THIS WEEK

Science without borders

The idea of standardizing science and removing barriers to research mobility across Europe is simple, but putting it into practice has proved more challenging.

The turn of the millennium was a time of optimistic ideas of change. European heads of state agreed to establish a utopian cross-border system that would allow the free exchange of ideas, technology and, most crucially, researchers themselves.

According to the official timetable, this European Research Area (ERA) should be in place by next year. Fat chance. Writing in this publication last month, Paul Boyle, the president of Science Europe, the Brussels-based organization of research councils, outlined the not inconsiderable obstacles to putting some of the apparently simple changes into practice, and argued that the timetable imposed was much too short (P. Boyle *Nature* **501,** 157–158; 2013).

This week, *Nature* spoke to Robert-Jan Smits, director-general of the research commission of the European Union (EU), which published the 2013 ERA progress report on 20 September. He thinks otherwise: implementation is way overdue. "We've been talking about this for 13 years!"

The idea is devilishly simple, but the devil, as always, is in the detail. According to the commission's plan, each member state should distribute most of its national research funds competitively. Research agencies should allow some funds to be shared across borders to address grand challenges or to build research infrastructures. Recruitment should be open and merit-based. Universities and research institutes should promote gender equality. Information should be open access. The idea is for all countries to have the same standards within which their scientists can thrive, and for all the barriers to research mobility to be removed.

Not all of it is difficult, Smits says. "How hard can it be for a university to put together a gender action plan? Yet fewer than 20% have." He reels off a list of other failures detailed in the September report. Almost half of researchers surveyed say they are unhappy with the transparency of recruitment procedures at their institutes. In some nations, barely 40% of research funds are distributed through competitive calls (in the most ERA-compliant countries the figure is closer to 80%). Countries have been slow to participate in joint research programmes that pool national money, or to make it easy for all comers to use some European scientific infrastructures.

But as an EU representative, Smits spins positive, listing the improvements that have been achieved and insisting that there is still time for the ERA process to be completed before 2014 ends.

The Nordic countries, Germany and the United Kingdom are doing well in the ERA process. Smits would not be drawn on which countries are doing badly, although it is likely that many of the former communist countries who have joined the EU since 2007 are among them.

Next year, the research commission will bring out the sticks if slow progress continues. It will name and shame non-complying countries, as well as organizations such as Science Europe and the League of European Research Universities that signed up to deliver the ERA to timetable. And if that doesn't work, as research commissioner Máire Geoghegan-Quinn told European ministers last week, the commission will consider drafting legislation to legally require steps towards the ERA to be taken — a particularly painful prospect for the world of academia, which likes to be self-governing.

So who is to blame for the ERA being off schedule? The commission is scientists' favourite whipping boy, but cannot be handed the blame this time. The universities and research organizations must shoulder a large share of the responsibility for not putting into practice what they signed up to do. After all, it is these organizations' scientists who will benefit from all that the ERA stands to offer. But those organizations cannot be held responsible for the genuine difficulties in breaking through endemic corruption in countries such as Romania, which actively operates against merit and competition (although stands to gain the most from the ERA).

"Without raising the potential, the European research base will remain static."

Does it matter if the ERA is not fully in place next year, especially given that the research base is better than it was as a result of the exercise? In the short term, it perhaps does not. Europe will muddle through. In the long term, yes it does. Without raising the potential by spreading world-level excellence from rich countries such as Germany and the United Kingdom to the periphery, the European research base will remain static and will probably be overtaken by growing Asian economies.

The commission is right to keep up the pressure. Europe's academic community has not found a way to govern itself into a system that gives equal opportunity to all of its scientists. Maybe it never will. It seems that the carrot — a more secure and high-flying future — is not enough. The stick of having rules enforced from above may prove more effective. ■

Dangerous work

Behavioural geneticists must tread carefully to prevent their research being misinterpreted.

Intelligence tests were first devised in the early twentieth century as a way to identify children who needed extra help in school. It was only later that the growing eugenics movement began to promote use of the tests to weed out the less intelligent and eliminate them from society, sparking a debate over the appropriateness of the study of intelligence that carries on to this day. But it was not the research that was problematic: it was the intended use of the results.

As the News Feature on page 26 details, this history is never far from the minds of scientists who work in the most fraught areas of behavioural genetics. Although the ability to investigate the genetic factors that underlie the heritability of traits such as intelligence, violent

behaviour, race and sexual orientation is new, arguments and attitudes about the significance of these traits are not. Scientists have a responsibility to do what they can to prevent abuses of their work, including the way it is communicated. Here are some pointers.

First: be patient. Do not speculate about the possibility of finding certain results, or about the implications of those results, before your data have even been analysed. The BGI Cognitive Genomics group in Shenzhen, China, is studying thousands of people to find genes that underlie intelligence, but group members sparked a furore by predicting that studies such as theirs could one day let parents select embryos with genetic predispositions to high intelligence. Many other geneticists are sceptical that the project will even find genes linked to this trait.

Second: be accurate. Researchers should design studies on the basis of sound scientific reasoning. For instance, in light of increasing evidence that race is biologically meaningless, research into genetic traits that underlie differences in intelligence between races, or that predispose some races to act more aggressively than others, will produce little. Furthermore, it is common for small studies of behavioural genetics to go unreplicated, and there are increasing concerns that the science of behaviour more generally suffers from poor practice, exaggeration and irreproducibility (see *Nature* http://doi.org/n2m; 2013). Scientists should refrain from claiming that they have found a basis for any complex trait until the results have been replicated and confirmed in large, definitive studies, such as multiple meta-analyses.

Third: be sensitive. Even if scientists have truly honourable intentions, they must realize how easy it can be for studies on socially favoured groups to seem self-serving. For instance, BGI's study of exceptionally intelligent individuals is itself led by people who are unusually bright, even in the cognitively enriched domain of science: there is a child prodigy who dropped out of high school to work on genomics; a physicist who graduated from university at age 19; and an International Mathematical Olympiad gold medallist. When such people make statements in favour of selecting embryos for intelligence, it can seem to the public as if the researchers think that society would benefit from the birth of more people just like them — even if this is not what they have in mind.

Finally: be proactive. Once scientists are sure of their results, they usually do their best to explain the significance of their work in academic publications. But these texts are often impenetrable to the public and may include technical terms that can be misinterpreted by non-specialists. To provide clarity, scientists would do well to follow the example of the Social Science Genetic Association Consortium. In June, this group published a paper on genetic variants associated with educational attainment (C. A. Rietveld *et al. Science* **340,** 1467–1471; 2013). Accompanying this was a nine-page Frequently Asked Questions document that, in plain, easy-to-understand language, addressed such questions as why the researchers did the study, what they found and what the implications of the work are — and are not (see go.nature.com/7mov2j). The document spelled out that the consortium had not found 'the gene' for educational attainment, that each genetic marker found has only a very small effect on length of schooling, and that any policy response based on that single study would be premature.

"Scientists have a responsibility to do what they can to prevent abuses of their work."

Scientists cannot be held responsible every time someone misinterprets their work. But simple steps such as these could help to prevent and address some of the potential distortions of behavioural genetics — and could help to ensure that society continues to support the work. ■

Cross the road

Research on chickens is legitimate — but scientists and funders must learn to justify it.

Taxpayers underwrite many public services, including the funding of science. So it is entirely right for them to question funding decisions. If they do, granting agencies should have mechanisms for responding in ways that are informed but not patronizing.

On 18 September, the UK Arts and Humanities Research Council (AHRC) announced nine grants, most of which aimed to bridge the gap between science and the humanities. The majority were uncontroversial. Nobody blinked, for example, at the £1.95 million (US$3.1 million) given to Colin Blakemore of the Institute of Philosophy in London for a project entitled 'Rethinking the Senses: Uniting the Philosophy and Neuroscience of Perception'. No eyebrow was raised when Randolph Donahue at the University of Bradford got £1.98 million to study 'Fragmented Heritage: From the kilometre to the nanometre: Automated 3D Technology to Revolutionise Landscape, Site and Artefact Analyses'.

But when Mark Maltby at Bournemouth University was awarded £1.94 million for 'Cultural and Scientific Perceptions of Human–Chicken Interactions', the reaction from some tabloid newspapers was predictable. "A birdbrained idea? Outrage as academics are handed £2m to study how humans interact with CHICKENS," crowed *The Daily Mail*. "Chicken study costing £1.9million of taxpayers' funds causes a flap," squawked *The Daily Express*.

Why the outrage? Could it be that journalists came across the AHRC press release, recognized the word 'chicken' in the morass of science-speak and went for an easy sell — lambasting the indulgence of barmy boffinry with taxpayers' money at a time of austerity? Why 'easy'? Well, whereas not many people know much about neuroscience or nanometres, everyone knows what chickens are. So much so that they feel they can take interactions with the birds for granted, and ask what more we would learn by spending almost £2 million on the subject. It is in that familiarity, however, that the questions lie. We know surprisingly little about the history of human–chicken relations, such as how chickens first came to Britain.

Behind the over-excited headlines lies a legitimate question about accountability. If it is right and proper for researchers, rather than politicians, to decide how public funds should be spent (the 'Haldane principle'), then those researchers should be ready to justify such decisions, promptly and simply. For example, after Greger Larson of Durham University appeared on radio and television this year to talk about his work on the domestication of dogs, he received an e-mail that demanded, bluntly, whether the £1 million being spent on such a subject came from the taxpayer. Larson replied with a polite, informative and, most importantly, personal e-mail explaining where the money came from — and how it fitted into the context of UK government funding.

The denigration of science by media outlets and some politicians relies on an us-against-them mentality. This can be weakened by individual personal engagement such as Larson's. Many corporations are breaking down barriers by interacting with customers through social media such as Twitter and Facebook, replying to comments much faster than they would through more conventional, formal channels. Customers appreciate the speed of service and the fact that it can be personalized, and come to feel more engaged with that corporation's aims.

Research bodies have not been slow to use such media. The AHRC, for example, has a Twitter feed (@ahrcpress), as does the Natural Environment Research Council, which funded Larson (@NERCscience). It is only a matter of time before taxpayers communicate routinely with researchers using such methods. Informal networks will help the public to become more engaged with the work that their money funds — demonstrating the value, if you like, of human–human interactions. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

PHIL ROBERTS

The Anthropocene could raise biological diversity

Humanity has wrought an age of ecological transformations. It is time to rethink our irrational dislike of invading species, argues **Chris D. Thomas**.

Human activity changes the environment, as last week's release of a report by the Intergovernmental Panel on Climate Change reminds us. But not all change is bad. One way in which animals and plants respond to warming temperatures, for example, is to move beyond their historical distributions, just as they do when they are transported to new regions by humans. The response of people who find themselves 'invaded' by such 'displaced' species is often irrational. Deliberate persecution of the new — just because it is new — is no longer sustainable in a world of rapid global change.

It is true that some invasive species damage ecosystems and can eradicate resident species. As a result, the European Commission, for example, is planning laws to control the 'adverse' impacts of species introduced through human activities, albeit without quite saying how those impacts should be defined. But the same process can also increase ecological diversity. On average, less than one native species dies out for each introduced species that arrives. Britain, for instance, has gained 1,875 established non-native species without yet losing anything to the invaders.

Human development — dubbed the age of the Anthropocene — boosts biodiversity in other ways too. New anthropogenic habitats, such as farmland and cities, usually support fewer species than the original ones, but they contain some that were previously rare or absent. The ensemble of new and old habitats holds more species than the original vegetation — habitat diversity is one of the strongest predictors of ecological diversity. Climate change also tends to boost regional diversity, because diversity increases with temperature and precipitation, both of which are rising (on average, but not everywhere). Global-diversity gradients dictate that more warm-adapted species are available to colonize new areas than cold-adapted species retreat from those areas as the climate warms.

Evolutionary origination is also accelerating. Populations and species have begun to evolve, diverge, hybridize and even speciate in new man-made surroundings. Evolutionary divergence will eventually generate large numbers of sister species on the continents and islands to which single species have been introduced. For example, marked reproductive incompatibility has developed in just 200 years between source populations of *Centaurea* plants in Spain and introduced populations of the same species in California. When should the citizens of California regard these plants as native?

Hybridization is becoming particularly important as formerly separated species are brought into contact. The rates are astounding: 88 hybrids between native and introduced plant species are sufficiently widespread to be mapped in the British Isles flora, as are 26 hybrids between two or more introduced species (together equivalent to 8% of the 1,377 higher plant species that have become naturalized following introduction). For example, introduced European *Rhododendron ponticum* plants hybridized with North American *R. catawbiense*, producing a vigorous, self-sustaining population that is hated by conservationists and removed at great expense.

It is a mistake to misdirect valuable and increasingly scarce conservation funds into unwinnable wars, especially when the enemy is not especially damaging. Eradication programmes should concentrate on problematic non-native species, such as rats and goats on oceanic islands, where the investment can deliver long-term benefits and the re-establishment of native species. Trying to control Himalayan balsam throughout England, just because it is alien, is a waste of effort.

Speciation by hybridization is likely to be a signature of the Anthropocene. A new hybrid species of *Rhagoletis* fruitfly has colonized invasive honeysuckle in North America. A primrose species, *Primula kewensis*, arose by hybridization and continues to be propagated in London's Kew Gardens. And five species (*Spartina anglica* and four *Senecio* species) that have arisen by hybridization between native and introduced species in Britain have become naturalized. Remarkably, the introduction of plants to Britain seems to have increased the global species list. These five (out of a flora of 2,711 naturalized and native species) suggest a speciation rate (0.00184 per original species in the past 150 years) similar to the extinction rate reported for mammals over the past 100 years. If sustained, with no subsequent extinctions, it would be sufficient to increase the number of plant species by 20% within 15,000 years.

SPECIATION BY HYBRIDIZATION IS LIKELY TO BE A SIGNATURE OF THE ANTHROPOCENE.

Rather than the catastrophic declines often portrayed, empirical evidence points to ecological increases in the number of terrestrial species in most of the world's regions over recent decades and centuries, even though the total number of species on the planet is declining.

We need more-concerted scientific investigation of the rates at which different processes generate diversity. Together, they could plausibly result in a net increase in the number of species on Earth during the Anthropocene (say, over a million years), despite the fact that we are losing irreplaceable populations, races, species and evolutionarily distinct taxa. There are excellent arguments for conserving the wildlife we already have, but it is less clear why our default attitude to novel biodiversity is antagonism or ambivalence. One recent hybrid species, *Senecio eboracensis*, became extinct soon after it arose in York, arousing little concern. In practice, it seems that new Anthropocene species are regarded as far less valuable than those that went before. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/zxa3ta

Chris D. Thomas *is professor of conservation biology at the University of York, UK.*
e-mail: chris.thomas@york.ac.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

AGRICULTURE

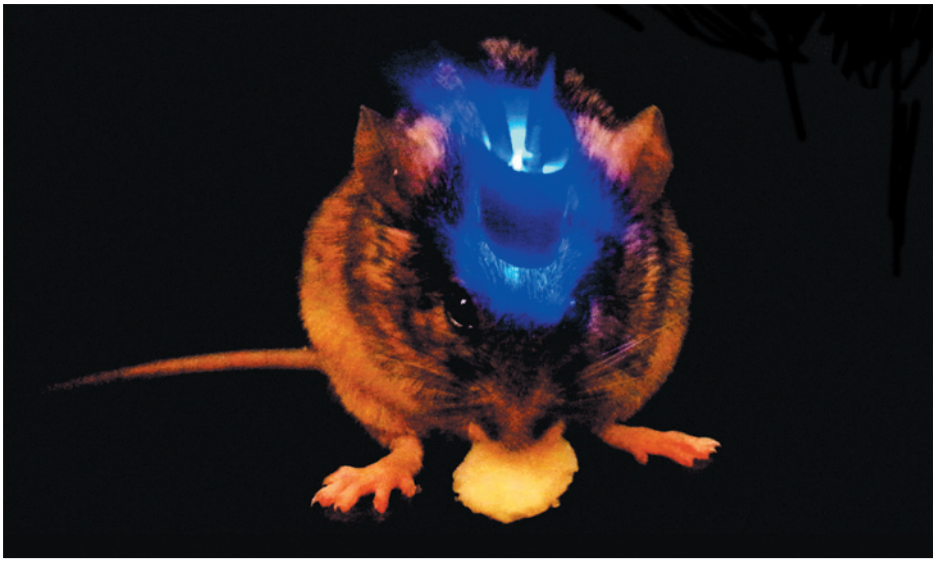
Easy to thresh and better to sow

When domesticating wheat, Neolithic humans preferred plants that kept the grain on the stalk until ripe. That not only made the crop easier to harvest, but also made it faster to separate grain from chaff.

Shahal Abbo at the Hebrew University of Jerusalem and his colleagues experimentally threshed nearly 200 types of wheat, including wild strains and traditional varieties. For wheat heads that did not shatter, threshing time was reduced by about 30% compared with more-brittle types. And for plants with resilient heads and weak husks, threshing time decreased by a further 85%. This resulted in many fewer broken kernels, making the seed more likely to be saved for future sowing.

Post-harvest processing, sometimes considered a barrier to the domestication of cereal crops, may have played an underappreciated part in promoting it, the authors say.

Ann. Bot. 112, **829–837 (2013)**



JOSH JENNINGS

NEUROSCIENCE

Hunger neurons hunted out

Scientists have revealed a key brain circuit that causes mice to eat uncontrollably.

Garret Stuber at the University of North Carolina, Chapel Hill, and his colleagues manipulated a precise set of brain cells using optogenetics — a technology that lets scientists control neurons by shining light into living brains (**pictured**). The neurons that they studied connect two brain areas: the lateral hypothalamus, which is involved in eating behaviour, and the BNST, which collects input on 'motivational states' such as hunger and thirst.

When these neurons were activated, well-fed mice began to gorge and showed a marked preference for high-calorie food. When the cells were inhibited, even mice that were previously deprived of food refrained from eating.

Science 341, **1517–1521 (2013)**

APPLIED PHYSICS

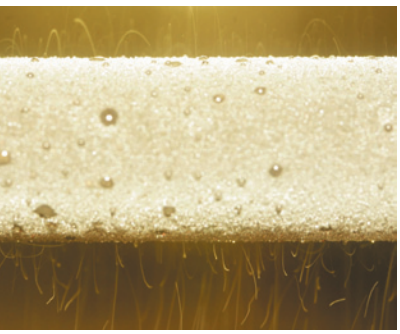
Jumping droplets repel each other

Water droplets that form on strongly repellent surfaces often coalesce and leap off. When they do, they carry an electric charge that can be used to control them.

Evelyn Wang at the Massachusetts Institute of Technology in Cambridge and her colleagues observed that jumping droplets sometimes repel each other in the air (**pictured**). To learn why, they studied droplets coalescing on a superhydrophobic copper oxide nanostructure, and found that the droplets sometimes gain a small positive charge as they merge and leap into the air.

The authors suggest that this electrostatic effect could be exploited to remove or manipulate the droplets, and so produce surfaces that can be easily cleaned or de-iced.

Nature Commun. 4, **2517 (2013)**



NENAD MILJKOVIC AND DANIEL J. PRESTON/MIT

CANCER

Tumour types have traits in common

Combining genomic data from tumours found in different organs has revealed previously unknown cancer-related genes, and has led to a system for classifying tumours that could predict a patient's response to therapy.

In recent years, studies of cancer genomics have focused on mapping the genetic changes that contribute to the disease. Rameen Beroukhim of the Broad Institute in Cambridge, Massachusetts, and his colleagues analysed how the number of copies of genes varied across almost 5,000 tumours representing 11 kinds of cancer. The team found 140 genome regions in which copy-number variations were associated with cancer. Of those, 102 contained no known cancer-related gene, suggesting a suite of new cancer drivers.

In related work, Chris Sander and his colleagues at

the Memorial Sloan-Kettering Cancer Center in New York used data from 12 cancer types to group tumours on the basis of genomic signatures rather than according to the organ in which the tumour originated. These signatures might one day be used to personalize cancer therapies.

Nature Genet. 45, **1127–1133; 1134–1140 (2013)**

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

ASTRONOMY

Fine weather on far-off planet

Skies are clear and blue on an extrasolar planet 14 times more massive than Earth.

