Table 4). On the basis of their association with the

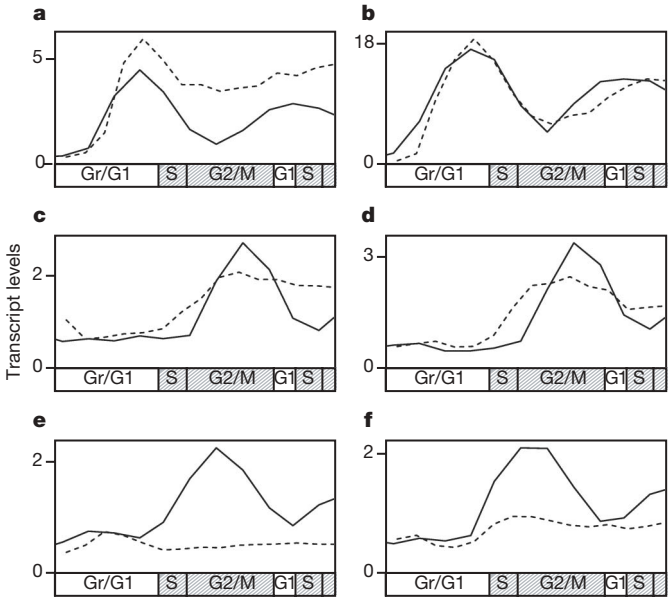


Figure 2 | Transcription dynamics of established cyclin–CDK-regulated genes. Absolute transcript levels (dChip-normalized Affymetrix intensity units/1,000) are shown for the genes *CLN2* (**a**) and *RNR1* (**b**), which are regulated by SBF and MBF, respectively; the Ace2/Swi5-regulated genes *SIC1* (**c**) and *NIS1* (**d**); and the Clb2-cluster genes *CDC20* (**e**) and *ACE2* (**f**). Solid lines, wild-type cells; dashed lines, cyclin-mutant cells.

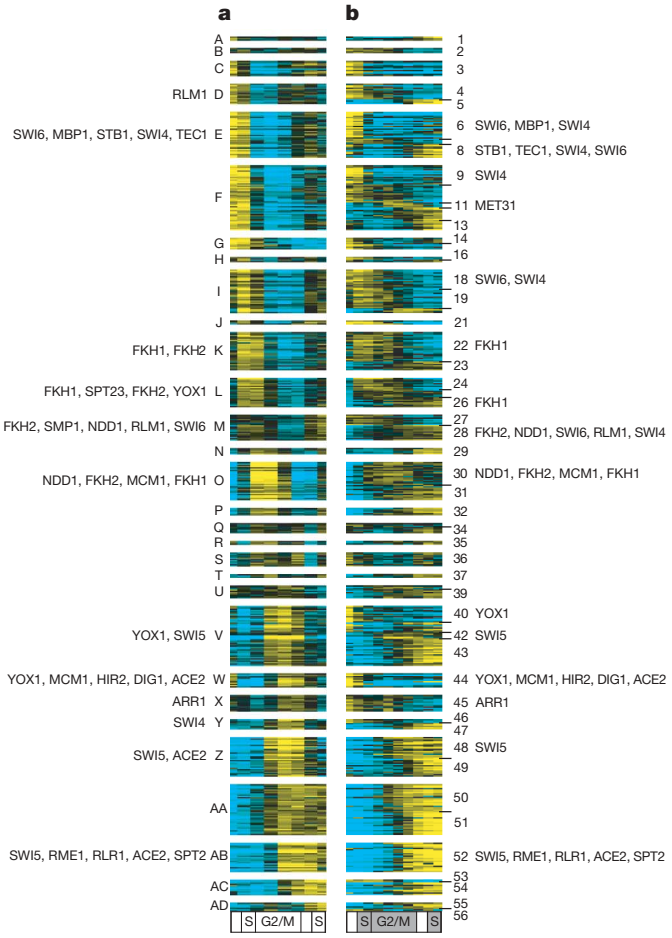


Figure 3 | Genes showing altered behaviours in cyclin-mutant cells. **a**, Clusters of genes with similar expression patterns in wild-type cells. **b**, Subclusters of genes with similarly altered expression patterns in cyclin-mutant cells. Each row in **a** and **b** represents data for the same gene (Supplementary Table 1). Transcript levels are depicted as in Fig. 1. Up to five over-represented transcription factors for each cluster are shown (see Methods and Supplementary Table 4 for complete lists).

promoters of genes in cyclin-regulated subclusters, these factors are likely to be directly or indirectly regulated by cyclins. Consistent with this hypothesis are previous demonstrations that several of these factors are CDK targets[14,15,18,19,24–29]. These findings lay the groundwork for elucidating the full range of mechanisms by which cyclin–CDKs regulate transcription during the cell cycle.

Of the genes identified as periodic in wild-type cells, 882 continued to be expressed on schedule in cyclin-mutant cells despite cell-cycle 'arrest' at the G1/S border (Fig. 4a, b). Some of these genes (450 in total) showed minor changes in transcript behaviour but continued to be expressed at the proper time, as shown above for *ACE2*. Therefore some genes that were cyclin-regulated are also included in the set of genes that maintain periodicity. Nevertheless, a statistical analysis of the dynamic range of expression of these genes in wild-type and cyclin-mutant cells indicates that the amplitude changes for most of these genes are quite modest (see Supplementary Figs 7–9). The finding that nearly 70% of the genes identified as periodic in wild-type cells are still expressed on schedule in cyclin-mutant cells demonstrates the existence of a cyclin–CDK-independent mechanism that regulates temporal transcription dynamics during the cell cycle. This observation is supported by the analysis of the set of 440 consensus periodic genes, the bulk of which maintain periodicity in cyclin-mutant cells (Supplementary Fig. 10).

In principle, a transcription network defined by sequential waves of expression of transcription factors[1,3,9] might function independently of any extrinsic control by CDKs. To determine whether a transcription network could account for cyclin–CDK-independent periodic transcription, we constructed a synchronously updating boolean network model and determined that such a model can indeed explain the periodic expression patterns we observed in cyclin-mutant cells (Fig. 4c). Transcription factors that maintained periodicity in the cyclin mutant were placed on a circularized cell-cycle timeline on the basis of their peak time of transcription in the cyclin mutant. Connections were drawn on the basis of documented physical interactions[23,30] (Supplementary Table 5) between a transcription factor and the promoter region of a gene encoding a transcription factor expressed subsequently (see Supplementary Information). The architecture of the network in cyclin-mutant cells is virtually identical to that in wild-type cells (Supplementary Fig. 11) and is also remarkably similar to models based on wild-type expression data from previous studies[1,3,9].

When the network is endowed with boolean logic functions (Supplementary Table 6a), synchronous updating of the model leads to a cycle that produces successive waves of transcription by progressing through five distinct states before returning to the initial state (Supplementary Fig. 12a, b). Thus, the model functions as an oscillator and produces a correctly sequenced temporal programme of transcription.

To examine the robustness of the network oscillator, we evaluated outcomes when initializing the network from all possible starting states. More than 80% of the 512 starting states entered the oscillatory cycle depicted in Fig. 4c, with the remainder terminating in a steady state in which all genes were transcriptionally inactive (Supplementary Table 6b, c). We also examined whether the oscillations were sensitive to the choice of the boolean logic functions assigned to nodes with multiple inputs, specifically the activating inputs to Cln3 and SFF, and the repressors of SBF and Cln3. For most of the logic functions, the predominant outcome was again the oscillatory cycle depicted in Fig. 4c, but in some cases the model entered two qualitatively similar cycles (Supplementary Fig. 12c, d, and Supplementary Table 6), with the remainder again terminating in a transcriptionally inactive steady state. Several boolean logic functions were found to produce the same cycles (Supplementary Table 6b), so the model cannot precisely determine the true logic of the network connections. Nevertheless, the fact that the model can produce qualitatively similar cycles, and that these cycles can be reached from many initial states, suggests that robust oscillation is an emergent property of the network architecture.

Previous studies proposed that a cyclin–CDK-independent oscillator could trigger some periodic events, including bud emergence[10]. The robust oscillating character of our model indicates that a transcription factor network may function as this cyclin–CDK-independent oscillator. Because cyclin genes are themselves among the periodic genes targeted by this network, and because cyclin–CDKs can, in turn, influence the behaviour of transcription factors in the network, precise cell-cycle control could be achieved by coupling a transcription factor network oscillator with the cyclin–CDK oscillator. The existence of coupled oscillators could explain why the cell cycle is so robust to significant perturbations in gene expression or cyclin–CDK activity.