Using the Large Binocular Telescope in Arizona, a team led by Valerio Nascimbeni at the University of Padua, Italy, studied ultraviolet and infrared light coming from the dwarf star GJ3470, 31 parsecs from Earth. By watching the star dim as a known planet passed in front of it, the astronomers could probe light scattered by the planet's atmosphere.

The data suggest that the planet has a blue, cloud-free sky, which could help to reveal the composition of its atmosphere. The data are also precise enough to suggest that ground-based telescopes can now be used to discover Earth-sized planets around similar stars, rather than relying on space-based observatories.

Astron. Astrophys. **http://doi.org/n2b (2013)**

CLIMATE SCIENCES

How plants helped Earth to stay cool

Plant growth spurred by rising levels of carbon dioxide pollution in the atmosphere has slowed the rate of global warming considerably.

Stephen Pacala of Princeton University in New Jersey and his colleagues used an Earth-system model to analyse historical emissions from industry and changes in land use, including deforestation and agricultural development. The team simulated the carbon cycle and climate from 1861 to 2005 with and without the fertilization effect of CO_2 on vegetation.

Although the model did not incorporate certain factors, such as pre-1960 land use, the results suggest that if vegetation had not soaked up so much of the gas, levels of CO_2 would have risen by an extra 85 parts per million — boosting temperatures by about 0.31 °C.

Proc. Natl Acad. Sci. USA **http://doi.org/n2c (2013)**

ORGANIC ELECTRONICS

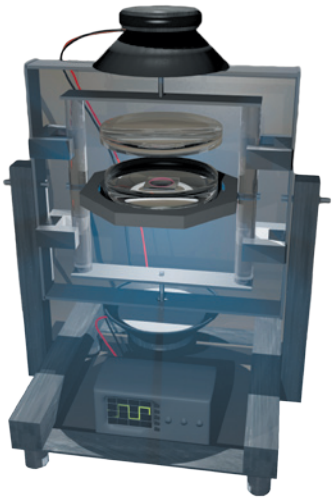
Sound shakes semiconductors

Gently vibrating a solution of semiconductor molecules as they crystallize into a conductive film helps to reduce structural flaws. This yields higher-quality organic transistors that can be used in flexible, lightweight electronic devices.

A team led by Oana Jurchescu of Wake Forest University in Winston-Salem, North Carolina, used audio speakers operating at a low-frequency of around 100 hertz to shake molecules as they formed a thin crystalline film (apparatus **pictured**). This boosted the film's semiconducting properties, making it almost as good a semiconductor as single crystals grown by diffusion from molecular vapour — a method that, unlike shaking liquid solutions, does not lend itself to high-throughput manufacture at room temperature.

Other methods such as heating or including additives in the film also improve superconductivity, but vibration might be cheaper and more scalable, the researchers think.

Adv. Mater. **http://dx.doi.org/10.1002/adma.201302838 (2013)**



PETER DIEMER/WAKE FOREST UNIV.

COMMUNITY CHOICE

The most viewed papers in science

ANIMAL BEHAVIOUR

Chimps ignore watching eyes

HIGHLY READ on www.elsevier.com/animal-behaviour in September

Unlike humans, chimpanzees do not alter their behaviour significantly when eyes are gazing down on them.

Daniel Nettle at Newcastle University, UK, and his colleagues observed chimps (*Pan troglodytes*) consuming shelled peanuts in front of a large image of a chimp (**pictured**). Although chimps adjust their eating habits in the presence of dominant chimps and recognize such stylized black-and-white cartoons as faces, the animals in these experiments did not hesitate to take peanuts when 'watched' by the image. Humans are more charitable and honest under images of watching eyes than in their absence, and may be unique in their extreme sensitivity to faces, the authors say.

Anim. Behav. 86, **595–602 (2013)**

DANIEL NETTLE

NANOTECHNOLOGY

Friction at the atomic scale

The effect of atomic bonds on friction has been demonstrated, for the first time, at the scale of just a few atoms.

Variations in atomic surfaces are thought to modulate the force of friction in a way that depends on the direction in which objects are moved, but this has been difficult to show experimentally. Jay Weymouth of the University of Regensburg in Germany and his colleagues have now done so.

They passed the tip of a tungsten lateral-force microscope between regions of a silicon crystal surface, on which pairs of silicon atoms were oriented at right angles to each other. Oscillations of the tip varied depending on whether it was sliding along or across the direction of the paired atoms.

Phys. Rev. Lett. 111, **126103 (2013)**

NATURE.COM

For the latest research published by *Nature* visit:

www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Climate report

In its latest assessment, released on 27 September, the Intergovernmental Panel on Climate Change gave its first formal estimates of the total carbon dioxide emissions that can be released while limiting CO_2-induced global warming. Total emissions must be kept below 1,800 gigatonnes for a 66% chance of keeping average global warming to 2 °C above pre-industrial levels, the report says. At current CO_2 emissions rates, about 37 gigatonnes per year globally, that limit will be exceeded in less than 50 years. See go.nature.com/yqx3lm for more.

Spanish recovery

Spain's government has proposed its first increase in science funding since 2009, as part of a 2014 budget presented on 27 September. Total spending on research and development (R&D) would go up by 1% to €6.1 billion (US$8.3 billion), with non-defence research up by 6%. But there is a catch: more than half the budget would be awarded in loans, which in past years have been promised to companies and not used. And funding remains depressed after a 39% cut to R&D over the past five years (not counting inflation). See go.nature.com/7x694q for more.

JOURNALISM AWARD

Science writer Jennifer Frazer last week won the American Meteorological Society's Award for Distinguished Science Journalism in the Atmospheric and Related Sciences for her News Feature in *Nature* 'Blowing in the wind' (*Nature* **484,** 21–23; 2012).

NATL INST. OCEANOGRAPHY

Pakistan quake throws up island

An island appeared in the Arabian Sea on 24 September, the apparent result of a magnitude-7.7 earthquake that shook south-central Pakistan (see go.nature.com/kmcchc). The low island (pictured) arose about 1 kilometre off the country's Gwadar coast, an area known for mud volcanoes that appear in coastal waters and are usually washed away within months.

The new island measures about 50 metres long by 20 metres wide and 10 metres high, says Asif Inam, a marine geologist at the National Institute of Oceanography (NIO) in Karachi. Seismic shaking probably caused mud, mixed with methane gas, to extrude from the sea floor and form the island, he says. The NIO plans to survey the area to search for similar features.

French budget

Research and higher-education funding remained largely unchanged in France's draft 2014 budget, released on 25 September. The proposal fulfils pledges made by French President François Hollande and Prime Minister Jean-Marc Ayrault to introduce a long-awaited carbon tax, which is set to begin next year. The tax is expected to bring in €340 million (US$460 million) in 2014, and about €4 billion by 2016. See go.nature.com/wepepo for more.

E-cigarette control

Pressure intensified last week for the US Food and Drug Administration (FDA) to take tough action on electronic cigarettes, when 40 state attorneys general called for "immediate regulatory oversight" of the products. 'E-cigarettes' have proved controversial among tobacco-control advocates (see *Nature* **501,** 473; 2013), and the FDA is expected to issue proposed regulations next month. In a letter dated 24 September, the attorneys general warned that the "increasingly widespread, addictive product" is being marketed to young people.

Helium reserve

The US Congress finalized on 26 September a bill to avert an imminent shutdown of the federal helium reserve, which provides more than one-third of the world's supply. President Barack Obama is expected to sign the bill, which will ramp down the reserve over several years instead of closing it abruptly in October. Experts had predicted that the sudden shutdown would cause a spike in helium prices. Liquid helium is commonly used as a coolant for ultra-low-temperature physics experiments and magnetic resonance imaging machines. See go.nature.com/vlmgbt for more.

EVENTS

US shutdown

Numerous research agencies have been affected by the US government's shutdown on 1 October, as lawmakers failed

to agree a 2014 spending plan. Government scientists stayed at home, grant-making ceased at the National Institutes of Health and the National Science Foundation, and government research ships were called back to port. See page 13 for more.

Space delivery

The second commercial cargo vehicle ever to fly to the International Space Station has arrived. On 29 September, astronauts aboard the space station used a robotic arm to dock the crewless Cygnus craft, made by Orbital Sciences of Dulles, Virginia. The company is competing with SpaceX of Hawthorne, California, to provide NASA with for-profit space-transportation services. Cygnus, which launched on 18 September, carried about 700 kilograms of supplies for the astronauts. It is meant to stay at the space station for a month, being loaded with cargo for disposal, before returning to Earth, where it will burn up on re-entry.

PEOPLE

Genius grants

The MacArthur Foundation, based in Chicago, Illinois, announced the 2013 recipients of its 'genius' grants on 25 September. Thirteen scientists were among

JOHN D. & CATHERINE T. MACARTHUR FOUNDATION

the 24 winners, including experimental physicist Carl Haber at the Lawrence Berkeley National Laboratory in California and statistician Susan Murphy (**pictured**) at the University of Michigan in Ann Arbor. Winners receive 'no-strings-attached' awards of US$625,000 paid over five years. See go.nature.com/4mr7qo for more.

FUNDING

Poaching pushback

The Clinton Global Initiative in New York launched on 26 September a US$80-million effort to clamp down on elephant poaching in Africa. The programme brings together several conservation groups and African nations to increase law enforcement at 50 sites across Africa, and to detect and prosecute smugglers. Elephant poaching has surged in recent years, driven particularly by ivory demand in Asia. Officials in Zimbabwe reported last week that more than 80 elephants had been killed with cyanide in a national park.

RESEARCH

Protein project out

The Protein Structure Initiative, a high-throughput US pipeline to solve thousands of protein structures, will end after 2015 (see *Nature* **466,** 544; 2010). Jon Lorsch, director of the US National Institute of General Medical Sciences in Bethesda, Maryland, announced the institute's plan for the 13-year-old programme in a blog post on 24 September. An independent review recently concluded that the project was not making its work relevant enough to biology, and Lorsch cited the need to free up more funds for investigator-initiated work in tight fiscal times.

Fouchier flu fight

Influenza expert Ron Fouchier protested on 26 September against a Dutch court's decision to uphold government oversight of his research, calling the court's arguments "weak". Fouchier, of the Erasmus Medical Center in Rotterdam, the Netherlands, has sparked controversy by creating mammalian-transmissible strains of the H5N1 avian flu virus. On 20 September, he lost his case to exempt such research from export control laws, which are aimed at restricting biological weapons. The regulations require researchers to obtain export permits before disseminating 'dual-use' materials and information that could have both legitimate and malicious uses. See go.nature.com/x1p9ea for more.

Defence contest

The US Department of Defense is taking heat from genome scientists after announcing last week the winners of a US$1-million software competition. Participants submitted algorithms designed to quickly identify pathogens from DNA sequencing data. Several researchers have criticized the contest as being poorly run, citing changes made to the scoring system during the competition. The solution by the winning team, from Germany and Singapore, will be used to address biological threats to the US military, the agency says.

COMING UP

6–11 OCTOBER
The latest findings from space missions such as MESSENGER, Cassini, Curiosity and Kepler are discussed at the American Astronomical Society's 45th annual meeting of the Division for Planetary Sciences, in Denver, Colorado.
go.nature.com/epvznl

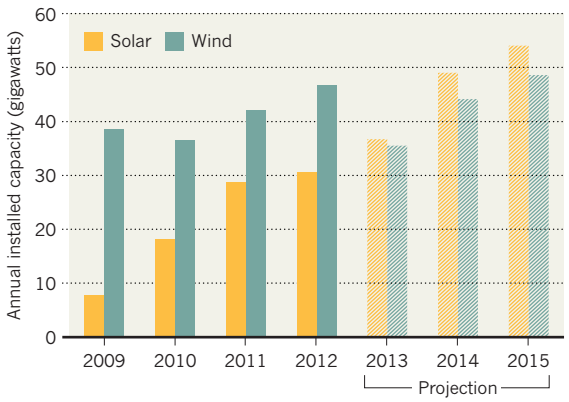
7–9 OCTOBER
The winners of the 2013 Nobel prizes for physiology or medicine, physics and chemistry are announced in Stockholm.
www.nobelprize.org

TREND WATCH

Owing to uncertainty over wind-energy policies in the United States and China, the number of new wind-power installations will drop by 25% worldwide this year. For the first time, solar energy will overtake wind energy in new installations (see chart). The forecast comes from analysts at Bloomberg New Energy Finance, who add that the number of solar installations is increasing in Japan and China because of financial incentives, but dwindling in western Europe, where subsidies are being cut.

MORE SOLAR THAN WIND

The number of new photovoltaic panels will overtake the number of new wind-power installations this year, say analysts.



SOURCE: BLOOMBERG NEW ENERGY FINANCE

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

NEWS IN FOCUS

CHIP SOMODEVILLA/GETTY

A protester in Washington DC harangues members of Congress over their failure to pass the next budget.

POLITICS

US government shuts down

Research disrupted as lawmakers spar over funding.

BY LAUREN MORELLO

The US government entered a state of suspended animation on 1 October after Congress failed to agree on a budget for the next fiscal year, causing federal agencies — including those overseeing science policy and research — to shut down indefinitely.

Most government scientists were ordered to stay at home, their offices and labs closed or run by a skeleton staff of 'essential' workers. The National Institutes of Health (NIH) and the National Science Foundation (NSF) stopped processing grants, some government websites were made inaccessible and many important research programmes were left hanging, potentially putting lives at risk in the case of some disease studies. Use of government telephones and e-mail was also suspended. The restrictions were still in place as *Nature* went to press.

The shutdown is the product of fierce political infighting in Congress, and comes as a cruel blow to researchers who were already struggling with successive rounds of federal cuts — including the 5.1% across-the-board 'sequestration' that took effect on 1 March.

"This is ridiculous," says Jennifer Zeitzer, director of legislative affairs at the Federation of American Societies for Experimental Biology in Washington DC. "We can't continue to survive as a research community this way."

At the NIH, headquartered in Bethesda, Maryland, 73% of the agency's 18,646 employees were immediately placed on furlough, or enforced leave. The agency also stopped accepting patients for its clinical trials or initiating new studies. Minimal staff remain to care for lab animals and to protect NIH facilities.

The NSF, in Arlington, Virginia, ordered 98.5% of its roughly 2,000 employees to stay at home. One notable exception was staff in the Division of Polar Programs, which oversees the agency's trio of Antarctic research stations and its remote facilities in Greenland.

The National Oceanic and Atmospheric Administration (NOAA), based in Washington DC, retained almost 5,400 of its 12,000 employees, largely to support the essential work of the National Weather Service. Most of its scientists were put on leave, with some exceptions; for example, roughly a dozen people will stay on to maintain the agency's six greenhouse-gas monitoring stations, including sites in Hawaii, Alaska, Greenland and Antarctica. But the team of scientists in Boulder, Colorado, that analyses the data collected by those stations has been told to stay at home, even as flasks of air samples shipped from NOAA field sites begin to pile up.

Pieter Tans, who heads NOAA's Carbon Cycle and Greenhouse Gases group in Boulder, was planning to bring work home to help occupy him during the shutdown, because he would be unable to access agency computers or his work e-mail. "I'll be busy, but you bet I'm angry," he said, hours before the government began winding down its operations. Tans said that he and his colleagues are treating the shutdown with a sense of resignation after several years of uncertain budgets. "In ten years, our programme will be totally gutted if this continues."

Some agencies are more fortunate. The Food and Drug Administration (FDA) in Silver ▶

Spring, Maryland, receives substantial user fees from the pharmaceutical industry that fund an estimated two-thirds of its drug-review process. Although the FDA has put 45% of its staff on leave and will cut back on food-safety programmes, user fees might keep its drug-review pipeline open — albeit operating more slowly than normal, says Timothy Coté, founder of Coté Orphan Consulting in Silver Spring and a former director of the FDA's Office of Orphan Products Development.

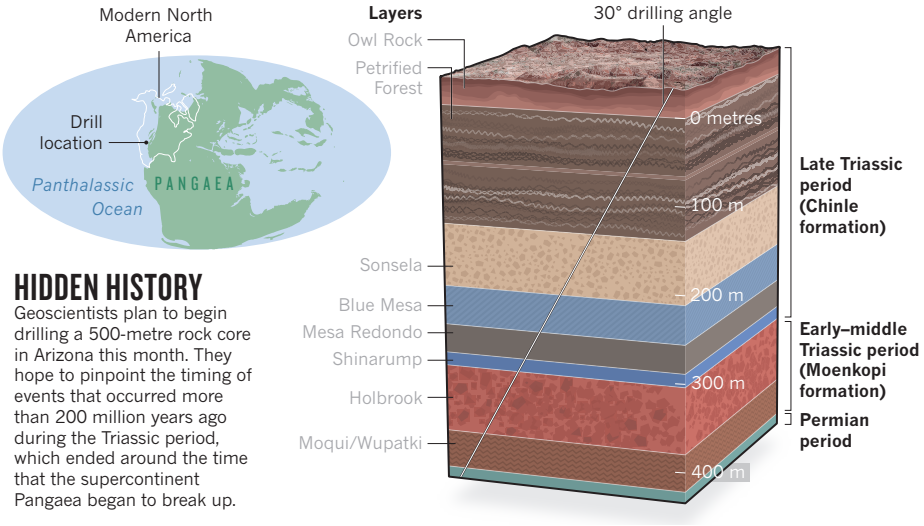
But a lingering shutdown would have knock-on effects. For instance, if the NSF misses one or two weekly payments to the National Radio Astronomy Observatory in Charlottesville, Virginia, the facility would be forced to close, disrupting long-term research, says facility director Tony Beasley.

At NASA, one casualty could be the Mars Atmosphere and Volatile Evolution (MAVEN) mission, which until 1 October was being prepared at Cape Canaveral in Florida for an 18 November launch. MAVEN's principal investigator, Bruce Jakosky of the University of Colorado Boulder, says that his team can accommodate a brief work stoppage. But if MAVEN, which will study the Martian atmosphere, misses its three-week launch window, it will be delayed until 2016, when Mars and Earth will again be favourably positioned in their orbits. Under NASA's contingency plans, operational missions, such as work on board the International Space Station, will continue.

Meanwhile, the shutdown forced the Centers for Disease Control and Prevention in Atlanta, Georgia, to halt its tracking of influenza cases just when the US flu season normally begins. The agency will also cut back on surveillance of emerging infectious diseases such as the Middle East respiratory syndrome coronavirus, says spokeswoman Barbara Reynolds. The Environmental Protection Agency has retained some staff to look after its environmental-health and security functions and to care for animals being used in studies.

Predicting when and how the shutdown might end is difficult. The last such event, in early 1996, continued for a record 21 days, although shutdowns have historically been much shorter. The latest dispute is rooted in attempts by the Republican-controlled House of Representatives to withhold funding for the health-care reform programme championed by President Barack Obama. The Democrat-controlled Senate has repeatedly rejected the House manoeuvres, with little sign that the two bodies are willing to compromise. ■

Additional reporting by Erika Check Hayden, Heidi Ledford, Brendan Maher, Maryn McKenna, Jeff Tollefson, Alexandra Witze and Sarah Zhang.



HIDDEN HISTORY
Geoscientists plan to begin drilling a 500-metre rock core in Arizona this month. They hope to pinpoint the timing of events that occurred more than 200 million years ago during the Triassic period, which ended around the time that the supercontinent Pangaea began to break up.

GEOSCIENCE

Geologists take drill to Triassic park

Arizona rock core to yield coherent picture of turbulent period.

BY ALEXANDRA WITZE

Tourists flock to Petrified Forest National Park in Arizona to marvel at great glittering logs of petrified wood. But geologists hope to flock there this month in search of something less visible and more scientifically significant: a core obtained by drilling half a kilometre into rock more than 200 million years old.

Drillers will spend several weeks boring through layers of rock that house the fossils of tiny early dinosaurs and giant crocodile-like phytosaurs, as well as the leaves and pollen of an entire fossilized ecosystem. The goal of the US$970,000 drilling project is to stitch together a complete picture of most of the middle and late Triassic period, a turbulent interval that saw both a mass-extinction event and the emergence of dinosaurs. Geoscientists hope to use the decay of radioactive uranium in layers of volcanic ash in the core to precisely date events between about 205 million and 235 million years ago, just before the supercontinent Pangaea began to break apart.