Our findings also indicate that the properly scheduled expression of genes required for cell-cycle regulated processes such as DNA synthesis and mitosis is not sufficient for triggering these events. The execution of cell-cycle events in wild-type cells is likely to require both properly timed transcription and post-transcriptional modifications mediated by CDKs.

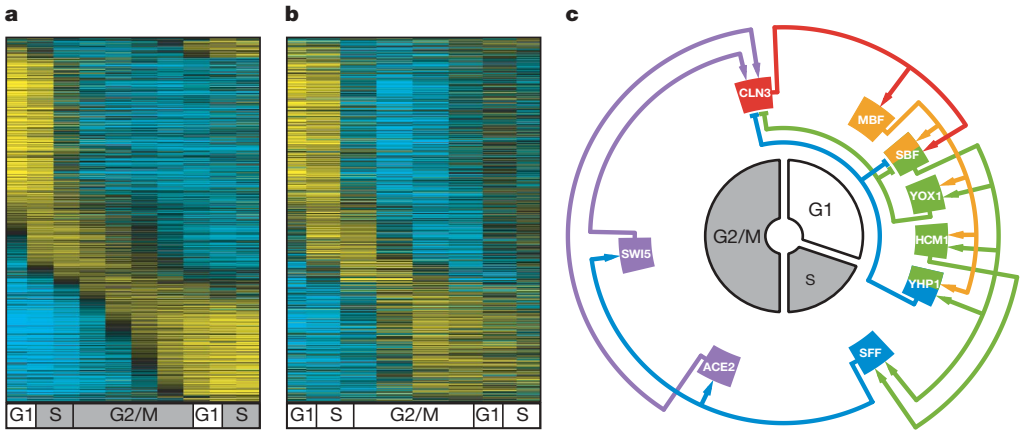


Figure 4 | The periodic transcription programme is largely intact in cyclin-mutant cells that arrest at the G1/S border. **a**, **b**, Genes maintaining periodic expression in cyclin-mutant cells (**a**) show similar dynamics in wild-type cells (**b**). Each row in **a** and **b** represents the same gene (Supplementary Table 1). Transcript levels are depicted as in Fig. 1. **c**, Synchronously updating boolean network model. Transcription factors are arranged on the basis of the time of peak transcript levels in cyclin-mutant cells. Arrows indicate transcription factor/promoter interaction. Activating interactions, outer rings; repressive interactions, inner rings. Colouring indicates activity in one of five successive states; SBF and YHP1 are active in two states (Supplementary Table 6).

METHODS SUMMARY

Strains and cell synchronization. Wild-type and cyclin-mutant strains of *S. cerevisiae* are derivatives of BF264-15Dau, and they were constructed by standard yeast methods. The *clb1,2,3,4,5,6 GAL1–CLB1* mutant strain, along with its growth conditions and synchrony procedures, was described previously[10,11].

RNA isolation and microarray analysis. Total RNA was isolated at intervals (every 16 min for a total of 15 time points) as described previously[10]. mRNA was amplified and fluorescently labelled by GeneChip One-Cycle Target Labelling (Affymetrix). Hybridization to Yeast 2.0 oligonucleotide arrays (Affymetrix) and image collection were performed at the Duke Microarray Core Facility (http://microarray.genome.duke.edu/) in accordance with standard Affymetrix protocols.

Data analysis. A workflow diagram for data analysis is shown in Supplementary Fig. 13.

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

Received 22 November 2007; accepted 31 March 2008.
Published online 7 May 2008.

1. Pramila, T., Wu, W., Miles, S., Noble, W. S. & Breeden, L. L. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.* **20,** 2266–2278 (2006).
2. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297 (1998).
3. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106,** 697–708 (2001).
4. Wittenberg, C. & Reed, S. I. Cell cycle-dependent transcription in yeast: Promoters, transcription factors, and transcriptomes. *Oncogene* **24,** 2746–2755 (2005).
5. Murray, A. W. Recycling the cell cycle: Cyclins revisited. *Cell* **116,** 221–234 (2004).
6. Zwicker, J. & Muller, R. Cell cycle-regulated transcription in mammalian cells. *Prog. Cell Cycle Res.* **1,** 91–99 (1995).
7. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2,** 65–73 (1998).
8. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
9. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
10. Haase, S. B. & Reed, S. I. Evidence that a free-running oscillator drives G1 events in the budding yeast cell cycle. *Nature* **401,** 394–397 (1999).
11. Haase, S. B., Winey, M. & Reed, S. I. Multi-step control of spindle pole body duplication by cyclin-dependent kinase. *Nature Cell Biol.* **3,** 38–42 (2001).
12. Lew, D. J. & Reed, S. I. Morphogenesis in the yeast cell cycle: Regulation by Cdc28 and cyclins. *J. Cell Biol.* **120,** 1305–1320 (1993).
13. de Lichtenberg, U. *et al.* Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* **21,** 1164–1171 (2005).
14. Amon, A., Tyers, M., Futcher, B. & Nasmyth, K. Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins. *Cell* **74,** 993–1007 (1993).
15. Koch, C., Schleiffer, A., Ammerer, G. & Nasmyth, K. Switching transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound transcription factor SBF (Swi4/Swi6) at start, whereas Clb/Cdc28 kinases displace it from the promoter in G2. *Genes Dev.* **10,** 129–141 (1996).
16. Toyn, J. H., Johnson, A. L., Donovan, J. D., Toone, W. M. & Johnston, L. H. The Swi5 transcription factor of *Saccharomyces cerevisiae* has a role in exit from mitosis through induction of the CDK-inhibitor Sic1 in telophase. *Genetics* **145,** 85–96 (1997).
17. Knapp, D., Bhoite, L., Stillman, D. J. & Nasmyth, K. The transcription factor Swi5 regulates expression of the cyclin kinase inhibitor p40SIC1. *Mol. Cell. Biol.* **16,** 5701–5707 (1996).
18. Moll, T., Tebb, G., Surana, U., Robitsch, H. & Nasmyth, K. The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the *S. cerevisiae* transcription factor, SWI5. *Cell* **66,** 743–758 (1991).
19. O'Conallain, C., Doolin, M. T., Taggart, C., Thornton, F. & Butler, G. Regulated nuclear localisation of the yeast transcription factor Ace2p controls expression of chitinase (CTS1) in *Saccharomyces cerevisiae. Mol. Gen. Genet.* **262,** 275–282 (1999).
20. Zhu, G. *et al.* Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* **406,** 90–94 (2000).
21. Kumar, R. *et al.* Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr. Biol.* **10,** 896–906 (2000).
22. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–976 (2007).
23. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104 (2004).
24. Costanzo, M. *et al.* CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast. *Cell* **117,** 899–913 (2004).
25. de Bruin, R. A. *et al.* Constraining G1-specific transcription to late G1 phase: the MBF-associated corepressor Nrm1 acts via negative feedback. *Mol. Cell* **23,** 483–496 (2006).
26. Ho, Y., Costanzo, M., Moore, L., Kobayashi, R. & Andrews, B. J. Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein. *Mol. Cell. Biol.* **19,** 5267–5278 (1999).
27. Pic-Taylor, A., Darieva, Z., Morgan, B. A. & Sharrocks, A. D. Regulation of cell cycle-specific gene expression through cyclin-dependent kinase-mediated phosphorylation of the forkhead transcription factor Fkh2p. *Mol. Cell. Biol.* **24,** 10036–10046 (2004).
28. Sidorova, J. M., Mikesell, G. E. & Breeden, L. L. Cell cycle-regulated phosphorylation of Swi6 controls its nuclear localization. *Mol. Biol. Cell* **6,** 1641–1658 (1995).
29. Ubersax, J. A. *et al.* Targets of the cyclin-dependent kinase Cdk1. *Nature* **425,** 859–864 (2003).
30. Teixeira, M. C. *et al.* The YEASTRACT database: A tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae. Nucleic Acids Res.* **34,** D446–D451 (2006).

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

Acknowledgements We thank D. Lew and L. Simmons Kovacs for discussions and critical reading of the manuscript, and P. Benfey for helpful discussions and support. Financial support was provided by the American Cancer Society (to S.B.H.), the Alfred P. Sloan Foundation (to A.J.H.), the National Science Foundation (to A.J.H. and J.E.S.S.) and the National Institutes of Health (to S.B.H., A.J.H. and J.E.S.S.).

Author Contributions D.A.O., C.Y.L. and S.B.H. designed and performed the experiments. J.Y.W. provided technical expertise. D.A.O., C.Y.L., A.B., E.S.I. and A.J.H. performed the computational analyses, with contributions from J.E.S.S. and S.B.H. to the boolean model. D.A.O. and S.B.H. prepared the manuscript with contributions from C.Y.L., A.B., E.S.I. and A.J.H.

Author Information The microarray data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO series accession number GSE8799. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.B.H. (shaase@duke.edu).

METHODS

Strains and cell synchronization. Yeast strains were grown in rich YEP medium (1% yeast extract, 2% peptone, 0.012% adenine, 0.006% uracil) containing 2% galactose. At 45 min before elutriation, dextrose was added to YEP 2% galactose medium to terminate *CLB1* expression from the *GAL1* promoter. After elutriation, wild-type and *clb1,2,3,4,5,6 GAL1–CLB1* cells were grown in rich YEP 2% dextrose, 1 M sorbitol at 30 °C at a density of 10^7 ml^{-1}. Sorbitol was added to stabilize cells with elongated buds. Aliquots of 50 ml (cell density 10^7 ml^{-1}) were harvested every 8 min for 4 h. Budding index was determined microscopically by counting at least 200 cells for each time point.

Data analysis. CEL files from all 60 oligonucleotide arrays were normalized, and summarized with the dChip method[31] as implemented in the *affy* package (v.1.8.1) within Bioconductor using default parameters. The output of this package is a measure of absolute expression levels for each probe in arbitrary expression units (Fig. 2). Data presented in heat maps and centroid line graphs (Figs 1, 3 and 4, and Supplementary Figs 3, 5, 8, 11, 12 and 15–18) are expressed as $\log_2$-fold change for each gene relative to its mean expression over the interval from the first G1 to the second S phase.

The CLOCCS population synchrony model[32] was used to align expression data temporally from our two wild-type and two cyclin-mutant experiments. In brief, the CLOCCS model allows the alignment of data from multiple synchrony/time-series experiments to a common cell-cycle timeline with budding as a parameter measured on single cells at each time point[32]. Although cyclin-mutant cells undergo arrest at the G1/S border by conventional measures, G1 events, such as bud emergence, are activated periodically with a cycle time similar to that of wild-type cells[10]. Thus, the CLOCCS model can be used to align cycles temporally in wild-type and cyclin-mutant cells. Because the kinetics of synchrony/release experiments can vary, and wild-type and cyclin-mutant cells have marginally different cycle times, alignment is imperative for meaningful comparison of data. We used the CLOCCS parameter estimates to align all four data sets such that the population-level measurements were mapped to a common cell-cycle timeline. The timeline uses standard cell-cycle phases (as determined by measured parameters) and an additional phase (Gr) corresponding to a period of recovery from the initial synchrony procedure that overlaps with early G1 (ref. 31). The recovery period (Gr) was eliminated from most of the data displayed, because the genes expressed in this period tend to be specific to this period and are not expressed again in the next cell cycle in either wild-type or cyclin-mutant cells. The CLOCCS model was designed for wild-type yeast populations but can, with minor modifications, accommodate data from the cyclin mutant (see Supplementary Information). Fits to the CLOCCS model for both the wild-type and cyclin-mutant data sets are shown in Supplementary Fig. 14, and the corresponding parameter estimates are shown in Supplementary Table 7.

A modification of the method described in ref. 13 was used to determine the subset of genes showing periodic transcription (see Supplementary Information for details). The methods used to identify genes with altered transcriptional profiles (Fig. 3), and similar profiles (Fig. 4a, b) in cyclin-mutant cells with respect to wild-type cells are also described in detail in Supplementary Information. Methods for determining over-represented transcription factors in the clusters (Fig. 3) and details regarding the construction of the synchronously updating boolean network model (Fig. 4c) can also be found in Supplementary Information. Two additional analyses similar to that performed in Fig. 3 were performed on consensus periodic genes (Supplemental Table 2) identified as changing expression as well as those identified as maintaining periodicity. Details and results of those analyses can be found in Supplementary Figs 15–19 and Supplementary Tables 8 and 9.

31. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, research0032.1–0032.11, doi:10.1186/gb-2001-2-8-research0032 (2001).
32. Orlando, D. A. *et al.* A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle* **6**, 478–488 (2007).

Vol 453|12 June 2008|doi:10.1038/nature06947

LETTERS

Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions

Lars Guelen[1], Ludo Pagie[1], Emilie Brasset[2], Wouter Meuleman[1,4], Marius B. Faza[1], Wendy Talhout[1], Bert H. Eussen[3], Annelies de Klein[3], Lodewyk Wessels[1,4], Wouter de Laat[2] & Bas van Steensel[1]

The architecture of human chromosomes in interphase nuclei is still largely unknown. Microscopy studies have indicated that specific regions of chromosomes are located in close proximity to the nuclear lamina (NL)[1–3]. This has led to the idea that certain genomic elements may be attached to the NL, which may contribute to the spatial organization of chromosomes inside the nucleus. However, sequences in the human genome that interact with the NL *in vivo* have not been identified. Here we construct a high-resolution map of the interaction sites of the entire genome with NL components in human fibroblasts. This map shows that genome–lamina interactions occur through more than 1,300 sharply defined large domains 0.1–10 megabases in size. These lamina-associated domains (LADs) are typified by low gene-expression levels, indicating that LADs represent a repressive chromatin environment. The borders of LADs are demarcated by the insulator protein CTCF, by promoters that are oriented away from LADs, or by CpG islands, suggesting possible mechanisms of LAD confinement. Taken together, these results demonstrate that the human genome is divided into large, discrete domains that are units of chromosome organization within the nucleus.

A recent genome-wide mapping study in *Drosophila* cells identified hundreds of genes that interact *in vivo* with the NL[4]. These genes tend to be clustered in the linear genome and are mostly transcriptionally repressed. We applied the DamID technique to generate a high-resolution map of NL interactions for the entire human genome. DamID is based on targeted adenine methylation of DNA sequences that interact *in vivo* with a protein of interest[5]. We introduced a chimaeric protein consisting of DNA adenine methyltransferase (Dam) fused to human lamin B1 into cultured human lung fibroblasts. Immunofluorescence microscopy confirmed that this fusion protein was correctly targeted to the NL (Supplementary Fig. 2). We expected that genomic regions in contact with the lamina would be preferentially methylated by the tethered Dam. To identify such regions at high resolution, we selectively amplified adenine-methylated DNA from a pool of about 10^6 cells and initially hybridized it to a high-density microarray probing all of human chromosome 4 (HSA4) with a median probe spacing of about 200 base pairs. Adenine-methylated DNA from cells expressing unfused Dam was processed in parallel and served as a reference[5].

The resulting interaction map shows a distinct pattern: large chromosomal domains of consistently high levels of lamin B1 association alternate with regions of very low levels (Fig. 1a). The lamin B1-associated domains vary in size and are frequently 1 megabase (Mb) in size or larger (Fig. 1b). By repeating the experiments with a set of eight microarrays that interrogate the entire human genome with a median probe spacing of about 750 base pairs, we found this domain pattern to exist on all chromosomes (Supplementary Fig. 3). The transitions between regions of low and high levels are generally very sharp (often occurring over less than 10 kilobases (kb)), a feature that was not observed in the *Drosophila* study[4] because of insufficient mapping resolution.

Six validation analyses indicate that the lamin B1 association maps do indeed reflect interactions of the genome with the NL. First, a DamID profile of emerin, which is a lamina-interacting integral component of the inner nuclear membrane[6] (Supplementary Fig. 2), is nearly identical to that of lamin B1 (Fig. 1c). Minor differences between the profiles of the two proteins may be due to lamin-independent interactions of emerin[7] or vice versa. Second, the fraction of lamin-B1-positive probes per chromosome correlates well with previously reported radial positioning of chromosomes in fibroblast nuclei as determined by three-dimensional fluorescence *in situ* hybridization (3D-FISH) microscopy[8] (Fig. 2a). We also found that HSA18 has a much higher density of lamin B1 interactions than

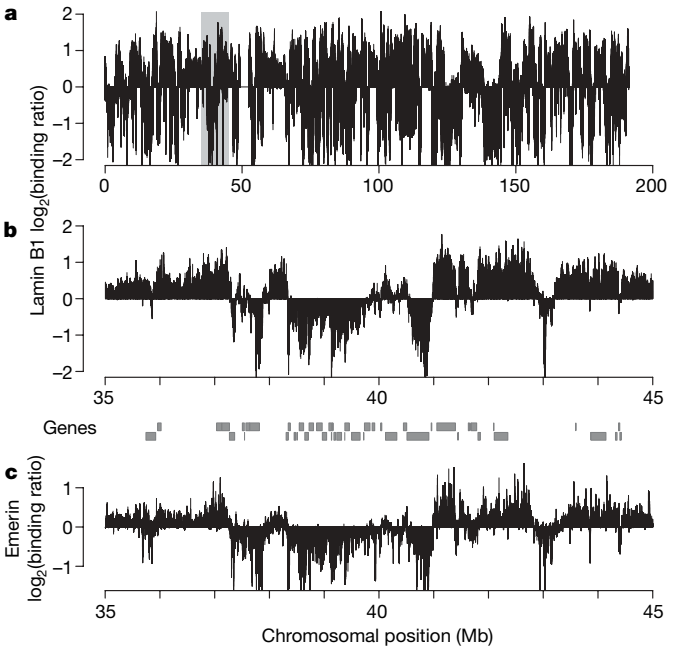


Figure 1 | Mapping of genome–NL interactions in human fibroblasts. **a**, DamID map of lamin B1 interactions for the whole of chromosome 4. **b**, Detailed view of a 10-Mb region (grey box in **a**). **c**, Map of emerin interactions in the same region as in **b**. Genes are indicated by grey rectangles between **b** and **c**.