"It's a unique opportunity to put together a coherent time framework for a critical part of the Triassic," says John Geissman, a geologist at the University of Texas at Dallas and one of the project's leaders. "Sure, we have other continental Triassic records, but the Petrified Forest area is pretty darn good when it comes to details."

The Petrified Forest effort has been years in the making. It is a follow-up to a project in which a Triassic core was drilled from New Jersey's Newark sediment basin between 1990 and 1993 (ref. 1). That project aimed to tease out changes in the amount of sediment that was deposited as Earth went through cyclical shifts in the shape of its orbital path around the Sun. "If we can show that the Newark timescale is correct, we can empirically calibrate the Solar System's behaviour," says Paul Olsen, a geologist at the Lamont–Doherty Earth Observatory in Palisades, New York, and a member of the project team. "That's probably the most exciting aspect for me."

The effort, funded by the US National Science Foundation and the International Continental Scientific Drilling Program, might also help to resolve a simmering dispute. Comparisons of the Newark data with data from Triassic rocks in the Mediterranean have led some researchers to suggest radically revising the period's history. This reworking would lead to one subdivision — the Norian stage — taking up nearly half of the entire Triassic period, drastically changing dates of key evolutionary events, including the emergence of certain dinosaurs.

The idea of a 'long Norian' remains fiercely controversial[2], and the Petrified Forest core would need to capture a sufficient record to settle the debate. But the rocks have plenty of chronological gaps, owing to weathering or abrupt geological events. Because of surface erosion, for example, the core will not capture the very end of the Triassic around 200 million years ago, when a mass extinction swept across the planet, killing many dinosaurian relatives. The core will instead start in rocks dating to around 205 million years ago, in layers known as

the Chinle formation (see 'Hidden history'). It will travel, with several breaks in time, through the Moenkopi formation and stop in rocks about 235 million years old. The record then skips tens of millions of years into rocks from the Permian period that preceded the Triassic.

"We know parts will be missing," says Geissman. But getting a nearly complete record for much of the Triassic, in such well-studied rock layers, is bound to offer a trove of information.

Geologists have explored the Petrified Forest area since the 1850s, most recently for its rich array of Triassic fossils. Since 2004, for instance, several skeletons have been unearthed of an extinct crocodile-like animal called *Revueltosaurus*, previously known only from its teeth. Early dinosaurs such as the dog-sized *Coelophysis* also roamed there, and radiometric dating has shown how these dinosaurs were related to those in other parts of the Americas[3].

Spectacular, fossil-bearing rocks sprawl almost everywhere in the park, says Bill Parker, Petrified Forest's palaeontologist. The challenge is tying separate discoveries into a coherent, well-dated story. Many surface rocks are weathered so badly that they distort fossil relationships and make radiometric dating all but impossible. "It's not like the Grand Canyon, where you can just hike down and see all the rocks in their proper order," says Parker. "A core eliminates all of that problem, when you get one single section all the way down." Drilling in US national parks is allowed at the discretion of the park superintendent. Petrified Forest is unusual in calling itself a science park and in having Parker on staff as a full-time palaeontologist.

The big question for researchers is when drilling can begin. The team had hoped to start on 8 October, but that plan is now in doubt. Petrified Forest National Park, along with the rest of the US government, shut down on 1 October. The park will not reopen until Congress agrees on a plan to keep the government funded. If delayed too long, drilling may have to be rescheduled for next spring.

Ultimately, if the project's science findings are strong, it will pave the way for further studies of the Triassic's buried history. The team already has its eye on other cores that it could drill. ■

1. Olsen, P. E. *et al. Geol. Soc. Am. Bull.* **108,** 40–77 (1996).
2. Lucas, S. G. *et al. Earth-Sci. Rev.* **114,** 1–18 (2012).
3. Irmis, R. B. *et al. Earth Planet. Sci. Lett.* **309,** 258–267 (2011).

POLICY

Overhauls set scientists on edge

Australian government axes carbon tax and designated science minister, but says it will not cut research funding.

BY CHERYL JONES

Changes made by Australia's newly elected conservative government, sworn in on 18 September, are unnerving scientists. Prime Minister Tony Abbott was quick to renew his pre-election pledge to axe the country's carbon-pricing scheme, and the Coalition administration has begun to kill off some of the main government agencies tasked with tackling climate change. Abbott opted not to appoint a single minister for science, deciding instead to spread responsibility for it over several ministries, including industry and education.

It is a tricky time for people working in carbon policy in Australia, and scientists are worried that research into global warming is under threat. David Karoly, an atmospheric scientist at the University of Melbourne, says that it is unclear whether funding for climate-change studies, including those conducted by the Bureau of Meteorology and Australia's national science agency, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), will be held at current levels.

University research may also be vulnerable. "There are fears about the funding of climate-change research, particularly on mitigation," says Karoly.

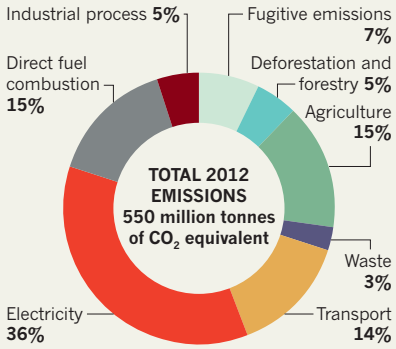
Australia is a world leader in climate-change research. As one of the world's biggest emitters of greenhouse gases per capita (see 'Australia's emissions'), its scientists have been a big part of the Intergovernmental Panel on Climate Change. Overall, the country has a high scientific output (see 'Punching above its weight'). But some fear that this work may be in jeopardy. The government has already closed the country's Climate Commission, a public-education body, although plans are afoot to revive it as a private entity. And the government wants to close the Climate Change Authority, which provides advice on emissions reductions, says Karoly, a member of the authority's board.

The Clean Energy Finance Corporation, or 'green bank', a Aus$10-billion (US$9.4-billion), 5-year programme established by the former administration to provide loans for the commercialization and deployment of clean-energy technologies, is also under threat.

Responsibility for carbon policy will form part of the environment portfolio, held by minister Greg Hunt. Until now, the centrepiece of the policy was the carbon 'tax' introduced in 2012 by the Labor government, with support from the Greens. The price was not strictly a tax, but, rather, was a permit system that functioned similarly to a tax, and was ▶

AUSTRALIA'S EMISSIONS

Australia's per capita greenhouse-gas emissions are among the highest in the world.



SOURCE: AUSTRALIAN DEPT OF CLIMATE CHANGE AND ENERGY EFFICIENCY

intended to become an emissions-trading scheme in 2015. Australia's 260 largest emitters faced a price of Aus$24 per tonne of carbon dioxide emitted, says Frank Jotzo, a climate-change economist at the Australian National University in Canberra. He says that the price is about three times that set by the current emissions-trading scheme of the European Union.

But the Coalition looks set to act on its promise to replace the system with a 'direct action' plan, which it hopes will meet Australia's target to cut greenhouse-gas emissions by 5% from 2000 levels by 2020. Direct action focuses on government payments to companies that cut their emissions below a specific level, Jotzo says.

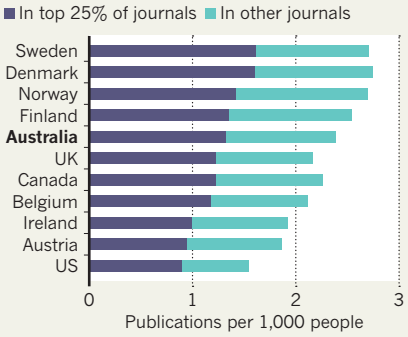
Last week, responding to the release of the IPCC's latest report, Hunt said in a statement that the report reinforces the government's "bipartisan support for the science and the targets set for emissions reductions". However, the government's stance on carbon policy and the science-policy vacuum since it came to power have fuelled fears about support for climate-science research.

"There are fears about the funding of climate-change research, particularly on mitigation."

The Coalition released few science policies during its election campaign or when it first came to power, but the scientific community was taken by surprise by Abbott's decision that his "back-to-basics" government would lack a designated minister of science. The government said that the move was aimed at simplifying ministerial and departmental titles. (Critics have pointed out that the incoming government has designated a sports minister.)

PUNCHING ABOVE ITS WEIGHT

Australia's research and development spending is moderate, but its scientific output is among the world's highest per capita.



SOURCE: AUSTRALIA'S CHIEF SCIENTIST

Responsibility for university research rests with the education ministry, but oversight of government research agencies such as the CSIRO now falls under the purview of the industry minister, Ian Macfarlane.

The industry portfolio will also include natural-resources policy, an area that will consume much of Macfarlane's time, says Kim Carr, former minister for innovation, industry, science and research, and for higher education, who is now the shadow minister for those portfolios.

"I championed the idea of building an innovation portfolio," Carr says. "It was about putting science and research at the centre of the transformation of Australian society. My concern was that if it was in the education area alone, there was a real chance that it would be marginalized. The political attention always goes to the teaching programme, not the research programme." Macfarlane was unavailable for comment.

John Rice, executive director of the Australian Council of Deans of Science, an organization that promotes the development of science in universities, says that the new portfolio configuration poses a "considerable challenge for the government in creating a strong interplay between basic research and innovation".

"I would have thought this government, more than any other, would have recognized the importance of science in supporting the economy," he says.

However, Michael Gallagher, executive director of the Group of Eight, an organization based in Canberra that represents Australia's research-intensive universities, welcomes the move. "There is a narrow culture of short-term, commercially oriented research prevailing in an industry portfolio, whereas Abbott has a broader view of what universities are about," he says.

Christopher Pyne, the education minister, says that the CSIRO and universities will continue to work closely together. "Changing the structure of portfolios will not have an impact on that," he says.

He adds that there will be no cuts to university research, but "some funding will be reprioritized for medical research". He did not respond to a question about the fate of climate-change research. ■

LAW

Uncertainty on trial

Former US drug-company chief appeals conviction for fraud over interpretation of results.

BY EWEN CALLAWAY

Once a pharmaceutical executive and socialite, Scott Harkonen now lives under house arrest and faces professional debarment. His crime: misrepresenting scientific data. But Harkonen is arguing to the US Supreme Court that he did not misrepresent anything.

Federal prosecutors convicted him in 2009 of wire fraud — using false communications to obtain money — for hyping the results of a clinical trial and encouraging the unapproved use of his now-former company's lung-disease drug. Eighteen months later, a judge sentenced him to six months' home confinement and a US$20,000 fine; in March this year, a federal appeals court upheld the conviction.

The United States' highest court will soon decide whether to hear Harkonen's final appeal. His supporters, who include statisticians, clinical researchers and legal scholars, say that his conviction relied on a poor grasp of statistics, and sets a precedent that could criminalize speculation in grant applications and papers.

"You don't want to have on the books a conviction for a practice that many scientists do, and in fact think is critical to medical research," says Steven Goodman, an epidemiologist at Stanford University in California who has filed a brief in support of Harkonen.

The US government sees the case as a warning to those who illegally promote medicines. "Mr Harkonen lied to the public about the results of a clinical trial," the lead investigator said after Harkonen's conviction.

The case centres on a clinical trial sponsored by InterMune, a company based in Brisbane, California, which Harkonen headed from 1998 to 2003. It tested whether a drug called γ-interferon, sold as Actimmune and already approved to treat a rare immune disease, helped people with idiopathic pulmonary fibrosis (IPF), an incurable lung condition.

The results were measured in terms of participants' survival and lung function — primary endpoints, or targets, that had been identified before the trial began. On 16 August 2002, Harkonen and other company executives learned that the 162 participants who had been given γ-interferon had fared no better than ▶

the 168-strong control group. Slightly fewer had died, but the difference was not deemed statistically significant, because the probability that it was not due to the drug was greater than 5%, a widely accepted statistical threshold.

It turned out that when InterMune analysed only the 254 participants who had mild and moderate IPF, the survival difference did meet this important threshold: there were 6 deaths among the 126 people on the drug, compared with 21 among the 128 people on the placebo. But the researchers had not decided to do this selective analysis before the trial, and statisticians consider such 'post hoc' analyses to be less reliable than pre-specified tests.

On 28 August 2002, the company issued a press release, approved by Harkonen, titled 'InterMune announces phase III data demonstrating survival benefit of Actimmune in IPF'. In it, Harkonen said: "We are extremely pleased with these results, which indicate Actimmune may extend the lives of patients." The company noted that the trial had failed to meet its primary endpoint, but did not say that the touted survival benefit had not been pre-specified.

SPL

Lungs affected by idiopathic pulmonary fibrosis become scarred, losing function.

GROWING CRITICISM

The press release quickly prompted concerns. Thomas Fleming, a biostatistician at the University of Washington in Seattle who had chaired the board that monitored the trial's safety, told InterMune that it was misleading, and an official at the US Food and Drug Administration (FDA) told the company that the positive results were inconclusive. The results were later published in the *New England Journal of Medicine* (G. Raghu *et al. N. Engl. J. Med.* **350,** 125–133; 2004).

In 2004, the US Department of Justice launched an investigation into allegations that Harkonen ran a campaign to promote γ-interferon to people with IPF and their doctors — illegal because the FDA had not approved that use of the drug. According to court documents, representatives had received bonuses for boosting sales of γ-interferon, which increased from US$11 million in 2000 to $141 million in 2003, largely owing to off-label prescriptions. Harkonen was charged with wire fraud for distributing "false and misleading" information in the press release, and with false labelling, a charge often used to prosecute off-label drug marketing.

Scott Harkonen says that he did not commit fraud.

The prosecution focused on proving that Harkonen knew that the claims were false and misleading. The jury heard that pre-specified endpoints are the main criteria used to judge the success of a clinical trial, and that post hoc analyses are less trusted. Harkonen was convicted of wire fraud but acquitted of false labelling.

In August this year, Harkonen's lawyers filed an appeal with the Supreme Court, contending — as they did in the original case — that freedom of speech protects the right to express scientific opinions. The US government is due to file a response by early November. If the Supreme Court then decides to take the case, it could hear the appeal in 2014.

Many physicians were encouraged by the results of the clinical trial, even though it did not meet its primary endpoint, says Joseph Zibrak, a pulmonologist at Beth Israel Deaconess Medical Center in Boston, Massachusetts. He used γ-interferon to treat some people with IPF, and says that insurance companies paid for the drug until 2007, when a follow-up trial was ended early because the drug was ineffective. The trial "sort of moved the study of the disease along quite a bit. And it certainly suggested that this was a direction we should continue to pursue," says Zibrak, whom InterMune paid to tell other physicians about his experience using γ-interferon to treat IPF. He has filed briefs in support of Harkonen's previous appeals.

Goodman, who was paid by Harkonen to consult on the case, contends that the government's case is based on faulty reasoning, incorrectly equating an arbitrary threshold of statistical significance with truth. "How high does probability have to be before you're thrown in jail?" he asks. "This would be a lot like throwing weathermen in jail if they predicted a 40% chance of rain, and it rained."

INTERPRETATION IMPLICATIONS

Gordon Guyatt, a researcher at McMaster University in Hamilton, Canada, who is not involved in the case, agrees that a clinical trial failing to meet its primary endpoint does not mean that the drug does not work. But he thinks that Harkonen skewed the findings. "This guy gave a very unbalanced presentation; whether it is sufficiently unbalanced that you should send him to jail, I don't know," he says.

Patricia Zettler, a former FDA attorney who was not involved in the case and is now a fellow at Stanford Law School's Center for Law and Biosciences, doubts that the case will make a difference to most scientists. She adds that the Supreme Court is unlikely to hear a fraud case, for which free speech is not usually protected.

Harkonen faces professional sanctions: the US government is seeking to prevent him working for companies that receive federal health funding or that develop drugs that require FDA review. In 2010, he stepped down as chief executive of Comentis, a San Francisco biotechnology company. But Harkonen tells *Nature* that he is most worried about the implications of his conviction for research. "I'm committed to going forward until the courts get the science straightened out," he says. ■

PHYSICS

Rethinking particle dynamics

Theoretical physicists are pursuing competing ways to calculate how particles interact.

BY EUGENIE SAMUEL REICH

Forty-odd years after US physicist Richard Feynman decorated his van with the wiggly diagrams that had won him a share of the 1965 Nobel Prize in Physics, the hunt is on for ways to improve on those representations.

Feynman's images depict ways in which subatomic particles interact during collisions, giving physicists an easy-to-visualize way to calculate 'scattering amplitudes': mathematical expressions that define the probabilities of various outcomes. High-energy physicists and quantum-gravity theorists alike are seeking a deeper understanding of those interactions, in work that ranges from the development of geometric ways of representing the terms used to calculate the scattering of particles, to the practical development of better algorithms.

The effort has potentially broad consequences. The emergence of a definitive set of tools would make life easier for researchers attempting to calculate the events that occur in particle colliders. And, on a more profound level, it might reveal the structure of space-time in a future quantum theory of gravity — a breakthrough that would unify modern physics.

Scrutinizing the structure of such calculations for clues to the geometry of space-time — and how that might be warped under the influence of gravity — is a relatively new trend. Research into the amplitudes of scattering particles has already allowed physicists at CERN, Europe's particle-physics laboratory near Geneva in Switzerland, to calculate complex particle interactions occurring in the centre's Large Hadron Collider (LHC)[1]. And elsewhere, quantum-gravity theorists are seeking to formulate calculations that can combine quantum mechanics, which describes small particles, with Einstein's general theory of relativity, which describes gravity.

In the 1940s, Feynman drew diagrams to depict how interactions between particles could take place — for example, two electrons scattering off each other could exchange a single photon. Today, Feynman's diagrams are being replaced by more efficient algorithms, and researchers are finding fresh geometric ways to represent terms in those algorithms.

Last year, David Skinner, a theoretical physicist at the University of Cambridge, UK, and Freddy Cachazo of the Perimeter Institute for Theoretical Physics in Waterloo, Canada, found a way to represent scattering amplitudes in a theory called 'supergravity'. This incorporates the effects of gravity and the theory of supersymmetry, in which particles are paired with heavier 'superpartners' in a way that makes calculations more tractable[2]. By doing so, Skinner says, he found that scattering amplitudes in the theory can be translated into strange geometric objects termed twistor spaces, first described by physicist Roger Penrose in the late 1960s. Skinner says that his calculations support Penrose's hunch that twistor spaces — in which light rays are represented as points — can form the basis of a quantum theory of gravity. "It's telling you something about the nature of the theory of gravity," he says.

CERN/SPL

Scientists are working to improve on Richard Feynman's representations of particle interactions.

Others are constructing alternative geometric shapes that also represent the scattering of particles. Nima Arkani-Hamed of the Institute for Advanced Study in Princeton, New Jersey, and his colleagues are preparing a paper on an object that they are calling the amplituhedron, which represents scattering amplitudes as the volume of a polygon with the same number of vertices as particles. The relationship of the object to quantum gravity is still opaque, but Arkani-Hamed hopes that the work is pointing the way to an elegant, universal geometric picture of gravity.

And there are other games in town. Calculations of scattering amplitudes usually contain a step in which all the possibilities are integrated, but in August, Pedro Vieira and his colleagues at the Perimeter Institute reported[3] a way to calculate scattering amplitudes directly, without going through the integration step. Vieira says that scattering amplitudes can be calculated from the area of a curved surface bounded by a polygon — a shape not unlike that a soap bubble might make as it was blown through a polygon-shaped hole. This is a very different geometric object to the amplituhedron, but the two are not mutually exclusive. "We are using different approaches, but both are right," Vieira says.

Such advances have practical implications. In July, Zvi Bern at the University of California, Los Angeles, and his co-workers reported their use of an algorithm that had been developed to improve on Feynman diagrams. They harnessed it to calculate the scattering of quarks and gluons in the proton beams at the LHC that could produce a W boson and five jets of particles[1]. This is three jets more than it has proved possible to calculate using Feynman diagrams alone.

Physicist Joe Incandela, a CERN spokesman, says that the calculation has enabled scientists working at the LHC to search for more subtle signals of supersymmetry in their experiments than they had initially thought possible.

Supersymmetry is a prime candidate for a theory that would go beyond today's standard model of particle physics, which explains the behaviour of most known particles and forces. However, the simplest searches — those that are sensitive to the most glaring signs of new particles — have so far failed to turn up any evidence for it in the form of 'extra' events caused by those particles' interactions. A more complex search this year for actual events in which a W boson was produced together with five jets of particles also failed to identify any such instances; these would be expected if supersymmetric particles were being created. The CERN team would next like to try to model events with six or even seven jets, Incandela says. "It's just phenomenal to be able to look at events that complicated."