[1]Division of Molecular Biology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [2]Department of Cell Biology and Genetics and [3]Department of Clinical Genetics, Erasmus Medical Centre, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [4]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.

HSA19, which is consistent with the reported more peripheral location of HSA18 (ref. 9) (Supplementary Fig. 3). Third, for a set of nine chromosomal segments (about 2–14 Mb in size) we observed a strong positive correlation between the fraction of lamin-B1-positive probes and the average intranuclear radial position in fibroblasts[10] (Fig. 2b and Supplementary Fig. 4a). Fourth, we used 3D-FISH with six probe pairs to confirm that loci with high levels of lamin B1 association are generally closer to the nuclear rim than neighbouring loci with low levels of lamin B1 association (Fig. 2c–e and Supplementary Fig. 4b). Fifth, our DamID maps indicate that pericentromeric regions are highly enriched for lamin-B1-positive probes (Supplementary Fig. 4c), which agrees with the preferentially peripheral location of centromeres in human fibroblast nuclei[11,12]. Sixth, we found that the genomic regions that associate with the NL are depleted of Alu elements (Supplementary Figs 4d and 8), which is consistent with previous FISH data[8]. Taken together, these results provide compelling evidence that the DamID maps predominantly reflect the association frequency of probed loci with the NL.

Next, we set out to identify genomic features or chromatin marks that specify lamina-associated domains or their sharp boundaries. A two-step algorithm applied to the genome-wide DamID map (Fig. 3a; see Methods) identified 1,344 sharply defined LADs (Supplementary Data 1). About 74% of array probes with positive lamin B1 log ratios are located in these LADs; the remainder are located in regions with inconsistent binding levels or with fuzzy boundaries that are difficult to define. LADs vary in size from about 0.1 to 10 Mb, with a median size of 553 kb (Fig. 3b). The largest lamina-associated gene cluster found in *Drosophila* is only about 180 kb, and most others are substantially smaller[4]. Thus, LADs in the human genome tend to be much (up to about 50-fold) larger than their fly counterparts.

To characterize the general properties of human LADs and their borders, we aligned all 1,344 LADs by their left or right border and calculated average profiles of various genomic and chromatin parameters across the combined 2,688 borders (Fig. 4 and Supplementary Fig. 5). The average profile of lamin B1 recapitulates the abrupt change in signal at these LAD borders (Fig. 4a). Using this analysis to investigate the positioning of genes relative to LADs, we found that the mean gene density outside LADs is about double that inside (Fig. 4b). This is consistent with earlier FISH data indicating that gene-poor regions tend to be located towards the nuclear periphery[8,13,14]. Previous DamID experiments in *Drosophila* also indicated that lamina-associated genes are separated by large intergenic regions[4]. The LAD borders mark a clear transition in average gene density. Inside LADs, gene density is further reduced with increasing distance from the nearest border (Fig. 4b and Supplementary Fig. 5a). This prompted us to investigate gene deserts, which are exceptionally large gene-free regions of hitherto unknown function[15]. Nearly 75% of the gene deserts interact at least partly with lamin B1, and nearly half of all gene deserts do so along more than 90% of their length (Supplementary Fig. 6). We propose that gene deserts may provide structure to interphase chromosomes by association with the NL.

Previous studies have indicated a link between gene repression and proximity to the NL[1–4,16]. Alignment of microarray mRNA expression data revealed that genes in LADs are generally 5–10-fold less active than genes outside LADs (Fig. 4c). Consistently, levels of RNA polymerase II (PolII) and dimethylated lysine 4 on histone H3 (H3K4me2, a mark typical of active transcription[17]) are very low at most promoters located within LADs (Fig. 4d, e). LADs therefore represent a strongly repressive chromatin environment. Strikingly abrupt transitions in gene expression, PolII and H3K4me2 profiles occur exactly at the LAD borders, emphasizing that LADs are functionally discrete and sharply demarcated chromatin units. However, we note that a small number of genes within LADs are transcribed and occupied by H3K4me2 and PolII (Supplementary Fig. 5b–d), indicating that, for a subset of genes, transcription is compatible with localization within LADs.

We also investigated H3K9me2 and H3K27me3, two histone modifications found in different types of heterochromatin[17]. H3K27me3 is substantially enriched in LADs (Fig. 4f and Supplementary Fig. 5e), which is consistent with the frequent association of this mark with repressed genes[18]. H3K27me3 enrichment occurs predominantly in the outer approximately 200 kb of LADs, immediately adjacent to the LAD borders ($P < 2.2 \times 10^{-16}$; Wilcoxon test). We speculate that H3K27me3 may occupy this position to prevent the spreading of active chromatin from neighbouring regions into some LADs. H3K9me2 is also significantly enriched in LADs, although less prominently (Fig. 4g and Supplementary Fig. 5f).

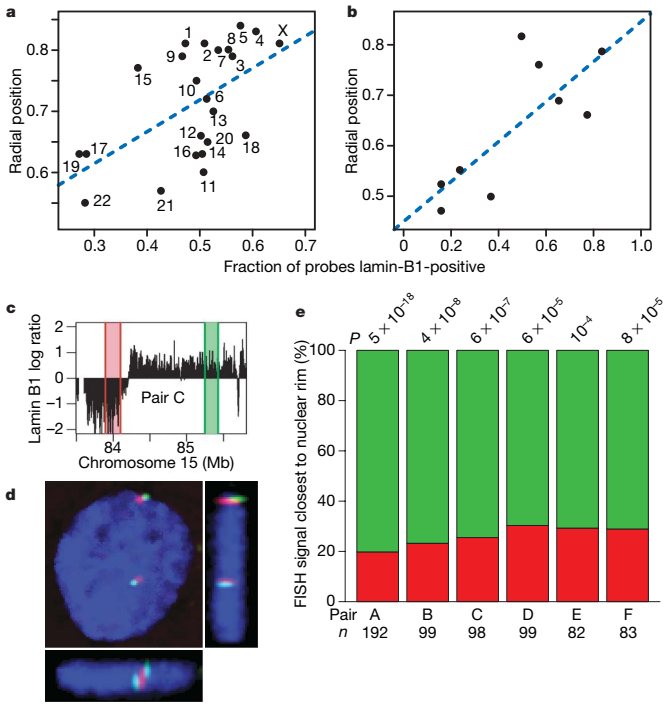


Figure 2 | Lamin B1 interaction maps reflect peripheral nuclear position. **a**, Correlation between the fraction of tiling array probes with positive lamin B1 DamID log ratios (f_{L+}) and the average intranuclear radial position of entire chromosomes (identified by labelled points)[8]; $r = 0.57$, $P = 0.0049$. **b**, Correlation between f_{L+} and the average intranuclear radial position of various chromosomal segments[10] (see also Supplementary Fig. 4a); $r = 0.77$, $P = 0.015$. **c–e**, Two-colour FISH with six probe pairs, each consisting of one probe (red) in a region with low lamin B1 association and one probe (green) in a neighbouring region (about 1 Mb away) with high lamin B1 association. **c**, Locations of a probe pair on HSA15. See Supplementary Fig. 4b for the localizations of all probe pairs. **d**, Projections along each axis of a 3D confocal FISH image stack showing the probe pairs. DNA is stained blue. **e**, Relative frequencies at which each of the two probes is closer to the nuclear rim. *P* values were calculated by a binomial test; *n* is the number of probe pairs scored.

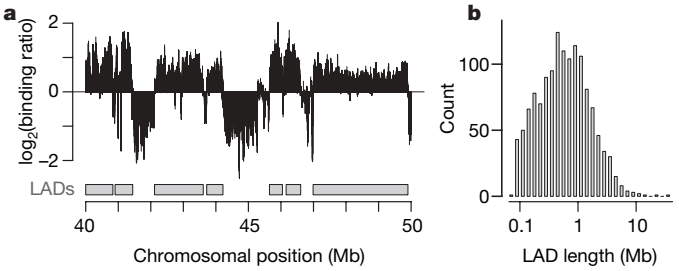


Figure 3 | Definition and size distribution of LADs. **a**, Profile of lamin B1 along a representative chromosome segment (HSA18). Grey rectangles depict LADs along the same segment, as identified by a domain detection algorithm (see Methods). **b**, Histogram of LAD sizes in the entire human genome; median = 553 kb; a total of 1,344 LADs were identified.

The sharp transitions of transcription status and chromatin composition across LAD borders indicated that these borders might contain particular elements that separate the active and inactive chromatin types. A systematic computational search among a wide range of annotated genomic elements, transcription factor motifs and repetitive elements (Supplementary Figs 7 and 8) identified three distinct features that preferentially mark LAD border regions. First, we discovered a striking enrichment of the insulator protein CTCF near LAD borders, peaking at about 5–10 kb outside LADs (Fig. 4h and Supplementary Fig. 5g). This enrichment is observed by analysis of CTCF *in vivo* binding maps[19] and also by the occurrence of the CTCF consensus motif[19] (Supplementary Fig. 5h). Within a range of 10 kb, 301 LADs are flanked by CTCF binding on a single side, and 32 LADs are flanked by CTCF on both sides. CTCF is known to have insulator activity[20] and is therefore a prime candidate to have a function in the demarcation of some LADs. CTCF has been described to interact weakly with A-type lamins[21], but because CTCF enrichment occurs 5–10 kb outside LADs (where binding of lamin B1 is low) it is unlikely that CTCF promotes genome–lamina interactions directly.

Second, we noticed a prominent peak of promoter density just outside LADs, again at about 5–10 kb from the borders (Supplementary Fig. 5i). This enrichment is found only for promoters that drive transcription away from LADs (Fig. 4i, j). Binding motifs for E2F proteins and a few other transcription factors are also somewhat enriched within 10 kb from LADs (Supplementary Fig. 7), which we attribute to the fact that these proteins tend to bind near promoters[22]. Most of the promoters at LAD borders are marked by PolII and H3K4me2 (Supplementary Fig. 5c, d), indicating that they are active. We suggest that active promoters form barriers that prevent the spreading of inactive chromatin from LADs into neighbouring regions. These promoters are presumably preferentially oriented away from LADs because LADs are generally incompatible with transcription.

Third, we discovered a prominent enrichment of CpG islands near LAD borders, just outside LADs (Fig. 4k). Most oriented promoters near LAD borders are accompanied by a CpG island (Fig. 4l). In addition, more than 200 LAD borders are marked by a CpG island in the absence of a promoter (Fig. 4l), indicating that CpG islands might also have an independent function in LAD demarcation. Relatively few LAD borders are marked by the combination of a CTCF site and a promoter (Fig. 4l), indicating that these elements act mostly independently. Oriented promoters, CpG islands and CTCF binding together mark about 30% of all LAD borders within 10 kb, indicating that additional, as yet unknown, demarcating elements might exist.

We also identified many sequence motifs and repeat elements that are enriched in either LADs or inter-LAD regions but not at LAD borders (Supplementary Figs 7 and 8). Among these, we find that Oct-1-binding motifs occur preferentially inside LADs. In combination with the reported localization of Oct-1 protein at the NL[23], this indicates that Oct-1 might help to target LADs to the NL.

The resolution of DamID is about 1 kb (ref. 5), which corresponds to about 50 nm linear distance if chromatin has an extended beads-on-a-string configuration. Genomic loci must therefore be at least transiently within tens of nanometres from the NL for detection by DamID. In this light, it is perhaps surprising that nearly 40% of the human genome consists of LADs (Fig. 3a and data not shown). Compaction of chromatin at the nuclear periphery[2] may allow a relatively large fraction of the genome to be located at the lamina. Furthermore, human fibroblast nuclei in culture are flattened, which

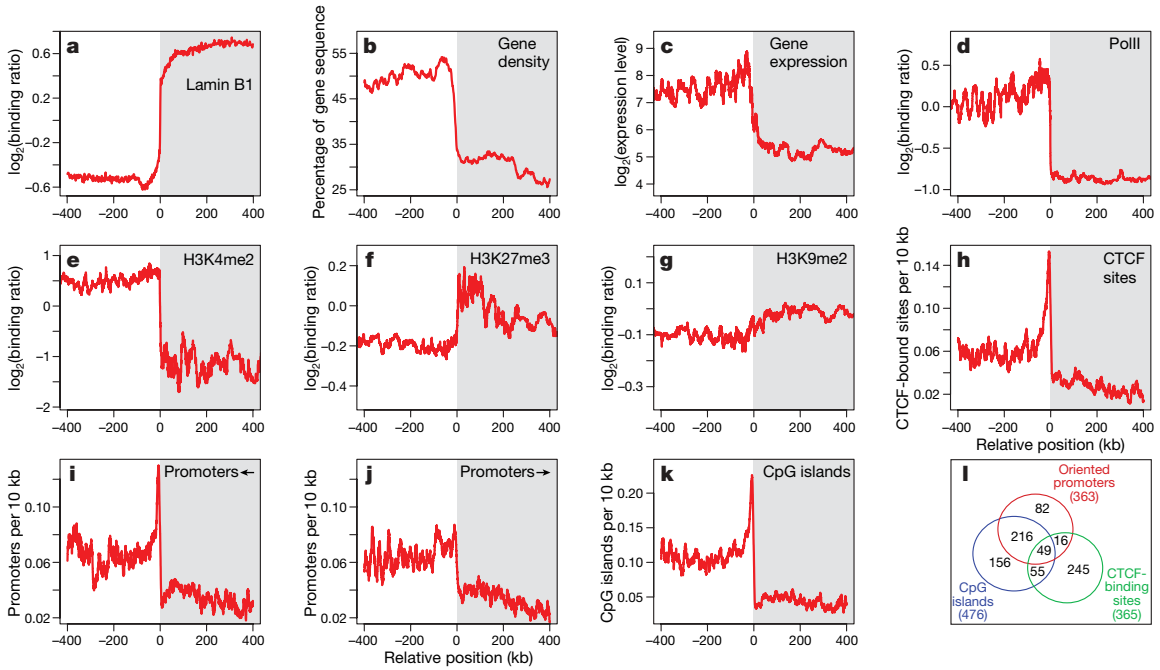


Figure 4 | Profiles of genomic and chromatin features around LAD borders. **a**–**k**, Profiles of aligned LAD border regions (all 2,688 borders; left and mirrored right border regions combined) are shown for lamin B1 binding (**a**); local gene density (**b**); gene expression levels[18] (**c**); promoter occupancy[18,29,30] of PolII (**d**), H3K4me2 (**e**), H3K27me3 (**f**) and H3K9me2 (**g**); density of CTCF-binding sites detected by ChIP[19] (**h**); density of promoters oriented away from (**i**) or towards (**j**) LADs; and density of CpG islands (**k**). To align LAD borders, genome-wide positions of all analysed features were converted to coordinates relative to the nearest border. Grey area and positive coordinates, inside LADs; white area and negative coordinates, outside LADs. Coloured lines show moving-window averages with window sizes of 10 kb (**a**, **b**, **h**–**k**) or moving medians with window sizes equal to 1/50 of the total number of genes or promoters analysed (**c**–**g**). Vertical axes in **c**–**g** are scaled to cover the 10th to 80th centiles of the respective genome-wide value distributions, to normalize for differences in dynamic range. Differences between observed values outside and inside LADs in each of **b**–**k**, and the increased frequencies of CTCF sites, promoters and CpG islands in the first 10 kb just outside of LADs relative to the remainder of inter-LAD regions (**h**–**k**), are all highly significant ($P < 2.2 \times 10^{-16}$; Wilcoxon test). **l**, Venn diagram showing the numbers of LAD borders demarcated within 10 kb by CTCF-binding sites (green; total 365), oriented promoters (red; total 363), CpG islands (blue; total 476) and combinations thereof.

increases the nuclear surface/volume ratio. Most microscopy data (including our FISH experiments in Fig. 2e) indicate that the positioning of genomic loci within the nucleus varies between individual cells and should be regarded as stochastic[24]. It is therefore unlikely that all LADs interact simultaneously with the lamina. Previous FISH studies indicated that these stochastic NL associations can occur throughout most of interphase[25,26].