Thus far, theorists have been paving the way for experimentalists. But Incandela expects that, eventually, data collected at the LHC will inform the explorations of theorists, allowing them to compare complex calculations of scattering with events actually seen in the machine. "They will need some reality check," he says. ■

1. Bern, Z. *Phys. Rev. D* **88,** 014025 (2013).
2. Cachazo, F. & Skinner, D. Preprint at http://arxiv.org/abs/arXiv:1207.0741 (2012).
3. Basso, B., Sever, A. & Vieira, P. *Phys. Rev. Lett.* **111,** 091602 (2013).

DRUG DEVELOPMENT

Pharma scrambles to fast-track drugs

'Breakthrough therapy' status is much sought after, but there is confusion about its definition and impact.

BY HEIDI LEDFORD

The experimental cancer drug ibrutinib has wowed in clinical trials, beating deadly blood cancers without the painful side effects of currently approved therapies. And it has raced through development and regulatory hurdles, in part thanks to a US programme to accelerate the development of particularly promising drugs, says its developer Pharmacyclics, based in Sunnyvale, California.

The US Food and Drug Administration (FDA) launched the 'breakthrough therapy' designation in 2012, and the label has been eagerly embraced by the pharmaceutical industry. Recent months have seen a steady stream of drugs being submitted for review. For some firms — particularly young ones — the designation can bring an extra boost of cash by raising investor confidence.

But for all the fanfare, the industry is also watching closely to see exactly what benefits can be gained by having a drug reviewed through this route. "It's like winning a beauty pageant," says Timothy Coté, a former director of the FDA's Office of Orphan Products Development who now runs a consultancy called Coté Orphan Consulting in Silver Spring, Maryland. "It doesn't have specific tangible outcomes, but it does appear to have enlivened the community."

The breakthrough therapy designation was created by the FDA Safety and Innovation Act, a law that requires the agency to fast-track promising drugs for serious or life-threatening conditions. The FDA aims to do this by meeting early and often with developers, as well as working with them to design clinical trials that deliver the needed data quickly and efficiently.

The industry leapt on the opportunity, so far submitting 99 applications for the designation. But the flurry of applications may partly be a product of confusion, says Coté: the FDA has avoided laying out detailed descriptions of what constitutes a breakthrough, and some companies are unsure of the criteria. "Most biotech chief executives with something in the clinic think that they're already there," Coté says — but 47 of the applications submitted in the past year have been denied. In most cases the denials are due to insufficient clinical data, the FDA says.

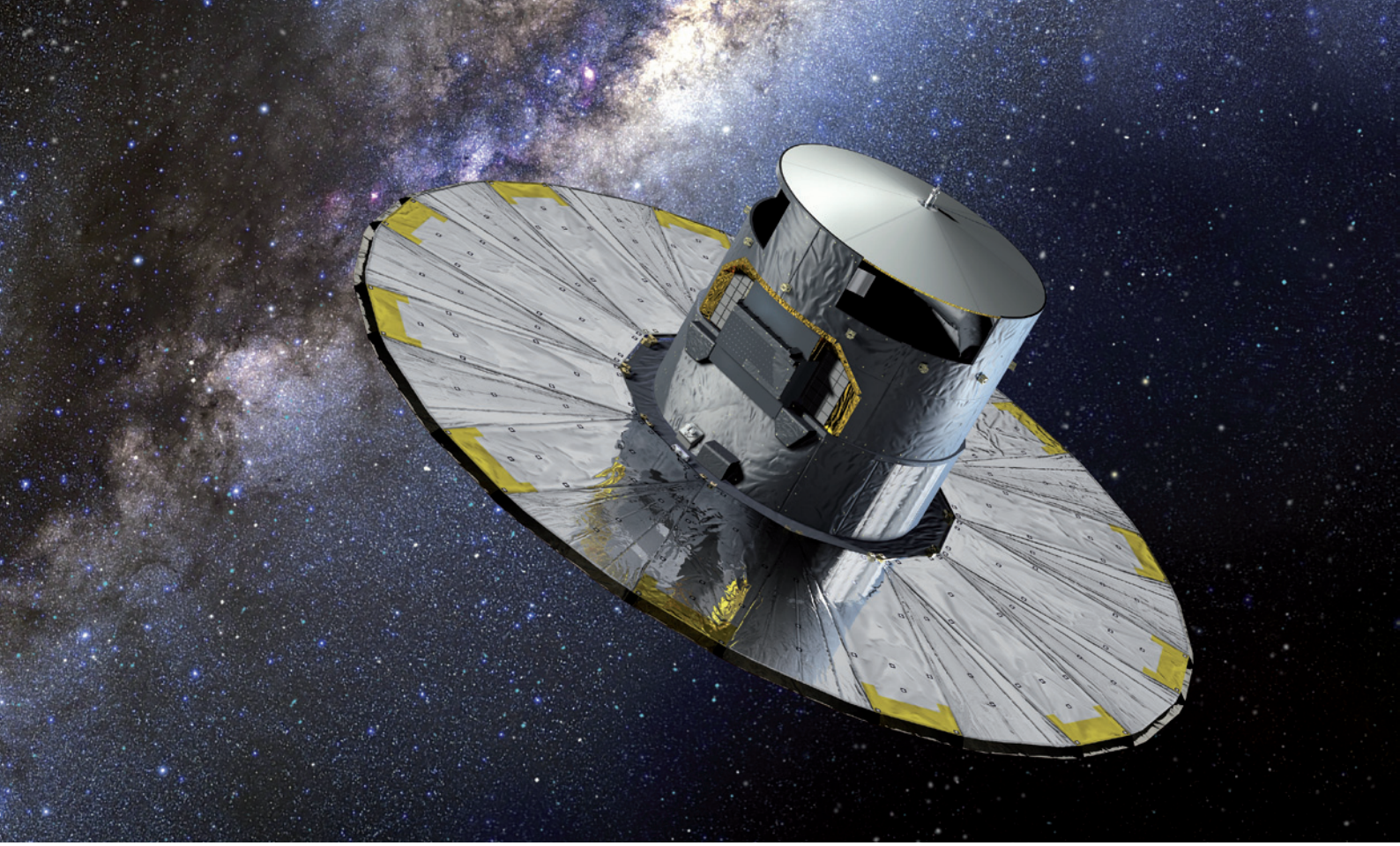
Although the lack of clear guidelines could be deemed confusing, the FDA's avoidance of hard-and-fast criteria can actually be an advantage for some drugs, says Keith Flaherty, an oncologist at Massachusetts General Hospital in Boston. He was pleased, for example, to see the FDA bestow breakthrough status on a melanoma therapy called lambrolizumab. The drug, which is made by Merck (based in Whitehouse Station, New Jersey), is one of several in development that stimulate the immune system to fight cancer by blocking a protein called PD1. Lambrolizumab works in only about 38% of patients, which is well below the response rate for some other cancer drugs in development. However, doctors champion it because it has tolerable side effects and can yield unusually long-lasting responses. "Having it get that designation really put a spring in the step of many people in our community," says Flaherty. "It showed us that the FDA really gets the importance of these drugs."

"The FDA has avoided detailed description of what constitutes a breakthrough."

There are lingering concerns that other aspects of the drug-development process might delay the ultimate impact of the breakthrough-designated compounds. Jeff Allen, executive director of the patient-advocacy group Friends of Cancer Research in Washington DC, notes that drugs are increasingly developed alongside medical tests that will select the patients who are most likely to benefit from them. The new law does not address the development of such tests, but unless their evaluation and approval is accelerated, a breakthrough drug — even if approved — may not achieve its full potential in the clinic, he says.

Coté, despite being a fan of the programme, says that it might not have much of an impact on review times because the FDA has always prioritized promising applications. The biggest benefit, says Steven Grossman, founder of the consultancy HPS Group in Silver Spring, might be for small companies that can use the designation to get the FDA's attention earlier in development than they normally might.

Financial analysts expect ibrutinib to be approved by the end of the year, and Coté thinks that most of the breakthrough designees will ultimately prevail. "If the FDA likes what you're doing, that can't be a bad thing," he says. ■



The Gaia mission (artist's impression) will pinpoint the locations of up to 1 billion stars.

EUROPE'S STAR POWER

The Gaia spacecraft will soon launch on a mission to chart the heavens in unprecedented detail.

BY DEVIN POWELL

The century-old brass telescope was broken in places and long past its useful life — but it captured Lennart Lindegren's heart.

Forty years ago, when Lindegren was a graduate student at the Lund Observatory in Sweden, he fell in love with the elaborate, once cutting-edge technology that had allowed nineteenth-century astronomers to track and time the motion of the stars. The telescope had an ingenious mechanical stopwatch — originally invented to time race horses — and a large metal wheel that could adjust its angle. "I got so fascinated by the beauty of the instrument that I wanted to get it working again," says Lindegren, who is now on the Lund Observatory's staff.

He might as well have fallen in love with a sundial. Astrometry — mapping the locations and movements of celestial objects — was once a central concern in astronomy, with roots going back to ancient Babylon and China. But by the 1970s it had long fallen out of fashion. Astronomers had just about reached the limit for improving the precision of such measurements taken from the ground, and most had moved on to other questions. Astrometry, says Lindegren, "was not regarded as a field that would offer any great prospects for young scientists".

He eventually gave up on repairing the telescope, but never abandoned the idea of reviving astrometry. Better star maps, he argued, could help astronomers to answer some fundamental questions, from how the Milky Way evolved to what makes up the dark matter that accounts for most of the Universe's mass. All researchers would need to do would be to get their astrometric instruments into space, above Earth's turbulent atmosphere, which subtly distorts starlight and limits the precision of measurements.

In November, a proposal by Lindegren and like-minded scientists will bear fruit when the European Space Agency (ESA) launches Gaia: an astrometric mission that required many compromises and 13 years to complete, and will cost about €1 billion (US$1.4 billion). Gaia will make observations for the next 5 years; the results will extend the reach of high-precision maps from the roughly 2.5 million stars near Earth to at least 1 billion stretching to the edge of the Milky Way or beyond. For an estimated 10 million of those objects, Gaia's map will be fully

D. DUCROS/ESA

SOURCE: CSIRO

three-dimensional: the spacecraft will measure not just the stars' locations on the sky, but also their distances from Earth, accurate to less than 1%. For now, the distances to only a few hundred stars are known at this level of precision.

Michael Perryman, an astronomer at the University of Bristol, UK, and former project scientist for the mission, is optimistic. "Gaia will be tremendous and transformational, a huge leap forward both in terms of the number of stars measured and the accuracy of those measurements," he says.

STELLAR CARTOGRAPHY

The keen eyesight that will make this leap possible starts with Gaia's digital camera, which uses light-gathering sensors similar to those found in consumer cameras — but 106 of them, providing a resolution of more than 900 megapixels. By contrast, the main camera on NASA's Hubble Space Telescope has two sensors with a resolution of just over 16 megapixels.

Guiding starlight into the camera are two telescopes that point 106.5° apart, to take in a wide field of view. As the spacecraft spins, completing a full revolution once every 6 hours, that view will sweep across the same stars, month after month. Each star will be photographed about 70 times, producing roughly twice as much imaging data in 5 years as Hubble generated during its first 21 years in orbit.

When all the data have been analysed, they will provide a pair of coordinates for each star, pinpointing its position in the sky with an error as small as 6 microarcseconds — the size of a small coin sitting on the Moon as viewed from Earth. That is hundreds of times better than today's best catalogue, and millions of times better than the first known Western star atlas, compiled more than 2,000 years ago through naked-eye observations by the ancient Greek astronomer Hipparchus of Nicaea.

Finding a star's position in three dimensions will require further measurements. Because of a geometric phenomenon known as parallax (see 'The parallax effect'), stars appear to move from side to side as Earth orbits the Sun. The closer a star is to Earth, the larger its apparent movement, for much the same reason that trees on the side of a road seem to whiz past a speeding car, whereas a mountain in the distance barely seems to move at all. If astronomers can measure that side-to-side motion precisely, simple geometry will allow them to use the known size of Earth's orbit to calculate the star's distance.

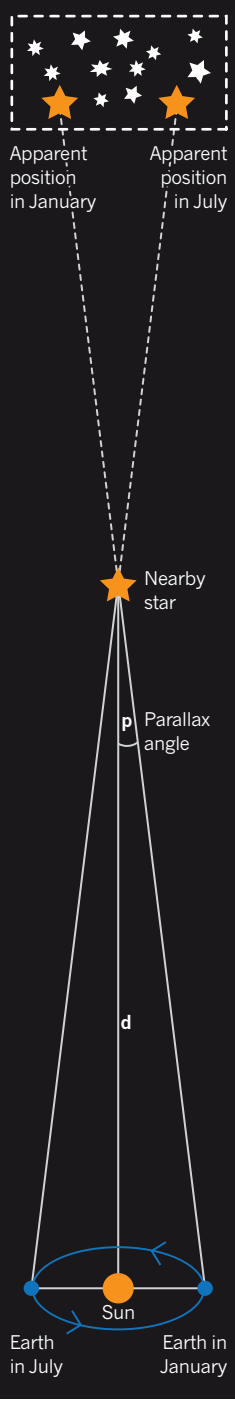
Atmospheric turbulence so compromises such efforts that even the best modern ground-based visible-light telescopes can see parallaxes up to only about 100 parsecs (a few hundred light years). Radio telescopes are less affected, so they can see much farther — but only for objects that emit strong radio waves. From its place outside the atmosphere, Gaia, which is destined for a stable orbit that will remain fixed relative to both the Sun and Earth, should be able to obtain parallax measurements for stars up to about 10,000 parsecs away.

The same precision should let it measure a star's 'proper motion' across the sky at even greater distances. Proper motion — the result of a star's actual movement through space perpendicular to the line of sight — will show up as a steady sideways drift in the star's position, superposed on its annual side-to-side motion.

Finally, Gaia should be able to use changes in the spectrum of light emitted by each star to measure the star's velocity towards or away from Earth. The result will be a complete portrait of the star's position and velocity in three-dimensional space.

THE PARALLAX EFFECT

As Earth travels around the Sun, nearby stars seem to move back and forth relative to distant ones. Using simple geometry, the known size of Earth's orbit and a measurement of the parallax angle (p), astronomers can determine the star's distance (d). For an angle of 1 arcsecond, the distance is 1 parsec (3.26 light years.) Smaller angles mean larger distances.



ASTROMETRY REBORN

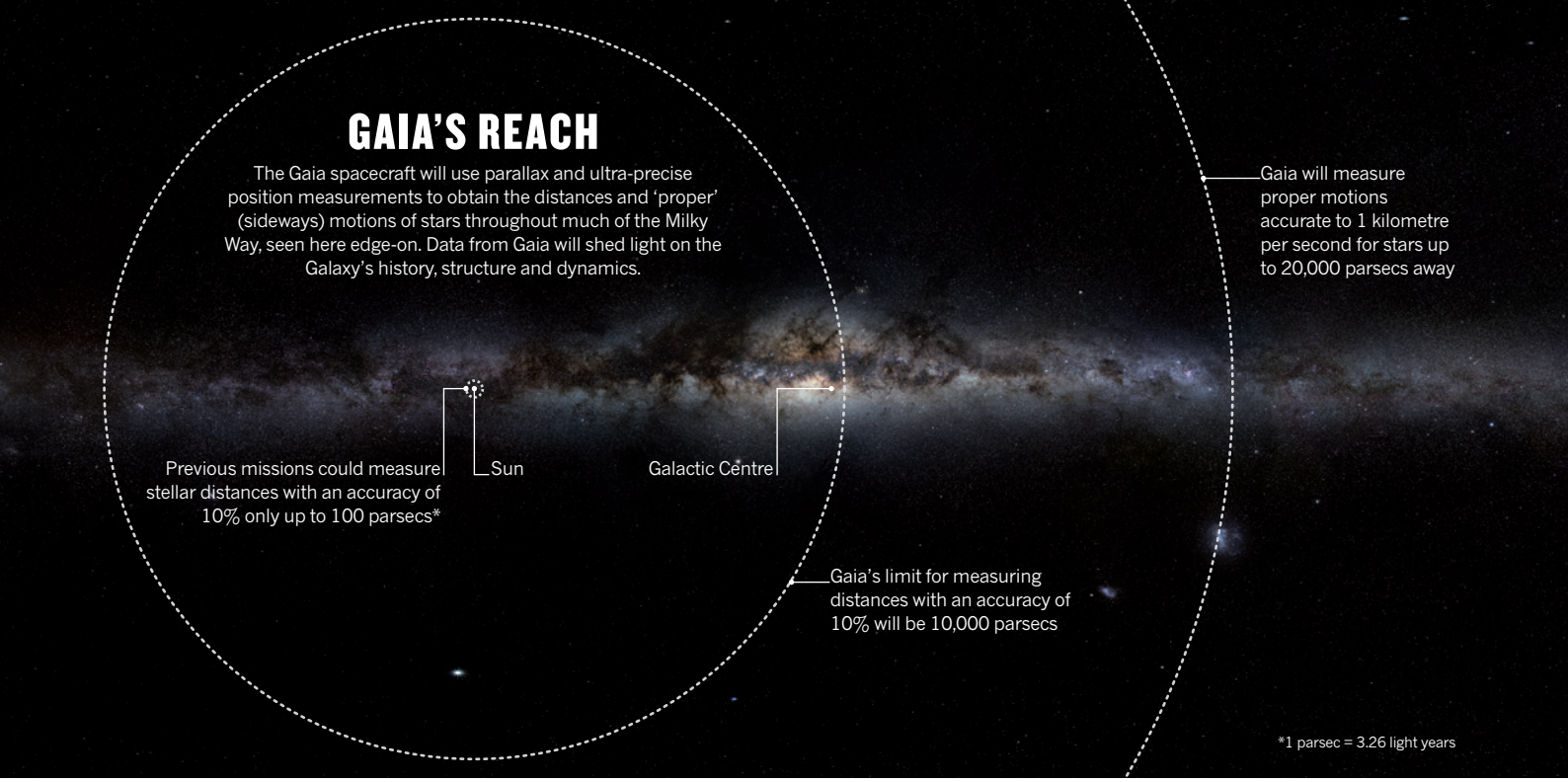
Gaia is not the first high-precision space-astrometry instrument. The feasibility of the exercise was demonstrated two decades ago by Gaia's predecessor, ESA's High Precision Parallax Collecting Satellite (Hipparcos). Launched in 1989, the €580-million spacecraft ran into trouble almost immediately, when a failed booster rocket left it in the wrong orbit. But even so, it worked: by the time the mission ended in 1993, Hipparcos had provided the distances to about 118,000 stars. Of those, some 400 were measured with an error of 1%. Only 50 had been measured so well from the ground. The Hipparcos star catalogue is still the best one available.

However, the mission focused on relatively nearby stars, says Shrinivas Kulkarni, an astrometer at the California Institute of Technology in Pasadena, so the Hipparcos catalogue was more an evolution in the science than a revolution. "It showed us that our basic understanding of stars was sound," he says. As a proof of principle, says Erik Høg, an emeritus astronomer at the Niels Bohr Institute in Copenhagen who drew up the first blueprints for the mission, "Hipparcos's success was really invigorating for astrometry. People could gather around this project."

What they could not do, until now, was get a worthy successor off the ground. One after another, advanced astrometry missions were proposed and then failed because they overran their budgets. One of the most spectacular examples was NASA's Space Interferometry Mission (SIM). It had a deliberately narrow focus: it would have concentrated on pinning down the positions and movements of a mere 10,000 stars located fairly nearby. The trade-off was that SIM could have detected wobbles in those stars caused by the gravitational pull of orbiting planets as small as Earth. But the project was postponed several times, and its original $600-million budget ballooned. After NASA had already spent hundreds of millions of dollars on development, projections suggested that the mission would need a further $1.2 billion and SIM was cancelled in 2010. "The money just wasn't there," says Michael Shao, an astronomer at NASA's Jet Propulsion Laboratory in Pasadena and formerly project scientist for SIM.