Thus, we report here that the human genome is divided into large, sharply demarcated domains about 0.1–10 Mb in size, defined by interaction with NL components (Supplementary Fig. 1) and punctuated by specific regulatory elements at their borders. Generally, gene expression levels are low within LADs. Artificial tethering to the NL in human cells can decrease the expression of some, but not all, genes[27,28]. Further experiments are needed to establish how NL associations can regulate human genes in their natural context. The molecular mechanisms by which CTCF sites, CpG islands and oriented promoters may contribute to LAD structure also remain to be explained. The discovery of LADs as architectural units of interphase chromosomes provides an important basis for the understanding of the spatial organization of the human genome inside the cell nucleus.

METHODS SUMMARY

DamID was performed with lentiviral transduction as described[5] in Tig3 human lung fibroblasts. Sample labelling and hybridizations were performed by NimbleGen Inc., on custom-designed oligonucleotide arrays with 385,000 probes each. Probe coordinates are based on genome assembly NCBI36. We limited the analysis of chromatin features around LAD borders to available ChIP-on-chip data from human lung fibroblasts, the cell type used in our DamID experiments. Gene and promoter positions are from Ensembl release 43 (http://www.ensembl.org). Gene deserts were defined as the 3% longest intergenic regions[15]. CpG island positions are from http://genome.ucsc.edu, release hg18.

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

Received 7 February; accepted 28 March 2008.
Published online 7 May 2008.

1. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128,** 787–800 (2007).
2. Akhtar, A. & Gasser, S. M. The nuclear envelope and transcriptional control. *Nature Rev. Genet.* **8,** 507–517 (2007).
3. Taddei, A., Hediger, F., Neumann, F. R. & Gasser, S. M. The function of nuclear architecture: a genetic approach. *Annu. Rev. Genet.* **38,** 305–345 (2004).
4. Pickersgill, H. *et al.* Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nature Genet.* **38,** 1005–1014 (2006).
5. Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of *in vivo* protein–DNA interactions using DamID in mammalian cells. *Nature Protocols* **2,** 1467–1478 (2007).
6. Manilal, S., Nguyen, T. M., Sewry, C. A. & Morris, G. E. The Emery–Dreifuss muscular dystrophy protein, emerin, is a nuclear membrane protein. *Hum. Mol. Genet.* **5,** 801–808 (1996).
7. Bengtsson, L. & Wilson, K. L. Multiple and surprising new functions for emerin, a nuclear membrane protein. *Curr. Opin. Cell Biol.* **16,** 73–79 (2004).
8. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3,** e157 (2005).
9. Croft, J. A. *et al.* Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.* **145,** 1119–1131 (1999).
10. Goetze, S. *et al.* The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Mol. Cell. Biol.* **27,** 4475–4487 (2007).
11. Gilchrist, S., Gilbert, N., Perry, P. & Bickmore, W. A. Nuclear organization of centromeric domains is not perturbed by inhibition of histone deacetylases. *Chromosome Res.* **12,** 505–516 (2004).
12. Wiblin, A. E., Cui, W., Clark, A. J. & Bickmore, W. A. Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *J. Cell Sci.* **118,** 3861–3868 (2005).
13. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10,** 211–219 (2001).
14. Murmann, A. E. *et al.* Local gene density predicts the spatial position of genetic loci in the interphase nucleus. *Exp. Cell Res.* **311,** 14–26 (2005).
15. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145 (2005).
16. Lanctot, C. *et al.* Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Rev. Genet.* **8,** 104–115 (2007).
17. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447,** 407–412 (2007).
18. Bracken, A. P. *et al.* Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* **20,** 1123–1136 (2006).
19. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231–1245 (2007).
20. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev. Genet.* **7,** 703–713 (2006).
21. Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell* **13,** 291–298 (2004).
22. Xu, X. *et al.* A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.* **17,** 1550–1561 (2007).
23. Imai, S. *et al.* Dissociation of Oct-1 from the nuclear peripheral structure induces the cellular aging-associated collagenase gene expression. *Mol. Biol. Cell* **8,** 2407–2419 (1997).
24. Parada, L. A., Roix, J. J. & Misteli, T. An uncertainty principle in chromosome positioning. *Trends Cell Biol.* **13,** 393–396 (2003).
25. Bridger, J. M., Boyle, S., Kill, I. R. & Bickmore, W. A. Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr. Biol.* **10,** 149–152 (2000).
26. Masny, P. S. *et al.* Localization of 4q35.2 to the nuclear periphery: is FSHD a nuclear envelope disease? *Hum. Mol. Genet.* **13,** 1857–1871 (2004).
27. Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452,** 243–247 (2008).
28. Kumaran, R. I. & Spector, D. L. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *J. Cell Biol.* **180,** 51–65 (2008).
29. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.* **39,** 457–466 (2007).
30. O'Geen, H. *et al.* Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* **3,** e89 (2007).

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

Acknowledgements We thank R. Versteeg, M. Indemans, R. van Driel and O. Giromus for help with FISH; L. Oomen for help with confocal microscopy; A. Bracken and K. Blahnik for data sets; the NKI microarray facility for technical support; E. de Wit for preliminary data analysis; and M. Fornerod, F. van Leeuwen, H.J. Bussemaker, M. van Lohuizen and members of the B.v.S. laboratory for helpful suggestions. This study was supported by the Netherlands Cancer Society, the Netherlands Genomics Initiative, and an European Young Investigator Award to B.v.S. E.B. was supported by NWO Rubicon.

Author Contributions L.G. performed DamID and related experiments; M.B.F. and L.G. performed immunofluorescence microscopy studies; E.B. conducted FISH experiments in the W.d.L. laboratory, with technical support from L.G., B.H.E., A.d.K. and W.T.; L.P. designed microarrays; L.P. and B.v.S. performed data analysis together with L.G.; W.M. and L.W. developed motif and repeat search algorithms. L.G. and B.v.S. conceived and designed the study and wrote the manuscript.

Author Information Microarray probe coordinates and data have been submitted to the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE8854. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.v.S. (b.v.steensel@nki.nl).

METHODS

Cell culture. Tig3ET normal human embryonic lung fibroblasts were maintained at 37 °C in 5% CO_2 in DMEM medium containing 4.5 g l^{-1} glucose, Glutamax-l-pyruvate (Gibco), 10% v/v FCS (Greiner Bio-One) and 1% v/v penicillin/streptomycin (Gibco).

DamID. DamID of lamin B1 and emerin was performed in Tig3ET cells by using lentiviral DamID vectors[5]. Because the DamID protocol involves cleavage of GATC motifs by DpnI/DpnII restriction endonucleases, all array oligonucleotide probes were designed to lack GATC motifs.

Microarray data analysis. Microarray data analysis was performed with R (http://www.r-project.org). Full-genome and HSA4 DamID data represent the average of two independent biological replicates. Separate arrays were first loess normalized, after $\log_2$ transformation. The 2×8 full-genome arrays were median and scale normalized with a set of 6,979 common probes. Subsequently, replicate arrays were averaged, and the resulting data were normalized to the genome-wide median.

For the definition of LADs, the full-genome lamin B1 DamID data set was binarized by setting tiling array probes with positive DamID log ratios to 1 and otherwise to −1. Next, a two-step algorithm was used to identify LADs. First, sharp transitions were identified with a sliding edge filter, which calculates the difference in average binary values in two windows of 99 neighbouring probes immediately left and right of a queried probe. The cutoff for this difference was chosen such that the number of edges detected in randomly permuted data sets was less than 5% of the number of edges detected in the original lamin B1 data set. Second, pairs of adjacent 'left' and 'right' edges were identified that together enclosed a region of arbitrary size with at least 70% of the enclosed probes reporting a positive $\log_2$ ratio. A total of 1,344 regions fulfilled these criteria and were termed LADs. In 20 randomly permuted data sets, fewer than 13 domains were identified by the same criteria. We note that there are also lamin-B1-positive domains flanked by one or two gradual or irregular transitions. Because it is difficult to define the borders of such domains precisely, we have not analysed these 'fuzzy' domains here.

We limited the analysis of chromatin features around LAD borders to available mapping data from human lung fibroblasts, the cell type used in our DamID experiments. For promoter ChIP data[18,29,30] the $\log_2$ ratios of tiling array probes belonging to the same promoter were averaged. Gene expression data and H3K27me2 ChIP data are from Tig3 lung fibroblasts[18]. H3K9me2 ChIP data are from human fetal lung fibroblasts[30]. CTCF target sites and the derived consensus motif locations were determined in IMR90 human lung fibroblasts[19]. H3K9me2 and PolII ChIP data are from WI38 human lung fibroblasts[29]. Genomic coordinates were converted to NCBI36 coordinates by using the UCSC LiftOver utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Gene and promoter positions are from Ensembl release 43 (http://www.ensembl.org). Gene deserts were defined as the 3% longest intergenic regions[15]. CpG island positions are from http://genome.ucsc.edu, release hg18.