Some researchers argue that Gaia will succeed where SIM and others failed because of Europe's style of building spacecraft. "In the European system," says Ken Seidelmann, an astronomer at the University of Virginia in Charlottesville, "the scientists write down the specifications they want, and the contractors come up with cheaper ways of doing things" — ways that sometimes involve cutting back the mission's capabilities. "In the United States, the scientists tend to stay more involved," he adds — and if they refuse to compromise on the objectives, costs can skyrocket.

In Gaia's case, keeping close to the budget set when ESA approved the mission in 2000 required a series of downgrades that substantially reduced the capabilities of the original design, mainly by halving the expected



PICTURE: S. BRUNIER/ESO; GRAPHIC SOURCE: ESA

accuracy of its parallax measurements. The cuts meant that some of the problems that Gaia had intended to tackle were now out of reach. For example, the mission would no longer be able to track potentially hazardous near-Earth objects such as asteroids well enough to predict their motion for the next century — a goal that had been named a top priority by a task force led by the UK minister of science. The downgrades led Perryman to quit the project in 2006, after six years as leader of the scientific team. "I was enormously frustrated by the decision to de-scope this project, which was not made on scientific grounds," he says.

But Gaia survived and is now scheduled for launch as early as 20 November. With a bit of distance and a graveyard of space astrometry missions to reflect on, Perryman now expects big things from the mission.

GAIA GOES GALACTIC

To start with, the cosmic census begun by Hipparcos will continue and expand. Millions of new binary stars are expected to show up, as are tens of thousands of brown dwarfs: 'failed' stars too small to ignite by fusing hydrogen. Gaia should also discover 1,000 Jupiter-sized planets — or, rather, the wobbles these objects cause in nearby stars. Closer to home, the spacecraft will get at least some data on the hundreds of thousands of Solar System asteroids expected to cross its field of view.

Where Gaia will really shine, however, will be in extending astrometry across the Milky Way (see 'Gaia's reach'). "Our unique science goal is to unravel the structure and dynamics and history of our Galaxy," says Jos de Bruijne, a systems scientist at ESA's European Space Research and Technology Centre in Noordwijk, the Netherlands, and Gaia's deputy project scientist.

Astronomers already know the basics, he says. The Milky Way is shaped something like a fried egg, with a bulge of stars in the middle surrounded by a flat stellar disk that tapers at the edges and contains the Galaxy's spiral arms. Around the disk is a diffuse sphere of old stars called the halo. But astronomers are not certain how these structures formed, or in what order (see *Nature* **490,** 24–27; 2012). Gaia will provide one important set of clues by measuring stellar composition and brightness — data that will reveal for the first time when many stars formed, and will help astronomers to work out the ages of the Galaxy's different parts.

Another set of clues will come with Gaia's measurements of stellar movements, which astronomers can extrapolate back in time to find out how the Galaxy has evolved. Typically this is difficult because tiny errors quickly accumulate into large uncertainties. "Exactly how far back we can get is very much an open question," says Lindegren. But the high precision of Gaia's measurements will certainly take the extrapolation much further than before.

➲ NATURE.COM
For more about the structure and history of the Milky Way, see:
go.nature.com/di7phk

The measurements will also help to illuminate the many episodes of violence in the Milky Way's history. The Galaxy has grown by cannibalizing other, smaller galaxies; when they got too close, the Milky Way's gravity ripped them apart into long streams of stars that were then pulled towards the galactic centre at various angles. One such stream, torn from a dying object known as the Sagittarius dwarf galaxy billions of years ago, was found in 2002. "There are other streams out there that encode information about how the Galaxy has been developing," says Andrew Gould, an astronomer at the Ohio State University in Columbus. "Gaia will discover those streams" — and use its measurements of stellar motions to reveal how the dismemberments unfolded.

Knowing precise stellar movements should also help researchers to map out the distribution of invisible dark matter, which permeates the whole Galaxy. Dark matter emits no light, but it exerts a gravitational pull on stars, causing perturbations that should reveal themselves in Gaia's data. Those will allow astronomers to test how clumpy the dark matter is and whether it forms into disks, as theorists have proposed.

Whatever Gaia finds, one thing seems certain: its star catalogue, due to be published in 2021, will remain unsurpassed for decades. ESA is considering a planet-hunting spacecraft similar to SIM for a future mission but has yet to choose a successor to carry on Gaia's astrometric work. "We have to start thinking about it now if we want to realize something in 15 years," says Lindegren. "But we don't really know what exactly is the best way to proceed." Boosting the precision significantly would be an enormous technological challenge. An easier path would be to fly another Gaia mission with the same specifications in 20 years, after the stars have moved noticeably, to better pin down their positions and velocities.

Another proposed follow-on mission would examine parts of the Milky Way to which Gaia will be blind. Dust will obscure the Galaxy's bulge and some far-away parts of its disk from Gaia's visible-light gaze — but would pose no problem to an instrument looking for infrared radiation.

Or perhaps Gaia itself will upend the whole discussion. As astrometry sharpens its focus, there is always the exciting possibility that something wholly unexpected could be found. "Science often progresses by making detailed measurements," says Kulkarni. "Sometimes you see a deviation — something that turns out to be profound." ■

Devin Powell *is a freelance writer currently based in Singapore.*

TABOO GENETICS

Probing the biological basis of certain traits ignites controversy. But some scientists choose to cross the red line anyway.

BY ERIKA CHECK HAYDEN

Growing up in the college town of Ames, Iowa, during the 1970s, Stephen Hsu was surrounded by the precocious sons and daughters of professors. Around 2010, after years of work as a theoretical physicist at the University of Oregon in Eugene, Hsu thought that DNA-sequencing technology might finally have advanced enough to help to explain what made those kids so smart. He was hardly the first to consider the genetics of intelligence, but with the help of the Chinese sequencing powerhouse BGI in Shenzhen, he planned one of the largest studies of its kind, aiming to sequence DNA from 2,000 people, most of whom had IQs of more than 150.

He hadn't really considered how negative the public reaction might be until one of the study's participants, New York University psychologist Geoffrey Miller, made some inflammatory remarks to the press. Miller predicted that once the project turned up intelligence genes, the Chinese might begin testing embryos to find the most desirable ones. One article painted the venture as a state-endorsed experiment, selecting for genius kids, and Hsu and his colleagues soon found that their project, which had barely begun, was the target of fierce criticism.

There were scientific qualms over the value of Hsu's work (see *Nature* **497,** 297–299; 2013). As with other controversial fields of behavioural genetics, the influence of heredity on intelligence probably acts through myriad genes that each exert only a tiny effect, and these are difficult to find in small studies. But that was only part of the reason for the outrage. For decades, scientists have trodden carefully in certain areas of genetic study for social or political reasons.

At the root of this caution is the widespread but antiquated idea that genetics is destiny — that someone's genes can accurately predict complex behaviours and traits regardless of their environment. The public and many scientists have continued to misinterpret modern findings on the basis of this — fearing that the work will lead to a new age of eugenics, preemptive imprisonment and discrimination against already marginalized groups.

"People can take science and assume it is far more determinative than it is — and, by making that assumption, make choices that we will come to regret as a society," says Nita Farahany, a philosopher and lawyer at Duke University School of Law in Durham, North Carolina.

But trying to forestall such poor choices by drawing red lines around certain areas subverts science, says Christopher Chabris of Union College in Schenectady, New York. Funding for research in some areas dries up and researchers are dissuaded from entering promising fields. "Any time there's a taboo or norm against

ILLUSTRATION BY OLIVER MUNDAY

studying something for anything other than good scientific reasons, it distorts researchers' priorities and can harm the understanding of related topics," he says. "It's not just that we've ripped this page out of the book of science; it causes mistakes and distortions to appear in other areas as well."

Here, *Nature* looks at four controversial areas of behavioural genetics to find out why each field has been a flashpoint, and whether there are sound scientific reasons for pursuing such studies.

1 INTELLIGENCE

TABOO LEVEL: HIGH

The comments that Miller made about Chinese families and the government wanting to select for intelligent babies touched a nerve still raw after many years. In the nineteenth century, British anthropologist Francis Galton founded the eugenics movement on the premise that extraordinary abilities, as well as deficits, were inherited. The movement led to abuses, such as forced sterilization of people deemed mentally inferior — generally minorities, poor people and especially people with mental illnesses — in countries around the world, including Germany, the United States, Belgium, Canada and Sweden.

The term 'intelligence' is also a slippery one. Intelligence tests don't measure a wholly innate ability; it is possible, for example, to improve one's scores with practice. Nevertheless, about 50% of variability in intelligence seems to be inherited, posing an irresistible puzzle to some researchers. No one gene has been linked strongly to intelligence and many that have been weakly linked have also been questioned[1].

Earlier this year, in an attempt to find stronger genetic correlations, Chabris and a large international group of colleagues examined the genomes of more than 125,000 people and found three genetic variants, each of which had a small effect on the length of an individual's school career[2]. The authors speculated that the variants' influence on educational attainment came from their effect on intelligence. But the results triggered the usual rounds of condemnation and concerns over eugenics. Other detractors argued that such studies take the focus and funding away from other, non-genetic, factors such as poverty, which have a much greater effect on social mobility.

Chabris says that the work can actually contribute to greater social mobility — for instance, by helping to identify preschoolers who could be helped by more intensive early childhood education. "The fact that people in the past interpreted the results in a certain way doesn't mean that it shouldn't be studied," he says. But not everyone buys that potential misuses of the information can be divorced from gathering it. Anthropologist Anne Buchanan at Pennsylvania State University in University Park wrote on the blog The Mermaid's Tale that rather than being purely academic and detached, such studies are "dangerously immoral".

➲ NATURE.COM
Vote on whether this kind of research should be off limits:
go.nature.com/n4llub

Critics of the BGI project also point to signs that its data could be misused. After this summer's furore over Miller's interview, Hsu played down the potential for abuse. "There's a big gap between finding a few hits and finding thousands of hits — enough to predict the trait on the basis of the genotype — and we were never saying we were going to get to that point," he says. But in 2011, before the uproar over the study, Hsu told *Nature*: "I'm 100% sure that a technology will eventually exist for people to evaluate their embryos or zygotes for quantitative traits, like height or intelligence. I don't see anything wrong with that."

One of Hsu's collaborators, behavioural geneticist Robert Plomin of King's College London, says that such talk has not been helpful. But after studying intelligence for 40 years, he has high hopes that this project and other sequencing ventures will help to pinpoint the many genetic contributors to the trait. Like Chabris, he says that the work could be used to target educational interventions. Moreover, like all of the intelligence researchers interviewed for this story, he says it is a fundamentally human trait and that it is worth searching for a genetic contribution. "I'm optimistic that we will find it," he says. "I'm not going to quit until we do."

2 RACE

TABOO LEVEL: VERY HIGH

As far as genetic taboos go, race is probably one of the most heavily policed from within the scientific community, largely because of the way researchers have examined its intersection with other controversial traits, such as intelligence. This is due mostly to suspicion about what motivates the study. There is broad consensus across the social and biological sciences that groups of humans typically referred to as races are not very different from one another. Two individuals from the same race could have more genetic variation between them than individuals from different races. Race is therefore not a particularly useful category to use when searching for the genetics of biological traits or even medical vulnerabilities, despite widespread assumptions.

Most researchers who examine genetic differences between populations take care to point out that the differences they observe reflect the geographic origins, reproductive history and migrations of these groups, not markers of some essential differences between them.

However, some researchers have asked whether the taboo on the genetics of race has become so severe that it bars legitimate research. In 2005, for instance, geneticist Bruce Lahn of the University of Chicago in Illinois published studies[3,4] suggesting that variants of two brain-development genes possibly linked to intelligence are evolving differently in white Europeans and African ethnic groups. This provoked a wave of worried comments by scientists about how the studies might be interpreted. Among those who voiced concerns was then-director of the US National Human Genome Research Institute Francis Collins, now director of the National Institutes of Health (NIH) in Bethesda, Maryland.

Lahn and his co-authors eventually found that the gene variants under selection were not linked to elevated intelligence[5]. But that report garnered little attention compared with the explosive studies that came before it. Lahn says he felt "ambushed" during the debate over his findings. At meetings, even his co-authors did not defend him. "My friends said nothing," he says.

Some argue that Lahn should have been more cautious. "Science always plays out in a certain socio-political context, and you have to look at the consequences of how the science might play out," says John Horgan, a journalist who has written widely on the societal implications of science. "Research on race and intelligence is much more prone to supporting racist ideas about the inferiority of certain groups, which plays into racist policies." Horgan says that institutional review boards should ban or seriously question proposed studies on race and IQ.

Lahn no longer works on the genetics of race and has urged researchers to have a more transparent discussion about whether such studies should proceed at all. "Given the history of the way race has been used in this country, maybe the research shouldn't be encouraged because it touches too many raw nerves. I'm OK with that," he says. "But I'm not OK with being ambushed by political discussions masquerading as scientific discussions."

3 VIOLENCE

TABOO LEVEL: MILD

A decade ago, forensic psychiatrist Tracy Gunter of Indiana University in Indianapolis was spending her time trying to help people to overcome the behavioural and substance-abuse disorders that had led to their entanglement in the criminal-justice system. But it was becoming increasingly clear to her that

once a client fell into an abuse–crime spiral, it was very difficult to bring them back.

It was around this time that researchers reported that people with a certain version of a gene called monoamine oxidase A (*MAOA*) have some protection from the effects of childhood abuse[6]. Other people who express low levels of the protein it encodes are more likely to commit crimes if mistreated.

Gunter switched fields to work in behavioural genetics, hoping to find ways to identify and preemptively treat high-risk individuals. She soon found her work complicated by the difficulty of defining criminal behaviour precisely; the impossibility of separating environmental and innate influences; and, again, the emerging consensus that behaviour is influenced by numerous small genetic factors. Ten years on, she says, "the simplistic notions I had about behavioural genetics when I started this work are not true".

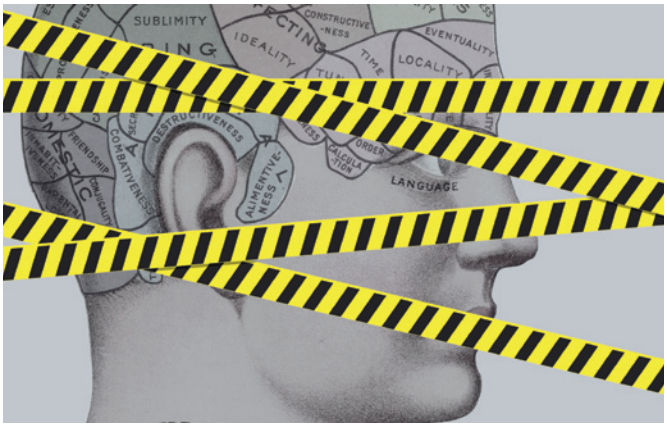
Despite these caveats — and the fact that some studies have failed to replicate the original *MAOA* finding[7] — some lawyers have used *MAOA* gene tests, combined with history of childhood abuse or life stress, to try to mitigate sentences.

In 2009, such testing led to a lesser charge for a Tennessee man who killed his wife's friend, and it convinced a judge in Italy to reduce a murderer's sentence by one year (see *Nature* http://doi.org/cttbjt; 2009). But juries are often underwhelmed by genetic testimony: in the United States in 2008, for instance, defence lawyers attempted to convince a jury to be lenient towards a boy who had shot a bus driver. They presented evidence that the boy had a variant of a promoter of a serotonin transporter gene, *SLC6A4*, that is linked to depression in people under stress. The jury found the boy guilty of first-degree murder anyway. Outcomes are mixed, Farahany says, perhaps because the research is so oblique. "It doesn't seem to be enough to persuade judges or juries to change guilt or sentencing," she says. William Bernet, a forensic psychiatrist in Nashville, Tennessee, adds that, "a genetic result does not directly cause a person to behave in a particular way. Juries seem to understand this".

That may change as the science progresses, but so far genetics has held no more sway than conventional mitigating factors, which often include the milieu in which a person grew up.

Those two domains are coming together as researchers look for more clues to the environmental factors that interact with genetics in influencing behaviour. Gunter was part of a team that showed that certain epigenetic modifications on the *MAOA* gene are linked to substance abuse in adult women[8], and these modifications are influenced by a history of smoking. "Every year that I work in this field has been a lesson that it's not just genes or environment," she now says. "It's genes and environment that matter."

Scientists continue to look at the genetics of violence, and of conditions such as psychopathy, although the tension between those who focus on just genes and those looking for genetic and environmental contributors is high, says James Tabery, a philosopher at the University of Utah in Salt Lake City. "My sense is that we're in a holding pattern; it's not clear what's going to happen next" — specifically because not many genes have been linked to violence and attempts to replicate the *MAOA* findings have produced mixed results.



"If we're worried that people will misuse science, we need to create safeguards — and an open public dialogue that ensures responsible use."

4 SEXUALITY

TABOO LEVEL: MILD

Sometimes, shifting political winds can destigmatize research. In 1993, for instance, geneticist Dean Hamer, then at the US National Cancer Institute in Bethesda, encountered a firestorm of criticism from political conservatives when he published a report suggesting that a region of the X chromosome might be linked to homosexuality[9]. Scientists also criticized some aspects of his work. Today, studies on the genetics of sexual orientation have been embraced by the US gay community. The successful campaign to strike down a 2008 California ballot measure that banned same-sex marriage enlisted evidence that homosexuality has some basis in genetics. And the NIH has designated research on lesbian, gay, bisexual, transgender and intersex people a priority. "The tables have turned tremendously," says geneticist Eric Vilain, director of the Institute for Society and Genetics at the University of California, Los Angeles.

But that does not mean that all research into the genetics of sexuality will be equally welcome, he adds. Vilain, for example, wants to study the epigenetics of homosexuality, in search of environmental influences that might affect the trait. The work hasn't been funded, but he predicts that if it is, it could upset some gay rights activists who have seen their cause benefit from the 'hardwiring' theory. He is keeping his fingers crossed. "I hope that now that there have been significant social advances, that scientists can do their work in peace," he says.

Such complexities are unavoidable in a democratic society in which citizens have a say on how public money is spent. Researchers must acknowledge that and take part in the broader conversation about the kinds of topics they want to pursue, Farahany says. "You hear this refrain in lots of areas of science, that because people will misuse science we shouldn't engage in scientific inquiry. I think that gets it backwards. If we're worried that people will misuse it, we need to create safeguards — and an open public dialogue that ensures responsible use." That, rather than censoring science or ignoring its implications, is perhaps the only way that Vilain and other researchers will get their wish: to do their work in peace. ■ SEE EDITORIAL P.5

Erika Check Hayden *reports for* Nature *from San Francisco, California.*

1. Chabris, C. F. *et al. Psychol. Sci.* **23,** 1314–1323 (2012).
2. Rietveld, C. A. *et al. Science* **340,** 1467–1471 (2013).
3. Evans, P. D. *et al. Science* **309,** 1717–1720 (2005).
4. Mekel-Bobrov, N. *et al. Science* **309,** 1720–1722 (2005).
5. Mekel-Bobrov, N. *et al. Hum. Mol. Genet.* **16,** 600–608 (2007).
6. Caspi, A. *et al. Science* **297,** 851–854 (2002).
7. Haberstcik, B. C. *et al. Biol. Psychiatry* http://dx.doi.org/10.1016/j.biopsych.2013.03.028 (2013).
8. Philibert, R. A., Gunter, T. D., Beach, S. R. H., Brody, G. H. & Madan, A. *Am. J. Med. Genet. B* **147B,** 565–570 (2008).
9. Hamer, D. H., Hu, S., Magnuson, V. L., Hu, N. & Pattatucci, A. M. *Science* **261,** 321–327 (1993).

ILLUSTRATION BY OLIVER MUNDAY

COMMENT

JUSTIN SULLIVAN/GETTY

US school children shelter under their desks during an earthquake drill.

Seconds count

The United States should install an earthquake early-warning system now — and before the next big one hits, says **Richard Allen**.

The United States will be hit by a major earthquake causing hundreds of fatalities within my lifetime. Whether it will be in 20 years, 2 years, or tomorrow, we do not know. Such an event will prompt the country to implement a public earthquake early-warning system, giving people seconds or minutes to prepare for the shaking. Rather than waiting until the next big quake galvanizes political action, I believe that we must build an alert system now.