FISH radial positioning data are from human skin fibroblasts[8] and human fibroblasts of unspecified origin[10]. Chromosome Y was not included in the analysis in Fig. 2a because our tiling array is limited to non-repetitive regions, which comprise only about 30% of this chromosome.

Motif and repeat searches. For motif searches, the TFBS Perl module[31] was used to scan sequences for the presence of motifs of all 588 vertebrate transcription-factor-binding sites included in the TRANSFAC database (version 11.1)[32]. The CTCF position weight matrix[19] was added to this set. A scanning threshold of 90% was used. Both the sense and antisense strands of sequences were considered.

Four categories of sequences, selected according to their location relative to LADs, were used for motif scanning. Category **Fo** contains sequences of 10 kb located just outside LADs, immediately flanking LAD borders. **Fi** contains sequences of 10 kb located just inside LADs, immediately flanking LAD borders. There are 1,344 LADs, and hence 2,688 **Fi** and 2,688 **Fo** sequences. Categories **I** and **O** each contain 2,688 sequences of length 10 kb that were randomly selected from LADs or inter-LAD regions, respectively, with the additional constraint that they should be more than 10 kb from any LAD border (that is, they do not overlap with the **Fo** and **Fi** sets).

For every vertebrate TRANSFAC motif the frequencies were determined for all sequences in all four categories. We then first identified motifs that were enriched in a single category only. For this, motifs were grouped according to the category with the highest number of motif occurrences, resulting in four groups (A–D in Supplementary Fig. 7). For each motif, the top-scoring category was tested against the second-best category, using a one-tailed Wilcoxon test. Benjamini–Hochberg multiple test correction was applied for each group separately, and results with $P < 0.05$ after correction were considered to be significant.

The remaining motifs are not significantly enriched in a single category. It is, however, of potential biological interest to identify motifs that are enriched in the combined **Fi** and **I** categories (that is, enriched inside LADs irrespective of border proximity) or in the combined **Fo** and **O** categories (that is, enriched outside LADs irrespective of border proximity). Motifs that were not significantly enriched in any of the four groups were therefore retested for pooled **Fi** and **I** sequences versus pooled **Fo** and **O** sequences as described above. Significantly enriched motifs are listed as groups E and F in Supplementary Fig. 7.

The same procedure was followed for all different repeat elements as listed in the RepeatMasker table from http://genome.ucsc.edu (hg18). Here, only repeats that are located completely within a 10-kb sequence were counted.

Immunofluorescence microscopy. To verify the correct subcellular localization of Dam-fusion proteins, they were expressed from the lentiviral LTR promoter by direct transfection of respective lentiviral transfer plasmids[5] using Lipofectamine 2000 transfection reagent (Invitrogen). Immunofluorescent staining was performed as described[33]. Epitope-tagged Dam-fusion proteins were detected with mouse anti-V5 antibody (Invitrogen) followed by donkey anti-mouse-FITC (Jackson ImmunoResearch Laboratories). Endogenous lamin B1 and emerin were stained with goat anti-lamin B1 antibody (Santa Cruz) and donkey anti-mouse-FITC (Jackson ImmunoResearch Laboratories), and with rabbit anti-emerin antibody (Abcam) and donkey anti-rabbit-FITC (Jackson ImmunoResearch Laboratories), respectively. DNA was counterstained with ToPro-3 (Invitrogen). Confocal images were made with a Leica TCS NT confocal system.

3D-FISH. Tig3ET cells grown on coverslips to 70% confluence were prepared for 3D-FISH in accordance with standard protocols[34], using bacterial artificial chromosome (BAC) clones as indicated in Supplementary Fig. 4b. The BAC probes were prepared as described[35]. Stacks of 8-bit greyscale optical sections (spaced 250–300 nm) through whole nuclei were collected with a Zeiss Meta 510 confocal microscope system equipped with Plan Apo 63×/1.4 numerical aperture oil-immersion objective lenses. For each optical section, images were collected sequentially for two fluorochromes. For segmentation of the DAPI (4,6-diamidino-2-phenylindole) image to define the nuclear rim, and for segmentation of FISH spots and calculation of their 3D centre-of-mass coordinates, we used ImageJ software with the Sync_Measure_3D plug-in (http://rsb.info.nih.gov/ij/). Calculation of the shortest distance of each FISH spot to the nuclear rim in three dimensions was performed by a custom-made R script.

31. Lenhard, B. & Wasserman, W. W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135–1136 (2002).
32. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
33. van Steensel, B. *et al.* Localization of the glucocorticoid receptor in discrete clusters in the cell nucleus. *J. Cell Sci.* **108**, 3003–3011 (1995).
34. Solovei, I. *et al.* Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Exp. Cell Res.* **276**, 10–23 (2002).
35. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genet.* **38**, 1348–1354 (2006).

CORRIGENDUM

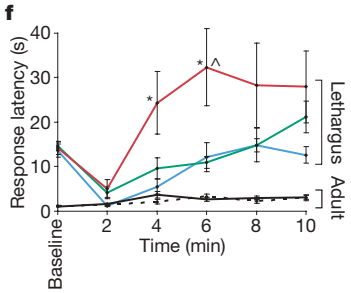
doi:10.1038/nature07044

Lethargus is a *Caenorhabditis elegans* sleep-like state

David M. Raizen, John E. Zimmerman, Matthew H. Maycock, Uyen D. Ta, Young-jai You, Meera V. Sundaram & Allan I. Pack

Nature 451, 569–572 (2008)

In Fig. 3f of this Letter, incorrect error bars were used. In the legend, the descriptions of the lines denoting adult responses were switched. The black solid line denotes non-deprived adults whereas the black dashed line denotes deprived adults. In addition, the *x* axis was inadvertently mislabelled during the production of the figure. The corrected figure is printed below. These changes do not affect conclusions drawn from these data.



naturejobs

JOBS OF THE WEEK

Last week saw the publication of yet another report on troubling trends in US science and engineering research. Sponsored by the American Academy of Arts and Sciences, *ARISE: Advancing Research In Science and Engineering* sounds a familiar refrain: universities and government should provide better support to early-career faculty members, and should encourage high-risk, high-reward research. Specific recommendations include creating multi-year awards for early-career faculty members, strengthening mentoring programmes, reconsidering promotion and tenure policies, and investing more in grant-reviewing officers.

But unlike most reports of this ilk, *ARISE* does not request more government funding. Tom Cech, president of the Howard Hughes Medical Institute and chair of the committee behind the document, says *ARISE* focuses on "modes and mechanisms" of improvement rather than a funding infusion. If the panel wants Congress to take the report seriously, this is a wise move — too often, lawmakers receive recommendations that simply advocate more money as the solution.

And in what Cech calls the committee's "boldest recommendation", the report says that universities should pay a greater proportion of the salaries of faculty members, rather than forcing investigators to rely almost entirely on government funds. This might be heresy to many university administrators, but it would help early-career researchers to avoid the ultra-competitive federal funding pool.

It could also provide an incentive for more practical planning. Consider places such as Florida, where local and state funding have sparked numerous bioscience initiatives. Administrators there expect newly hired young faculty members to support themselves with federal money once their multi-year start-up packages end (see *Nature* **449,** 371; 2007). Given the current trends, this means that many good scientists will struggle when the purse-strings are cut. Universities and institutions would be wise to rethink the extent to which they incorporate salary costs into their initiatives. The *ARISE* report's most heretical idea is one that merits serious consideration.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW: Nils Moeller (4953)

Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

NATURE|Vol 453|12 June 2008

MOVERS

Cato Laurencin, vice-president for health affairs, University of Connecticut Health Center, and dean, University of Connecticut School of Medicine, Farmington

2003–08: Professor and chair of orthopaedic surgery, professor of biomedical engineering and chemical engineering, University of Virginia, Charlottesville
2001–03: Professor of chemical engineering, vice-chairman and clinical professor of orthopaedic surgery, Drexel University, Philadelphia, Pennsylvania

Cato Laurencin says his mother's medical practice and small home laboratory inspired his career as a clinician-scientist. But a dinner sponsored by his high school steered him into combining medicine and engineering: there, New Jersey's Princeton University recruited Laurencin to study chemical engineering alongside his medical school prerequisites.

From Princeton, he went on to Harvard Medical School and, while there, did a joint engineering PhD at the Massachusetts Institute of Technology under famed bioengineer Robert Langer. For his dissertation, Laurencin developed a polymer-based drug-delivery system; these better preserve a molecule's bioactivity after it enters the body than do traditional pills. This led to a drug-delivery system for patients with brain cancer. "Cato's work established how to do Food and Drug Association-relevant work, including safety studies needed to take a new biomaterial into clinical practice," says Langer.