Earthquake early-warning technology is proven. Japan leads the way. When the magnitude-9 Tohoku-Oki earthquake hit the northeast of the country in March 2011, an automatic warning was issued within seconds. Trains stopped, students took shelter under desks, sensitive manufacturing equipment was paused and hazardous chemicals were isolated. Lives and money were saved. China, Taiwan, Mexico, Turkey and Romania also issue earthquake alerts.

The US public deserves access to warnings too. For the past 2 years, California has run a successful demonstration system — providing seismic alerts to 36 organizations, including the Bay Area Rapid Transit rail network. Now, the system must be made more robust and rolled out to the public, first across the West Coast and then nationwide. The benefits for security, business and science are manifold.

The cost of an initial West Coast system is reasonable: US$120 million for the first 5 years to build and operate it, and a further $16 million a year to run it. This is roughly twice the current earthquake-monitoring budget for the region. Private enterprise could deliver the alerts and tailored services. Seismology would also benefit from the hundreds of extra sensors that would need to be installed along high-hazard faults — those most likely to slip.

The main obstacle is political will. Politicians, business leaders and agency administrators need to recognize the significance and urgency of seismic risk and implement an early-warning system before the next ▶

big quake costs lives. Other regions, including Europe, should follow.

INEVITABLE HAZARD

The probability that a major earthquake will hit the western United States in coming decades is high. California has a 99% chance of experiencing a quake of at least magnitude 6.7 (the size of the 1994 Northridge earthquake near Los Angeles) in the next 30 years. Both San Francisco and Los Angeles have a two-in-three likelihood of such an event. There is a 50% chance that the next big Bay Area quake will be on the Hayward Fault, which is situated about 500 metres away from the seismological laboratory at the University of California, Berkeley, where I work.

The Pacific Northwest region must be prepared for even bigger earthquakes, measuring up to magnitude 9 — similar to the one that hit Japan in March 2011. The hazard across the rest of the United States is lower, but damaging earthquakes can also occur all the way to the East Coast, as illustrated by the August 2011 Virginia quake of magnitude 5.8 that rattled Washington DC and New York.

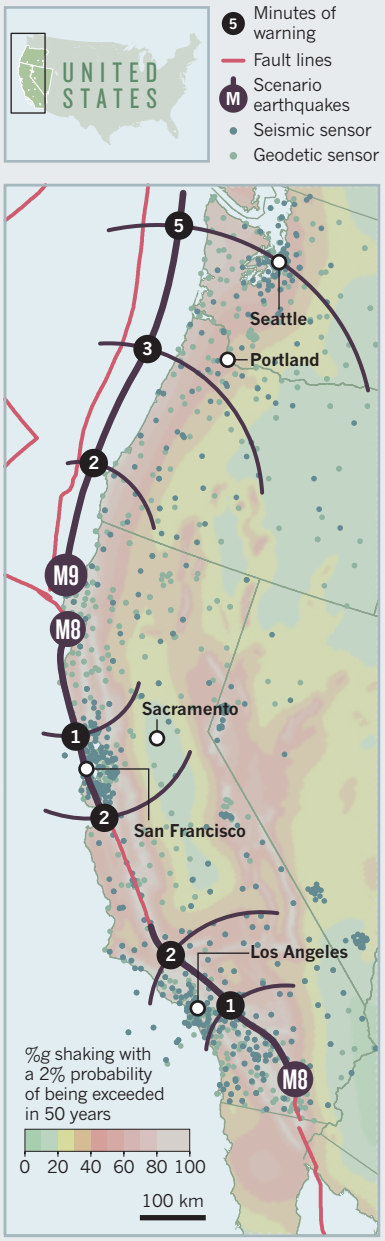
The first line of defence in the United States is a robust building code to prevent structures from collapsing. But now, the information revolution allows us to develop real-time responses to minimize casualties and damage. When seismic sensors pick up the first vibrations of a rupturing fault, automatic alerts can be issued within seconds to give people up to 5 minutes to react, depending on their distance from the epicentre[1].

Japan has pioneered such systems since the 1995 Kobe earthquake, which killed more than 6,000 people. The government invested billions of yen in seismic and geodetic networks to detect quake signals. In 2004, the Japan Meteorological Agency tested a limited earthquake-warning system. It delivered its first alert in 2005, and in 2007 the system went national and public. The first true test came in the Tohoku-Oki earthquake. Sendai, the closest major city to the epicentre, received a 15-second warning.

California's demonstration system, ShakeAlert[2], has been operating since 2011 but is yet to go public. Using existing seismic sensors, it detects earthquakes daily and, when magnitudes exceed 2.5, issues alerts to a limited group of organizations mainly involved in transport, manufacturing and emergency response. It is largely a public-sector and academic enterprise: a collaboration that includes the California Integrated Seismic Network, with researchers and funding provided by the University of California, Berkeley, the California Institute of Technology in Pasadena, the US Geological Survey (USGS), the Southern California Earthquake Center and the Swiss Federal Institute of Technology in Zurich. In 2012, the scheme expanded to include the Pacific Northwest, adding the University of Washington in Seattle, and gained support from the Gordon and Betty Moore Foundation, a private grant-making organization in Palo Alto, California.

US WEST COAST HAZARD MAP

Public alerts of large earthquakes could give the populations of major cities along the West Coast several minutes advance warning.



Although the California system has not yet been tested by a large earthquake, it successfully gave organizations in Pasadena a 5-second warning of ground shaking for a magnitude-4.2 earthquake in September 2011. In the San Francisco Bay Area, several small (magnitude-3) earthquakes located close to the epicentre of the 1989 Loma Prieta quake were detected, and alerts were provided around 20 seconds before peak ground motion in San Francisco, Oakland and Berkeley — illustrating what would be possible in a repeat of the devastating 1989 quake.

The detection algorithms have performed well so far, with few false alerts and no cases of small earthquakes being misclassified as large, dangerous ones. But before the California system can be trusted to go public, it needs more monitoring stations (using both seismic and Global Positioning System (GPS) techniques), more reliable communications and testing, multiple data paths and round-the-clock daily support. The infrastructure must be made resilient to hard shaking, to ensure that the system stays online during a big quake. More seismic and GPS stations will produce faster alerts in some regions and allow tracking of large-magnitude events as they tear along active faults.

Users can decide on the thresholds for alerts and choose whether they want to hear about only the big quakes that will cause damage, or all those that are felt. But the mechanisms to deliver the alerts in the United States — through smartphone and computer apps, television and radio — remain to be developed.

BUSINESS BOON

A public–private partnership is the most effective way to disseminate the warnings, as in Japan. There, the public sector pays for the installation and long-term operation of geophysical networks to detect earthquakes and generate basic alerts. The private sector enhances and delivers the alerts, and provides support and risk-reduction expertise to the public and to businesses.

The benefits for business are threefold. First, commercial opportunities will be created. Apps will raise the alert on mobile phones, count down the time until shaking and provide location-specific instructions of what to do: for example, get under the table, exit the building or remain inside the steel-framed, glass-clad skyscraper. Self-driving cars will slow and stop. Manufacturing plants, petrochemical facilities and biotechnology companies will need other services, including determination of money-saving actions, appropriate thresholds of when actions should be taken and devices to implement them.

Second, financial losses will be reduced. After two damaging earthquakes in 2003 caused $15 million in losses at Oki Electric Industry, a chip manufacturer in Miyagi prefecture, Japan, the company spent $600,000 on an early-warning system and improvements to its buildings. In two similar earthquakes that followed, its resulting losses fell to $200,000 because machine damage and chemical spills were reduced.

ERIC RISBERG/AP/PA

An earthquake early-warning drill at a Bay Area Rapid Transit station in California.

In California, the Bay Area Rapid Transit system has implemented an automated train-braking mechanism that is triggered by earthquake early warnings. It takes 24 seconds to bring a train travelling at 112 kilometres per hour (70 miles per hour) to a stop. During peak commuting times, about 64 trains are in operation, each carrying around 1,000 passengers, and up to 45 trains travel at 112 kilometres per hour at any one time. Even one derailment at such a speed would be devastating.

Third, the recovery time for businesses is reduced. With its warning system installed, Oki's fabrication plant was closed after the earthquakes for just a few days, rather than for weeks. Minimizing damage to trains and tracks will result in faster resumption of service, which in turn supports the restart of regional businesses after a quake.

Seismology will benefit from the improved instrumentation. The 2011 Tohoku quake — the fourth largest since 1900 — yielded scientific advances because it occurred in one of the most densely instrumented regions in the world[3–9]. This extreme event tested the limits of early-warning systems. The size of the earthquake and the area affected were underestimated, and as a result, improvements to observational arrays are being made. Fast and accurate GPS sensor networks are needed to detect ground surface deformation, and ocean-floor observatories could closely monitor undersea faults.

Tracking fault motion in real time is key to making accurate shaking predictions, avoiding the underestimation that occurred in Japan's 2011 quake. To improve estimates of earthquake strength from the first signals, my research group is exploring how kilometre-scale seismic arrays can be deployed and tuned to track the progression of large fault ruptures.

In the future, the accelerometers that are embedded in smartphones and computers could provide a source of shaking data, thus boosting the number of magnitude sensors by thousands[10].

NEXT STEPS

By providing earthquake early warnings, everyone wins: people, businesses and science. So what is the hold-up? The answer is the allocation of money — and responsibility. Although US federal and state budgets are tight, the $120-million price tag to build and operate a West Coast system over 5 years works out at roughly $2.44 per person for the populations of California, Oregon and Washington. My morning coffee costs me $2.40.

What is needed to move forward is a partnership between leaders from state and federal politics, businesses, government agencies and science. Some individuals are making headway, but more must do so.

"By providing earthquake early warnings, everyone wins."

California State Senator Alex Padilla (Democrat) introduced a bill to build an earthquake early-warning system in the state, and successfully manoeuvred it through the legislature, where it was passed unanimously this September and was signed into law by Governor Jerry Brown. The governor's office of emergency services is charged with finding the necessary funding. Start-up funds are needed now, to maintain momentum. To cover the West Coast, governors and state legislators in Oregon and Washington will need to take similar steps.

Federal legislators must also take action. A number of California representatives, led by Congressman Adam Schiff (Democrat), have expressed bipartisan support for a warning system. However, with the House Interior Appropriations Bill now stalled, there is no prospect of funding this year. The USGS stands ready to deliver earthquake alerts, but it needs an extra $16 million a year to operate and maintain the system.

Partnerships between the USGS and other federal agencies are needed. The US National Science Foundation should fund the expansion of geophysical networks in zones of high seismic hazard. The Department of Homeland Security and its Federal Emergency Management Agency should support earthquake early warning as a public-safety issue.

Business leaders must also advocate the value of a warning system more strongly. John McPartland, a member of the board of directors for Bay Area Rapid Transit, has recognized and spoken widely about the need for it. Others should step forward in the many sectors that are affected by severe earthquakes.

Other regions should implement early-warning systems before their next big quake. European researchers are poised to do so, but are waiting for funding. The United Nations Educational, Scientific and Cultural Organization should continue to help in promoting the technology in other earthquake-prone areas around the world.

The scientific community provides information about the likelihood of earthquakes and their effects. But I believe that it is also important to use its expertise and authority to apply moral pressure on leaders. Although some researchers will prefer not to step out of their ivory towers, earthquake scientists who are keen to see progress must gain knowledge of and access to the policy-making process.

As Padilla commented to a colleague during the California Senate hearings for his earthquake bill: "I know you don't want to be sitting here with me after the next big one if we have not deployed this system." I am happy to give up tomorrow's coffee in exchange for a warning before the next big shake. ■

Richard Allen *is director of the Berkeley Seismological Laboratory and a professor in the Department of Earth and Planetary Science at the University of California, Berkeley.*
e-mail: rallen@berkeley.edu

1. Allen, R. *Sci. Am.* **304,** 74–79 (2011).
2. Böse, M. *et al.* in *Early Warning for Geological Disasters: Scientific Methods and Current Practice* (eds Wenzel, F. & Zschau, J.) 49–70 (Springer, 2013).
3. Fujiwara, T. *et al. Science* **334,** 1240 (2011).
4. Ide, S., Baltay, A. & Beroza, G. C. *Science* **332,** 1426–1429 (2011).
5. Jones, N. *Nature* **479,** 16 (2011).
6. Sato, M. *et al. Science* **332,** 1395 (2011).
7. Kato, A. *et al. Science* **335,** 705–708 (2012).
8. Ozawa, S. *et al. Nature* **475,** 373–376 (2011).
9. Noda, H. & Lapusta, N. *Nature* **493,** 518–521 (2013).
10. Allen, R. M. *Science* **335,** 297–298 (2012).

Joseph Meister, the first person to receive a rabies vaccine, at the Pasteur Institute in Paris.

MARCH OF TIME/TIME LIFE PICTURES/GETTY

Great myths die hard

Finding that part of the story of Louis Pasteur's rabies vaccine is false, **Héloïse Dufour** and **Sean Carroll** explore how science fables are born, spread and die.

John Snow's ending of London's 1854 cholera outbreak, Joseph Lister's development of antiseptic surgery, Alexander Fleming's invention of the drug penicillin — the history of science and medicine is full of such stories of great deeds by heroic figures.

But these are myths. They are grounded in some reality, yet careful historical research has revealed them to be far from accurate[1,2]. And, despite having been exposed by historians, the fables live on — in books, on television, in classrooms and online.

We have discovered that another story from the history of science — the heroic death of Joseph Meister, the first person to be saved by Louis Pasteur's rabies vaccine — is also a myth. Here we dissect Meister's story to understand how such myths are born, why they die so reluctantly, and what could be done to puncture them.

HOW MEISTER DIED

In July 1885, a 9-year-old French boy named Joseph Meister was badly bitten by a rabid dog, and faced near-certain death. Instead, young Meister entered medical history: he was Louis Pasteur's first human patient to be treated and saved by a rabies vaccine.

For more than half a century[3], accounts of the story in both English[4–6] and French[7] have been given a dramatic ending. In 1940, 55 years after his life was saved, Meister was serving as a gatekeeper at the Pasteur Institute in Paris. The story goes that when German forces invaded Paris in June that year, soldiers arrived at the institute demanding access to Pasteur's tomb and, rather than surrender his saviour's resting place to the Nazis, the 64-year-old Meister killed himself.

Two years ago, while researching life in occupied Paris for a book on biologist Jacques Monod[8], we came across a contemporaneous diary by Eugene Wollman in the archives of the Pasteur Institute[9]. Wollman was head of the institute's bacteriophage lab and resident on site, and his entries directly contradict the popularized accounts of Meister's suicide. The diary reveals that the date, means and motive have each been altered in the making of a myth.

In the widely repeated narrative, Meister killed himself on 14 June[4] or 16 June[5], just after the German invasion of France. But on 24 June, ten days after the Germans entered Paris, Wollman wrote: "This morning, Meister was found dead." It is often reported that Meister shot himself[5,6], but Wollman stated: "He committed suicide with gas." Some sources note that Meister committed suicide because he could not bear the idea of the Nazis profaning Pasteur's tomb[3–7]. Wollman makes no mention of any such incident. Instead, he indicates that Meister was "very depressed" and that "his wife and children had left"[9]. Like millions of others, they had fled Paris ahead of the onrushing German army.

Our interest piqued, we scoured published accounts[10,11] of Meister's death, as well as several written sources in the Pasteur Institute's archives and museum[12,13]. Moreover, Marie-José Demouron, Meister's granddaughter, kindly granted us an interview. Together, these sources confirm Wollman's version and shed further light on the motive for Meister's suicide. Meister apparently believed that his family had perished in enemy bombing, and was overwhelmed with guilt for having sent them away (ref. 11, and M.-J. Demouron, personal communication). In the chaos of France's collapse,

it was almost impossible to get news from loved ones, so Meister was unaware that they were safe. His wife and daughters actually returned later on the very day that he killed himself. As Wollman noted[9]: "Life has an extraordinarily refined cruelty."

The story of a man in despair over the apparent loss of his family, and taking his life using a gas stove hours before they come home, is tragic. But it is far from the myth of a humble servant thwarting invaders. Our sources show that the truth was not smothered from the start[10,11]. So how was the myth born?

MYTH MAKING

A myth-making pattern seems to be emerging from the researches of many science historians over recent decades[1,2]. Stories such as Meister's are founded on some facts, which are then moulded to create or fit the 'great man' model.

Fleming, for example, did isolate the antibacterial product of a mould, and did name it penicillin. But he was not responsible for the development of the antibiotic drug used in humans 14 years later, nor was he even in contact with the scientists responsible for that[1,2]. The catalysts to the making of this myth are fairly easily identified. Successful clinical trials of penicillin were first reported in 1941, in the thick of the Second World War, when infected wounds caused enormous casualties. Wartime newspaper editors naturally looked for heroic stories to inspire and encourage readers. Accounts traced the miracle drug to Fleming's serendipitous discovery many years earlier — as *The Times* of 12 June 1944 put it: "Providence had been kind to us in letting us have this most powerful agent … when against our will we were plunged into a bloody war." Similarly, Meister's story was probably distorted in part because of the war. The heroic version of his suicide embellishes the Pasteur legend and doubles as a tale of resistance.

As the myth is repeated it can become more disconnected from reality, and takes on a life of its own. After the Meister narrative began to involve German soldiers and their access to the crypt[3], the suicide date shifted closer to the Nazis' arrival in Paris. Meister was then said to have shot himself[6], and with his First World War revolver[5]. In some accounts, he is even shot by the Nazis[14].

The Fleming legend too has seen attempts at further glorification, with the claim that Fleming saved the life of former British Prime Minister Winston Churchill — twice. As a child, Fleming purportedly rescued a young Churchill from drowning, and later he was said to have cured Britain's leader with penicillin. Both claims were eventually discredited (see go.nature.com/hfakhl). But if the truth is available, why are fabricated stories still perpetuated?

Historians have long recognized that the main reason that certain myths get repeated is because they contain the ingredients of good storytelling[1,2]. Tenacious myths have heroes and villains, portray tragedy and triumph, and present climactic actions and revelations. For example, Snow did map the London cholera outbreak and correctly attributed it to a contaminated public water pump. Inconveniently, he did not remove the pump handle and end the outbreak. A committee took that action, and only after the outbreak had abated[1].

"Tenacious myths have heroes and villains, and portray tragedy and triumph."

Myths also inflate admirable qualities in their protagonists. Lister is portrayed as an eccentric outsider defending the scientific truth against a hostile environment; in fact, his use of carbolic acid was hardly revolutionary for the time[1]. And many members of the Pasteur Institute did put up resistance against the Germans, but Meister's suicide made history because it was portrayed as a manifestation of Pasteur's eminence.

A myth is also perpetuated when authors rely on a self-referencing body of previous work, rather than on primary or secondary sources. The more often one sees the same version of a story, the more likely one is to accept it as truth. This phenomenon is now amplified by the Internet. So how do we get to the real story? And why is it important to do so?

MYTH BUSTING

Scientific myths are harmful. They distort the history and the process of science[1] by portraying researchers as extraordinary people making epic advances in a fast, linear fashion. Such tales are particularly damaging to the public's and to students' understanding of the pace and complexity of science. For example, the Fleming myth ignores the vast time, effort and extra data that are required to make a medically viable drug. And by crediting luminaries with fictional achievements, we create superheroes that no student can hope to match.

Storytellers — journalists, authors, filmmakers, scientists and educators — need to be vigilant when it comes to their sources. Of course, primary and well-documented secondary sources are optimal. We recognize, however, that the search for facts surrounding events long past can be difficult and time-consuming, and it is tempting to accept something that has been widely repeated. Indeed, were it not for the discovery of Wollman's diary, one of us (S.B.C.) would have been close to propagating the Meister myth. Still, if at least we are aware of the predisposition to embellish histories, that might discourage us from parroting them without solid evidence[1,2].

Another step is to ensure that once myths have been debunked, the truth gets exposure. The Internet is both an asset and a liability in this endeavour. In the Meister case, the myth has snowballed to the point at which an admittedly fake 'contemporary journal article' is now highly visible online (see go.nature.com/wqo2z6) and is sometimes referenced as legitimate. But this power can be used to advantage. Wikipedia, for example, is a widely accessed, often initial source of information that promotes the use of mostly primary and secondary literature. Myth-busters should therefore make sure that the results of their work, and especially their sources, are properly referenced in this encyclopaedia. Other tools, such as Google Books, can be used to scour vast amounts of the published literature. Such a survey led one reader to expose the Fleming–Churchill myth.