To graduate from Harvard with high honours, however, Laurencin needed a thesis project. He showed that inorganic polymers called polyphosphazenes could control drug delivery better than traditional polymers. "Reaching for that honours degree proved vital to my career — spawning a long-term collaboration, multiple papers and patents, and a long-running National Institutes of Health grant," says Laurencin.

After Harvard, Laurencin spent a year in sports medicine at Cornell University in Ithaca, New York, then joined the Medical College of Pennsylvania's Hahnemann University, which soon merged into Drexel University. There, he started the Center for Advanced Biomaterials and Tissue Engineering and helped usher in a new frontier in orthopaedic surgery — biodegradable polymer materials for fracture repair and tissue engineering.

Laurencin left Philadelphia in 2003 to join the University of Virginia, but carried his sports-medicine work with him. He developed therapies to repair soft tissues such as the anterior cruciate ligament in the knee. Laurencin's newest work uses polymer-based drug-delivery systems and nanotechnology to enhance bone and tissue regeneration.

Having held high-level positions in both medicine and engineering at several universities, Laurencin ultimately decided he could have the greatest clinical impact as dean of the University of Connecticut School of Medicine and as vice-president for health affairs at the university's Health Center. The university's programmes in musculoskeletal medicine and stem-cell technology made it a natural fit for him. ■

Virginia Gewin

NETWORKS & SUPPORT

A route to postdoc diversity

Many programmes aimed at drawing ethnic minorities into science emphasize the early stages of a scientific career. But universities are still struggling to increase faculty diversity because of a lack of candidates. A postdoc programme started almost 30 years ago at the University of North Carolina (UNC) has shown considerable success in bridging that gap. Of the programme's 132 graduates, 113 have gone on to land tenure-track faculty positions — 27 of them currently at UNC.

The philosophy behind the Carolina Postdoctoral Program for Faculty Diversity, based at UNC's Chapel Hill campus, is to groom newly minted PhD students for academia. And the programme has not only assisted its participants' professional development; it has also helped UNC to increase the minority representation in its faculty, says Tony Waldrop, the university's vice-chancellor for research and economic development.

The programme is funded by the state, and supports 10 postdoc fellows at a time for two-year periods. Each receives a $36,000 stipend, healthcare benefits and a small research fund. Competition for places is fierce, with five scholars selected a year from around 130 applicants. Because the quality of the chosen fellows is so high, the university's academic departments "fall over each other" to get one of them, Waldrop says.

The ten fellows meet monthly for social gatherings, workshops and speakers. Besides providing peer support, the small group meetings allow "great access to very high level officials at the university", says past-participant Keith Esch. A Native American and a biochemist, Esch now works at RTI International, a private research institute in North Carolina's Research Triangle Park, and is an adjunct assistant professor at UNC.

For others, the luxury of time that the programme affords is key. "My main professional weakness is a lack of publication records," says Fred Hall, an African-American based in the department of physics and astronomy who is one year into the programme. It allows him to focus on writing, unhindered by teaching responsibilities.

Past participants' academic track records show that the fellows do more than just get onto the tenure track. Many have advanced into management positions, in which minorities are even less well-represented nationally. Nine are department chairs and 18 have held higher administrative positions. ■

Jill U. Adams

POSTDOC JOURNAL

Winning a plant campaign

I'm going into battle — with tomatoes. I trudge daily through a field of 3,000 plants of varying genetic constitution with which I am trying to conquer and, ultimately, expose how inflorescences — the branching structures where flowers are produced — develop. My 'to do' lists comprise a plan of attack. My weapons are a pencil, tags, a pair of tweezers and the pollen that I move from plant to plant. And the plants and the field are my opponents, mounting perilous counter-attacks.

Just last week I lost precious crossbreeds to a blown-out irrigation pipe and an early infestation of a notorious virus-transmitting fly. This is a major setback, because, on average, I have only two growing seasons a year. My fellowship has taught me to think ahead, beyond the end of this season to the next, with a vision for the end of the war. In many cases, the crosses I make now will need to go through four more generations before I see any results, so having this vision is a key component of my research programme.

When I started my postdoc, I made a list of long-term research plans, which, not surprisingly, has been modified as battles have been lost and won. After three years, I am confident I have learned to fight my battles well, but only time will tell if I will win the war. ■

Zachary Lippman is a postdoctoral fellow at the Hebrew University of Jerusalem's faculty of agriculture.

Permanent position

A lesson from history.

John Gilbey

After running out of postdoc money, and then running out of sofas to sleep on, it was a great relief to get offered work by the Institute. Only a short-term contract, obviously. Like everywhere else these days you need to be practically a Nobel laureate to get a permanent post.

Before we go any further, I should remind everyone that the Institute for Temporal Sciences — despite its reputation for secrecy — has nothing to do with time travel, except in the strictly linear sense of the term. It is really all to do with a sort of pan-spectral archaeology — sucking swarms of ultra-fine contextual data from a historical artefact and building a holistic information structure that tells you everything from the history of its travel to the emotional health and genetic make-up of everyone who has handled it.

I've always enjoyed working nights: no interruptions, nobody hassles you, no idle chats at the Eats machine. You can just get in the lab and get the job done. Nice. The building, a low mushroom-shaped structure on the science park, watched me stump up to the door and decided not to call Security — which it sometimes did out of spite, or possibly boredom. It opened the main door and lit my way down the elegantly minimalist corridor towards my lab.

The space they'd given me was big — really big. But then, before practical room-temperature superconductors came along, the kit would have been the size of a power station, if you could have built it at all. Three months ago the lab was as stylish as the rest of the building — but all the hugely expensive systems had been installed by consultants, so everything worked exactly to specification while wholly failing to deliver what was needed. They recruited me to sort out the mess — and my time has been spent ripping stuff out, reinstalling, realigning and generally commissioning — so the floor is now stained with lubricant spills and burn marks, and the walls have gained marker-scribbled lists of parameters, phone numbers and direction arrows.

While the coffee brewed I ran a system check and called up some grungy old rock as background noise. On my desk was the pile of samples for the night — all in sealed hi-tech plastic lunchboxes marked with a single barcode. Security was a hot topic in the Institute and I'd yet to see a single sheet of paper that wasn't 'Commercial In Confidence'.

Imagine my delight then, when I noticed that the seal on one of the boxes wasn't quite stuck down. After a crafty look round, I snuck a peek inside.

JACEY

A single piece of ancient paper looked back at me, covered in not-quite-Cyrillic writing and carrying a woodcut of unsettlingly graphic violence. It depicted some sort of cloaked head honcho cheerfully supervising the bloody disassembly of a rather glum fellow by some enthusiastic henchmen. Not pleasant — but hey, that's history for you. Oddly disappointed — no jewels, no gold — I sealed up the box, locked all the samples into the purge tank and thought no more about it.

The message was waiting for me at the end of my shift — on paper, in an envelope no less. The Director and the Board wanted to see me the following evening to discuss my contract — which I hoped meant an extension and, who knows, maybe tenure. Bitterly resisting the temptation to go out and get completely lashed, I trotted off home and got some sleep. A smarter bloke would have wondered about the timing.

The boardroom was as expensively austere as the rest of the building; the formally gowned board members were typical old committee hands — little hair, less conversation and no obvious sense of humour. The Director welcomed me with an almost-smile and directed me to a chair. He told me how pleased they were with me, how much they valued my contribution and all the usual flannel — including a bit about my 'interest in the subject', which hinted that they knew about the sample.

Then he paused, looked at me squarely and said: "We want to offer you a permanent position — but before we discuss that, there are some things you must understand about the Institute. Our funding comes from a group that values its privacy — as you probably know. We are dedicated to the collection of important cultural information that was previously considered lost — to the great misfortune of our funders, who have been the subject of systematic and venal persecution as a result. You would be helping to right a great wrong — and would be amply rewarded."

I liked the sound of 'amply rewarded'. Sure, I had guessed that some band of well-resourced weirdos was behind the Institute — but for permanent money I'll work for anyone even moderately legal.

"Where do I sign?" I enthused, with a big grin. An admin person with a contract and a pen oozed up to my side. The contract ran to about 20 pages — I just flicked through and poised the nib over the last page.

"The contract is permanent, and in some senses irrevocable. I would encourage you to read it closely," intoned the Director courteously. I smiled to the crowd at the table. "Hey, I trust you," I beamed, and signed with a flourish.

With a sigh of corporate satisfaction, everyone in the room grinned openly. I looked with some shock from face to face — and finally at the impressive canine teeth of the Director. As he advanced towards me — gown flaring, jaws wide — I saw just how much he now resembled the figure in the woodcut, and realized that I probably should have read the small print. ■

John Gilbey is a writer and photographer who has held a number of short-term research contracts, although not in exactly these circumstances.