Myths are born because they fulfil our need for a good yarn, but a powerful way to eliminate them is to replace the fiction with facts that are equally satisfying. In the case of Joseph Meister, his suicide out of despair for his family less than 24 hours before they returned is a moving story. However, because it no longer burnishes Pasteur's legend, it remains to be seen whether Meister's death will be as widely mentioned by the next wave of scientific biographers. ■

Héloïse D. Dufour *and* **Sean B. Carroll** *are at the Howard Hughes Medical Institute, University of Wisconsin–Madison, 1525 Linden Drive, Madison, Wisconsin 53706, USA.*
e-mail: sbcarrol@wisc.edu

1. Waller, J. *Fabulous Science: Fact and Fiction in the History of Scientific Discovery* (Oxford Univ. Press, 2004).
2. Allchin, D. *Sci. Educ.* **87,** 329–351 (2003).
3. Dubos, R. *Louis Pasteur: Free Lance of Science* (Little, Brown, 1950).
4. Faunce, T. A. *Pilgrims in Medicine: Conscience, Legalism and Human Rights* (Martinus Nijhoff, 2004).
5. FitzGerald, J. *What Disturbs Our Blood: A Son's Quest to Redeem the Past* (Vintage Canada, 2010).
6. Gapp, M. *Chemical Heritage* **19,** 18–19 (2001).
7. Deville, P. *Peste et Cholera* (Seuil, 2012).
8. Carroll, S. B. *Brave Genius: A Scientist, A Philosopher, and their Daring Adventures from the French Resistance to the Nobel Prize* (Crown, 2013).
9. 'Journal d'Eugene Wollman', Pasteur Institute Archives, fond Eugene Wollman, cote WLL1.A.1.
10. *Journal des débats politiques et littéraires* **152,** 2 (16–17 August 1940).
11. *Vet. Med.* **35,** 5538 (1940).
12. 'Mort de Joseph Meister', note by Marneffe, H., Pasteur Institute Archives, fond Hubert Marnette, cote MRF. ARC.13 (copy by H. Marneffe of notes taken by Noel Bernard).
13. 'A propos du suicide de Joseph Meister', note by Perrot, A., Museum of the Pasteur Institute.
14. Magill, F. N. *Masterplots II* (Salem Press, 1993).

NATL GALLERY OF IRELAND

***David and Goliath*, by Orazio Gentileschi (c. 1605–1607).**

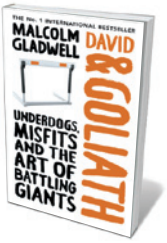
PSYCHOLOGY

Improbable heroes

Philip Ball finds much to engage and surprise in Malcolm Gladwell's study of power and how it is misinterpreted.

We think of David as the weedy foe of mighty Goliath, but he had the upper hand all along. The Israelite shepherd boy was nimble and could use his deadly weapon without getting close to his opponent. Given the skill of ancient slingers, this was more like fighting pistol against sword. David won because he changed the rules; Goliath, like everyone else on the battlefield, was anticipating hand-to-hand combat.

That biblical story about power and how it is used, misused and misinterpreted is the frame for Malcolm Gladwell's *David and Goliath*. "The powerful are not as powerful as they seem," he argues, "nor the weak as weak." Weaker sports teams can win by playing unconventionally; the children of rich families are handicapped by complacency; and smaller school-class sizes do not necessarily produce better results.



David and Goliath: Underdogs, Misfits and the Art of Battling Giants
MALCOLM GLADWELL
Little, Brown: 2013.

Gladwell describes a police chief who cuts crime by buying Thanksgiving turkeys for problem families, and the doctor who cured children with leukaemia using drug cocktails that others thought to be lethal. Conventional indicators of strength, such as wealth or military superiority, can prove to be weaknesses; what look like impediments, such as broken homes or dyslexia, can work to one's advantage. Students who are provincial high-flyers may underachieve at Harvard because they are not accustomed to being surrounded by even more brilliant peers, whereas at a mediocre university they might have excelled. Even if some of these conclusions seem obvious in retrospect, Gladwell is a consummate storyteller and you feel that you would never have articulated the point until he spelled it out.

But we all know of counter-examples. Whether someone is demoralized by or thrives on the stimulus of an academic hothouse depends on particular personal attributes and all kinds of other intangibles. More often than not, dyslexia and broken homes really are disadvantages. The achievement of a school class may depend more on what is taught, and how, and why, than on size.

The case of medic Emil J. Freireich, who developed an unconventional but ultimately successful treatment for childhood leukaemia, is particularly unsettling. If Freireich had good medical reasons for administering untested mixtures of aggressive anti-cancer drugs, they are not explained here. Instead, there is simply a description of his bullish determination to try them out come what may, seemingly engendered by his grim and impoverished upbringing. Yet determination alone can equally prove disastrous — as shown by bacteriologist Robert Koch's misguided conviction that the tuberculosis extract tuberculin would cure the disease.

Even the biblical meta-narrative is confusing. So was David not after all the plucky hero overcoming the odds, but more like Indiana Jones defeating the sword-twirling opponent by pulling out a pistol and shooting him? Was that cheating, or just thinking outside the box? In any case, there are endless examples of the stronger side winning, whether in sport, business or war, no matter how ingenious their opponents. Mostly, money does buy privilege and success. So why does David win sometimes and Goliath other times? Is it even

clear which is which (it seems that poor Goliath might have suffered from a vision impairment)?

These complications are becoming clear, for example in criminology. Gladwell is very interested in why some crime-prevention strategies work and others do not. But although his 'winning hearts and minds' case studies are surely part of the solution, recent results from behavioural economics and game theory suggest that there are no easy answers beyond the fact that some form of punishment (ideally centralized, not vigilante) is needed for social stability.

Some studies suggest that excessive punishment can be counter-productive; others show that people do not punish simply to guard their own interests, and will impose penalties on others even to their own detriment. Responses to punishment are culturally variable. In other words, punishment is a complex matter that resists simple prescriptions.

Besides, winning is itself a slippery concept. Gladwell's sympathies are for the underdog, the oppressed and the marginalized. But occasionally his stories celebrate a very narrow view of what constitutes success, such as becoming a Hollywood mogul or the president of an investment-banking firm — David turned Goliath, with little regard for what makes people genuinely inspiring, happy or worthy.

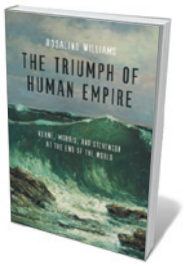
"Gladwell's sympathies are for the underdog, the oppressed and the marginalized."

None of this is a problem of Gladwell's writing, which is always intelligent and perceptive. It is a problem of form. His books, like those of legions of inferior imitators, present a 'big idea'. But it is an idea that works only selectively, and it is hard for him or anyone else to say why. These human stories are too context-dependent to deliver a take-home message, at least beyond the advice to not always expect the obvious outcome.

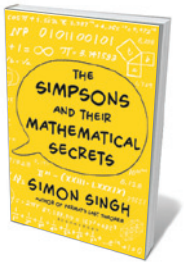
Perhaps Gladwell's approach does not lend itself to book-length exposition. In *The Tipping Point* (2000) he pulled it off, but his follow-ups *Blink* (2005), about the reliability of the gut response, and *Outliers* (2008), a previous take on what makes people succeed, similarly had theses that unravelled the more you thought about them. What remains in this case are ten examples of Gladwell's true forte: the long-form essay, as engaging, surprising and smooth as a New York latte. ■

Philip Ball *is a freelance science writer living in London. His latest book is* Serving the Reich.
e-mail: p.ball@btinternet.com

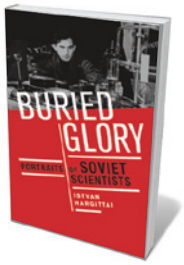
Books in brief



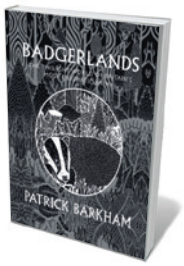
The Triumph of Human Empire: Verne, Morris, and Stevenson at the End of the World
Rosalind Williams UNIVERSITY OF CHICAGO PRESS *(2013)*
In *Nova Atlantis* (1624), the philosopher Francis Bacon characterized the human urge to dominate the globe as "the enlarging of the bounds of Human Empire, to the effecting of all things possible". By the late nineteenth century, a handful of luminaries recognized the destructive potential of that urge. Through the lives of three — Jules Verne, Robert Louis Stevenson and William Morris — science historian Rosalind Williams reveals how the transcendent power of the romantic impulse ignited environmental consciousness.



The Simpsons and Their Mathematical Secrets
Simon Singh BLOOMSBURY *(2013)*
As fans of *The Simpsons* know, the television programme's writing team is peppered with mathematicians. Physicist and writer Simon Singh skips joyously through key episodes of Matt Groening's saga, unpacking the maths embedded in each as he goes. Intoning "Be there or be a regular quadrilateral", Singh disentangles the link between pi and Homer as "Simple Simon, Your Friendly Neighborhood Pie Man"; explores Homer's "doughnut-shaped universe", admired by a cartoon Stephen Hawking in the episode 'They Saved Lisa's Brain'; and more. A chewy treat for maths geeks.



Buried Glory: Portraits of Soviet Scientists
Istvan Hargittai OXFORD UNIVERSITY PRESS *(2013)*
Nine of the fourteen Soviet scientists profiled in chemist Istvan Hargittai's tribute are buried in Moscow's Novodevichy Cemetery. But the lifting of the Iron Curtain led to a much grimmer interment, Hargittai argues: a golden era for science dimmed and died. Hargittai has delved into archives and personal recollections to bring its stars to life. A key chapter in twentieth-century research unfolds, embodied by the likes of Petr Kapitza, the low-temperature physicist who courageously supported persecuted colleagues, and the daringly original crystallographer Aleksandr Kitaigorodskii.



Badgerlands: The Twilight World of Britain's Most Enigmatic Animal
Patrick Barkham GRANTA *(2013)*
For a beast that few in Britain have seen alive, badgers have a powerful national presence — whether linked to bovine tuberculosis or place names such as Badgers Mount. Patrick Barkham revels in their ubiquity, ethology and "fright mask: the long white face burnished by two black stripes". As he visits scientists and enthusiasts, Barkham is both acute and engaging, noting, for example, the speculation that Roman emperor Marcus Aurelius inspired Kenneth Grahame's gruff Badger in his 1908 *The Wind in the Willows*.

The Secret Language of Color: Science, Nature, History, Culture, Beauty of Red, Orange, Yellow, Green, Blue & Violet
Joann Eckstut and Arielle Eckstut BLACK DOG & LEVENTHAL *(2013)*
The evanescent phenomenon of colour has gripped great minds from Plato to Isaac Newton, all the way through to researchers who now probe the links between blue light and circadian rhythms. In this many-hued tome, Joann and Arielle Eckstut zip through optics and electromagnetism. They then explore colour in art, such as the pointillist work of Georges-Pierre Seurat, and in nature, from minerals to nebulae. Fact-filled and flamboyantly illustrated. **Barbara Kiser**

MATHEMATICS

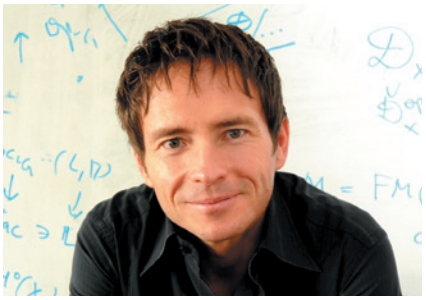
Proof of passion

Marcus du Sautoy is enthralled by a personal journey into mathematics centring on the Langlands program.

Two fascinating narratives are interwoven in *Love and Math*, one mathematical, the other personal. The love that Edward Frenkel alludes to in his title is the passion stirred in the mathematical heart by the extraordinary story of the research launched in the 1960s by eminent academic Robert Langlands, now at Princeton University, New Jersey. An unfinished saga, the 'Langlands program' is a far-reaching series of conjectures that connect number theory (the challenge of solving equations) with representation theory (part of the theory of symmetry) to create a "Grand Unified Theory of mathematics".

Frenkel grew up in Russia during the 1970s and 1980s. He initially found mathematics boring at school, and it was his fascination with quantum physics that lured him into the subject. A teacher, Evgeny Evgenievich, revealed to the youthful Frenkel that Murray Gell-Mann's prediction of quarks was actually a mathematical theory — and Frenkel's love affair with mathematics began.

Initially, it seemed his passion would be thwarted. In mid-1980s Russia, his Jewish ancestry was enough to prevent his admission to Moscow University. Frenkel's account of how his examiners at the university grilled him with increasingly difficult questions to try to provide a reason to deny him entry is as shocking as it is gripping to read.



Edward Frenkel in 2010.

PEG SKORPINSKI

But Frenkel was not put off. He enrolled at the Moscow Institute of Oil and Gas, which had an applied-mathematics programme. And he broke through the fences of the Moscow University compound to sneak into lectures and learn more about representation theory. Before long, he was doing his own original research. Publications developing the ideas of some of his mentors on braid groups and Kac–Moody algebras eventually provided his passport to the West. In 1989, when he was just 20 and still studying for his degree, he received an invitation to continue his research at Harvard University in Cambridge, Massachusetts.

Love and Math: The Heart of Hidden Reality
EDWARD FRENKEL
Basic Books: 2013.

Frenkel — now a media-feted mathematician at the University of California, Berkeley — first publicly aired his passion for the field in the 2010 short film *The Rites of Love and Math*, in which he tattoos formulae on a woman's naked body. His book brings out an almost symphonic aspect to the Langlands program. Through its refrains "solving equations" and "modular forms", Frenkel deftly takes the reader from the beginnings of this mathematical symphony to the far reaches of our current understanding.

As Frenkel has found, the Langlands program is a profound endeavour (see 'Clocking on'). But his mission, as explored in this book, is much broader. He seeks to lay bare the beauty of mathematics for everyone. As he writes, "There is nothing in this world that is so deep and exquisite and yet so readily available to all". ■

Marcus du Sautoy *is a professor of mathematics and the Simonyi Professor for the Public Understanding of Science at the University of Oxford, UK. He is the author of* The Music of the Primes.
e-mail: dusautoy@maths.ox.ac.uk

CLOCKING ON

The Langlands program

What numbers x and y make the equation $y^2+y=x^3-x^2$ true? Renaissance mathematicians approached these equations by introducing clock or modular arithmetic. On a conventional clock with 12 hours, we know that 9 o'clock plus 4 hours is 1 o'clock rather than 13 o'clock. We write this as $9+4=1$ modulo 12.

Consider a clock with 7 hours labelled 0, 1, 2, 3, 4, 5 and 6. The question now is how many pairs of numbers (x,y) chosen from the possible hours on this clock will make the equation $y^2+y=x^3-x^2$ true. For example, if we take $y=3$, then $3^2+3=9+3=12$. On the 7-hour clock, this comes out at 5 o'clock. But if we put $x=6$ in the other side of the equation then $6^3-6^2=216-36=180$, which also has remainder 5 on division by 7. So we say that the pair $(x,y)=(6,3)$ is a solution of the equation $y^2+y=x^3-x^2$ modulo 7.

Of the $7\times7=49$ possible pairs of hours on the 7-hour clock, there are 9 that make this equation true. The question that has obsessed mathematicians for generations is how this number of solutions varies as you change the number of hours on the clock. Interestingly, if you have a prime number p of hours on the clock, you will get approximately p pairs of numbers that solve this equation. For the 7-hour clock we get 2 more. With a 5-hour clock you get 1 less. For each prime p we call this error a_p. So $a_7=2$ and $a_5=-1$.

Remarkably, there is a function that allows us to predict what these errors will be as you change the prime number. This discovery, made by Martin Eichler in 1954, came from a completely different area of mathematics called modular forms:

$$q(1-q)^2(1-q^{11})^2(1-q^2)^2(1-q^{22})^2(1-q^3)^2(1-q^{33})^2(1-q^4)^2(1-q^{44})^2\ldots$$

Or, collecting the terms:

$$q-2q^2-q^3+2q^4+q^5+2q^6-2q^7-2q^9-2q^{10}+q^{11}-2q^{12}+4q^{13}+\ldots$$

The coefficient in front of q^7 is $-a_7$. The coefficient of q^5 is $-a_5$. Remarkably, the number of solutions there are of the equation $y^2+y=x^3-x^2$ on a clock with p hours is p minus the coefficient of q^p in this equation.

This is the seed from which the Langlands program grew. It was like uncovering a wormhole connecting opposite ends of the mathematical universe. Frenkel's work contributes to the possibility of another wormhole through to a more geometric corner of the mathematical landscape. The challenge of the Langlands program is to prove that these wormholes really exist.

Q&A Sandra Smith
Textile technologist

London's Victoria and Albert Museum holds more than 100,000 textile pieces. From next week, all tapestries, lace, kimonos and more not on display will be stored in the new Clothworkers' Centre for the Study and Conservation of Textiles and Fashion. Head conservator Sandra Smith talks about fabric-feasting insects, gas-emitting sequins and leaky, sticky PVC dresses.

ILLUSTRATION BY NICK HIGGINS BASED ON A PHOTO BY PETER KELLEHER

What are your main conservation challenges?
The best thing for these objects is to be kept in the dark and cold, but we want them to be visible and available. So we take a preventive role, managing the environment to slow down ageing processes. We use paper boxes that emit none of the gases or pollutants that accelerate deterioration. Relative humidity is kept below 70%, the point above which moulds kick in. We try to keep light damage, which is irreversible on textiles, to an absolute minimum. Organic dyes can be really affected by light. If you look at old tapestries, the trees can appear blue, because the dye was originally a combination of yellow and blue, and the more light-sensitive and fugitive yellow is gone.

How do you protect against insect attack?
Two species of carpet beetle and their larvae, as well as the webbing clothes moth [*Tineola bisselliella*] — the silvery-white one that people see fluttering out of their wardrobes — are of particular concern. Many chemicals used 15 years ago, such as the organophosphorus insecticide dichlorvos, are banned now, so our whole emphasis is on prevention. We avoid using partition walls to separate spaces, which can collect dust for insects to breed in. We also use insect traps in all the storage areas, and pheromone traps to lure the males.

Do embellishments such as sequins present particular problems?
In the 1920s, many sequins were made of cellulose nitrate. These might shrink, change colour, or give off acidic and oxidizing gases such as nitrous oxide. These gases can tarnish decorative metal threads and may weaken the core fabric, ultimately affecting surrounding fibres. Activated charcoal cloth, used in displays and storage containers, and other materials can absorb such 'off-gassing'. We also have tiny mass-produced beads from the late nineteenth and early twentieth centuries. These can go opaque, change colour or get sticky because they absorb moisture — a condition known as 'glass disease'. They can also give off alkali salts that accelerate fibre degradation. But there are times when we have to accept that there is no treatment. So we just photograph the object in case it goes.

VICTORIA AND ALBERT MUSEUM

A conserved 1954 Dior ensemble.

What about late-twentieth-century materials?
Polyurethane, an early foam used in shoes, can degrade into a dust. We have red vinyl Italian pumps from 1971 with polyurethane soles that are deteriorating and breaking away. Polyvinyl chloride [PVC] contains phthalate plasticizers — they are short-chain compounds which lubricate a stiff material by slipping in between the rigid polymer chains, allowing them to move more freely. But because the plasticizers are not chemically bonded to the PVC, they can migrate to the surface, forming a viscous layer.

Do you restore any pieces?
We carry out practical treatments on objects that will go on display in the galleries. This can be about strengthening them. For example, if we want to put a very fragile handkerchief on the wall, we might place another textile of the same size behind it. Then — using the tiniest threads and needles, normally used in medicine — we sew between the fibres of the handkerchief so that it is completely secured to the stronger material underneath.

Have any iconic designer outfits presented real conservation challenges?
For the museum's 2007 exhibition *The Golden Age of Couture*, we bought an extremely rare Christian Dior 1954 fuchsia-pink costume in the 'Zémire' design, made from synthetic cellulose acetate. The bodice, skirt and jacket ensemble was heavily soiled and creased, with a blackened hem and extensive water staining. Wash tests showed that cellulose acetate released its dye in acidic solutions, but the silk lining released colour in alkali solutions. So we dry-cleaned it with the solvent perchloroethylene, with limited success. We then decided to wash the costume in a pH close to neutral and with the addition of triammonium citrate, which is very effective at removing black soiling. The colour of the costume remained intense and, after hot steam removed the creasing, the fabric was once again pristine. The final task was to reinstate the complex box pleating of the heavily altered skirt. Once done, the stunning creation was close to Dior's original vision. ■

INTERVIEW BY JOSIE GLAUSIUSZ

Correspondence

Keep Australia's carbon pricing

Australia's new government may ditch the nation's pioneering carbon-pricing scheme in a move that would send a negative signal to other countries with plans for carbon trading or taxation (see E. Diringer *Nature* **501,** 307–309; 2013). Researchers and policy-makers need to be more effective in communicating such schemes to the public, to help guard against adversarial politics leading to inferior policy outcomes.

In Australia's present carbon-pricing system, emitters buy government permits for a fixed price of Aus$24 (US$22) per tonne of carbon dioxide produced. Under current law, this scheme is due to become a market-based emissions-trading scheme in 2015 and will allow trading with the cheaper emissions permits from the European Union.

The latest proposals retain Australia's target for lower emissions, but aim to replace the carbon-pricing scheme with government payments to companies that reduce emissions below a specified baseline. Critics include environmentalists, who fear that such a scheme would fail to meet Australia's emissions-reductions target, and economists, who say that it would compromise efficiency.

Although most experts regard a carbon price as the most efficient way of cutting emissions, it has been discredited in the popular discourse by opponents branding it as a punitive tax. In my view, inadequate communication with the public is partly to blame: Australia's former government failed to explain that carbon-pricing revenue is returned to low- and middle-income earners (under the new scheme, the taxpayer would pay to reduce emissions). Moreover, economists should have proclaimed their almost unanimous support for putting a price on carbon emissions.

Frank Jotzo *Australian National University, Canberra, Australia.*
frank.jotzo@anu.edu.au

Learn from China's local pilot schemes

I agree that international road maps for reducing carbon emissions will ultimately depend on stitching together inspired local initiatives (E. Diringer *Nature* **501,** 307–309; 2013). Good examples are the local pilot schemes set up nationwide in China (see, for example, *Nature* **498,** 145–146; 2013).

The National Development and Reform Commission of China selected five provinces and eight cities in 2010 for the first round of low-carbon pilots, and has now included another province and 28 more cities in the second round. The scheme encourages policy innovation by local governments for low-carbon development and includes timetables for reducing greenhouse-gas emissions.

A survey by my team has revealed that several of these pilots are using innovative measures to upgrade conventional manufacturing and agriculture, to ensure that new industries, buildings and transportation conform with these plans, and to improve energy efficiency and carbon-sink capacity through ecological programmes.

Xufeng Zhu *Tsinghua University, Beijing, China.*
zhuxufeng@tsinghua.edu.cn

Better drug access for terminal patients

The distinguished statistician Les Halpin died last month from motor neuron disease, aged 56. He was the founder of the Empower: Access to Medicine campaign to improve the availability of experimental therapies and to accelerate drug approval and licensing for people with life-threatening illnesses (see go.nature.com/v2nr3y).

After his diagnosis in May 2011, Halpin was surprised by the lack of treatments for people with his disease. This led him to ponder, from a statistical viewpoint, the regulatory systems that we apply to biomedical innovation, noting that they are much more risk-averse than the patients they are intended to serve. He believed that as a result, new medicines take longer to develop and are more costly than necessary.

His campaign has enabled drugs to get to market faster and more cheaply (see, for example, go.nature.com/gmgcyu). With support from academics, politicians and industrial scientists, he developed the Halpin Protocol (see go.nature.com/3z2wkd), which has sparked debate in the UK Parliament.

Halpin's campaign to overcome barriers to health-care translation will continue, aiming to lower them objectively and safely through increased flexibility in drug development and regulation.

David A. Brindley *University of Oxford, and CASMI, Oxford, UK; and Harvard Stem Cell Institute, Massachusetts, USA.*
david.brindley@ndorms.ox.ac.uk
Richard W. Barker *CASMI, Oxford, UK.*
Peter J. Lachmann *University of Cambridge, UK.*

Big data for a sustainable future

As discussions on Sustainable Development Goals end this week at the United Nations General Assembly in New York, I call for more 'big data' to help secure a sustainable future (see also D. Griggs *et al. Nature* **495,** 305–307; 2013). We should be collecting big data that can be used to model and test an array of different scenarios for sustainably transforming the production and consumption of energy, improving food and water security, and eradicating poverty.

Managing these issues will also help to rebalance important biogeochemical cycles (especially the carbon, nitrogen and phosphorus cycles), mitigate climate change, reverse ocean acidification and reduce the loss of biodiversity. Big data will help to illuminate the origins, nature and scale of these challenges, and how they relate to one another.

National databases and research centres can be linked to create huge databases. Initiatives similar to those of the Intergovernmental Panel on Climate Change and the Global Ocean Observing System could fill the gaps in scientific, technical and socio-economic data. New initiatives such as Global Pulse (www.unglobalpulse.org) could help in mining and mobilizing big data, which are available in real time as a result of the explosive growth in new media.

The collection and use of big data sets needs to be coordinated globally, between regions and countries, as well as between relevant agencies and institutions. The United Nations and the International Council for Science could help to forge these collaborative initiatives and networks.

Hubert Gijzen *UNESCO Regional Science Bureau for Asia and the Pacific, Jakarta, Indonesia.*
h.gijzen@unesco.org

Great scientists and society

Robert White and colleagues point out that most scientific luminaries from centuries past were religious (*Nature* **501,** 33; 2013). This says a great deal about the societies that these scientists lived and worked in, but little about the value or truth of a theistic world view, as can be seen by considering what else these great names have in common: they were all men.

Simon Vaughan *University of Leicester, UK.*
simon.vaughan@le.ac.uk

CONDENSED-MATTER PHYSICS

Quantum togetherness

Two independent experiments have demonstrated control of one mobile quantum of excitation by another. The results are likely to have ramifications for information processing and transfer. SEE LETTERS P.71 & P.76

SOUGATO BOSE

Rare are the instances when parallel experiments on completely different physical entities observe the entities' behaviour to be largely captured by the same plots. But this is precisely what has happened in the studies reported by Fukuhara *et al.*[1] (page 76) and Firstenberg *et al.*[2] (page 71) in this issue*. The researchers manipulated the smallest units of excitation of very different physical systems. Such units are called quanta. Fukuhara and colleagues focused on quanta called magnons, which carry the energy in magnets. These magnons were borne by chains of atoms, and their interactions were controlled by light. Firstenberg and co-workers dealt with photons, the quanta of light, and show that an interaction between the photons can be mediated by atoms. Despite their nearly reverse set-ups, both experiments show how such quanta clutch each other closely in 'bound states' while moving.

The remarkable similarity between the results is twofold. In both cases, the separation between the propagating quanta can be depicted by a narrow wavepacket (Fig. 1). Moreover, in both experiments the quanta are mobile: when one of the quanta moves, the other moves a step ahead or behind, while keeping close to its partner. This control of one mobile quantum by another is unprecedented for both magnons and photons, and is likely to open new doors in information processing and information transfer.

In Fukuhara and colleagues' set-up, a chain of atoms trapped by lasers and cooled to extremely low temperatures simulates a magnet. The interactions between the atoms are such that all of the atoms' spins, which are simulated by atomic energy levels, initially point the same way to lower the system's energy[3]. This configuration is that of a ferromagnet (Fig. 1a). A previous experiment had already achieved the feat of flipping the spin of a single atom in a chain, creating a localized packet of energy in the chain[4]. Such an energy packet is a mobile magnon, and was seen to propagate freely along the chain, much like a photon in a waveguide. Now the same group has flipped the spins of two neighbouring atoms, thereby producing two magnons in a chain (Fig. 1b). If these magnons did not interact, they would essentially move freely, oblivious to each other. In fact, because excitations in a spin chain can also be regarded as fermionic particles (particles that do not like to stay close to one another), the magnons would have had a tendency to run away from each other. Although this does happen to some extent, the authors also observe a strong tendency towards exactly the opposite behaviour.

The fact that the two magnons stick to each other in a pair while propagating indicates a strong interaction between them, and vindicates a classic result[5] by Hans Bethe suggesting that neighbouring magnons can form bound states. Although other evidence of bound states exists in the condensed-matter literature, owing to their unprecedented ability to image the state of each atom, Fukuhara *et al.* could for the first time track the propagation of bound magnons along a chain.

In their very different set-up, Firstenberg *et al.* shoot polarized photons in succession through a gas of rubidium atoms. The photons' energy is somewhat mismatched from the energy that would make them excite the atoms, so that the photons are not scattered or absorbed significantly as they traverse the gaseous medium. Normally, two such photons would simply pass through each other. The authors alter the photons' fundamentally 'non-interacting' nature by driving the gas with an appropriate laser. Such driving converts a lone photon to a slowly moving packet

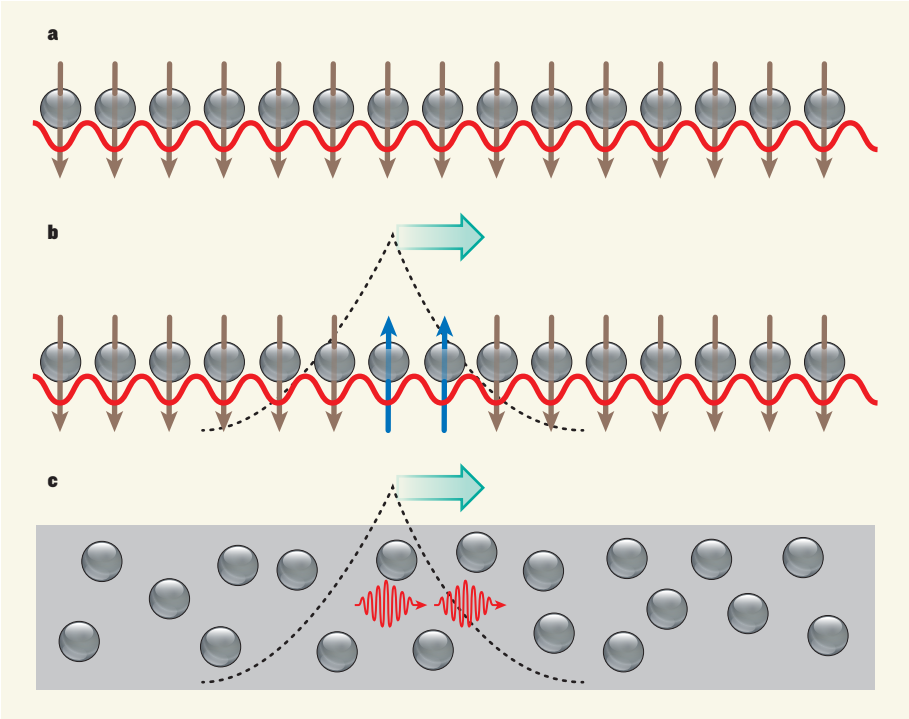


Figure 1 | Propagating pairs. **a**, Fukuhara *et al.*[1] trapped atoms (grey) in a chain using a lattice formed by light (red sinusoid) to simulate a ferromagnet, in which all atomic spins (brown arrows) point in the same direction (here, down). The barrier between the atoms, and hence their mutual interactions, is controlled by the light. **b**, The authors then flipped two spins (blue), thus creating two propagating quanta of excitation called magnons. The separation between the propagating quanta can be depicted by a narrow wavepacket (dashed line). **c**, Firstenberg *et al.*[2] produced a similar wavepacket, but one that involves a pair of mobile photons (red wiggly lines) in a gas of rubidium atoms (grey).

*This article and the papers under discussion[1,2] were published online on 25 September 2013.

of energy that is partly also an excitation distributed among some of the atoms. This excited state of the atoms is long lived, and is a member of a special class of state called a Rydberg state.

Two atoms within a certain distance of each other cannot both be in a Rydberg state because of strong van der Waals interactions. Thus, by energizing an atom to a Rydberg state, one photon in the medium can prevent another photon from doing the same, a phenomenon known as Rydberg blockade[6,7]. This creates a narrow region of gas of altered refractive index around one photon for any other photon that ventures close — effectively, a photon–photon interaction. This interaction permits a bound state for the two photons similar to Fukuhara and colleagues' two-magnon bound state. The presence of the bound state manifests itself as a strong tendency for the two photons to stay close to one another while propagating (Fig. 1c). Firstenberg *et al.* observe this tendency by recording the time interval between the clicks made by the two successive photons in the detectors after they exit the atomic medium.

Major implications of these studies may well relate to information technology. Firstenberg *et al.* also observe that their medium imparts a phase difference between photon pairs of a certain polarization and photon pairs with other polarizations. The authors use this feature to entangle the polarizations of their photons. Such entangled states of photons would be useful for connecting distinct quantum processors. Although the amount of the entanglement is rather low in the present experiment, shooting photons towards each other through the gas might increase this[8]. In general, the area of information processing that uses photons suffers greatly because photons normally do not see each other. Firstenberg and colleagues' work is a milestone in remedying that.

By the same analogy, Fukuhara and coworkers' study should open up the possibility of entangling magnons. Although there is no polarization degree of freedom here, another variable, such as the momenta of the magnons, might be entangled. However, what thrills me most about their study is that an exquisite control and observation of the long time dynamics of simulated magnets has been achieved for a case that cannot be mapped to models of independently propagating quanta. As a minimal application, the bound state of several magnons could be used to encode a bit of information. The authors' experiment provides a mechanism for moving this robust bit from place to place. More generally, it raises the hope of creating automata that exploit the dynamics of networks of continuously coupled spins to process information[9]. ■

Sougato Bose *is in the Department of Physics and Astronomy, University College London, London WC1E 6BT, UK.*
e-mail: sougato@theory.phys.ucl.ac.uk

1. Fukuhara, T. *et al. Nature* **502,** 76–79 (2013).
2. Firstenberg, O. *et al. Nature* **502,** 71–75 (2013).
3. Duan, L.-M., Demler, E. & Lukin, M. D. *Phys. Rev. Lett.* **91,** 090402 (2003).
4. Fukuhara, T. *et al. Nature Phys.* **9,** 235–241 (2013).
5. Bethe, H. *Z. Phys.* **71,** 205–226 (1931).
6. Jaksch, D. *et al. Phys. Rev. Lett.* **85,** 2208–2211 (2000).
7. Lukin, M. D. *et al. Phys. Rev. Lett.* **87,** 037901 (2001).
8. Gorshkov, A. V., Otterbach, J., Fleischhauer, M., Pohl, T. & Lukin, M. D. *Phys. Rev. Lett.* **107,** 133602 (2011).
9. Bose, S. *Contemp. Phys.* **48,** 13–30 (2007).

STEM CELLS

Close encounters with full potential

Conferring stem-cell potential on mature cells is not easy. A decisive impediment to this process has now been identified, and its elimination allows almost all mature cells to efficiently adopt a stem-cell identity. **See Article p.65**

KYLE M. LOH & BING LIM

Differentiated cells, such as skin cells or blood cells, are firmly committed to their lifelong vocation. They adopt an alternative fate only with great reluctance and through powerful coercion[1]. Such reprogramming events are typically infrequent (for example, the efficiency[1] of converting skin cells into induced pluripotent stem (iPS) cells is generally less than 1%), hampering the generation of various cell types for research and therapy. On page 65 of this issue, Rais *et al.*[2] show that, during reprogramming, cells enter an intermediary purgatory in which positive and negative influences duel to affect successful dedifferentiation*. The authors further demonstrate that eliminating a single negative influence, the protein Mbd3, allows reprogramming to iPS cells to occur with almost 100% efficiency. They thus provide salient insight into the molecular events that drive cell-fate changes.

Cellular identity is spearheaded by transcription factors, which dictate what genes should be expressed, and thus cell fate[1,3]. The introduction of just four such stem-cell regulators (Oct4, Sox2, Klf4 and Myc) into skin cells

*This article and the paper under discussion[2] were published online on 18 September 2013.

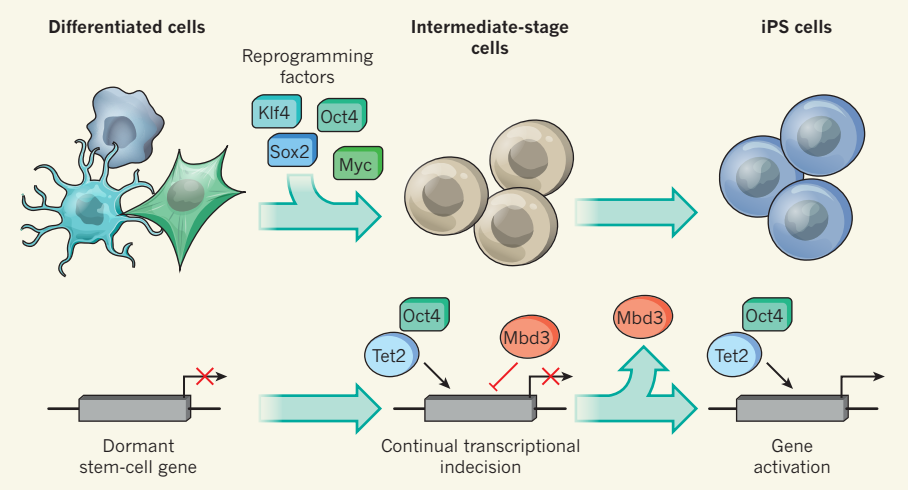


Figure 1 | The battle for pluripotency. The introduction of just four genes encoding transcription factors (*Oct4, Sox2, Klf4* and *Myc*) into differentiated cells triggers the reprogramming of these cells into induced pluripotent stem (iPS) cells, by reviving the expression of dormant stem-cell genes[1]. This process is highly inefficient, because most cells are halted at a dedifferentiated but not fully reprogrammed stage called the intermediate stage[10] (perhaps akin to pre-iPS-cells). Rais *et al.*[2] show that the reprogramming factors Oct4, Sox2 and Klf4 recruit not only transcriptional coactivators (such as Tet2)[5] to stem-cell genes, but also transcriptional repressors, including Mbd3, which potently inhibit gene reactivation. Eliminating Mbd3 enables the coactivators to efficiently resuscitate dormant stem-cell genes, allowing almost all cells to be converted into stem cells.

is theoretically sufficient to enkindle stem-cell identity[1], yet in practice this remains a sporadic process. Reprogramming inefficiency has been ascribed to either the absence of crucial activating signals during this process[4,5] or an inability of the cells to surmount inhibitory barriers[6,7]. Therefore, numerous stem-cell regulators have been tested for their ability to maximize reprogramming efficiency[1,4]. Rais *et al.*, however, took a different approach.

Genes active in stem cells are largely dormant in differentiated cells. To induce pluripotency (the ability to differentiate into all cell types of the body), reprogramming factors attempt to resuscitate the expression of these genes. For this, they recruit transcriptional coactivators[5]. But why is this effort ultimately futile in most cases[8]? Rais and colleagues find that, unexpectedly, reprogramming factors also recruit the transcriptional repressor Mbd3 (a member of the NuRD complex[9]), inadvertently directing it to suppress the very stem-cell genes they are trying to reactivate. The authors reasoned, therefore, that ablating Mbd3 — which has been implicated[6] in hindering iPS-cell reprogramming — might improve reprogramming efficiency (Fig. 1). They found that, indeed, deleting the *Mbd3* gene while introducing the reprogramming factors did the trick: when combined with optimal growth conditions[7], mouse skin, blood and brain cells could be reprogrammed with near-complete (more than 90%) efficiency, and even human cells could be reprogrammed with vastly improved efficiency[2].

These data suggest that reprogramming factors act dichotomously, trying to revive pluripotency genes while simultaneously repressing their expression (Fig. 1). This paradoxical transcriptional indecision leads to a power struggle between transcriptional coactivators and repressors (including Mbd3). Therefore, there is only stochastic reactivation of pluripotency genes, which explains why successful iPS-cell generation is such a rare event.

Potentially, after a protracted bout of many months[10], coactivators may eventually prevail. However, Mbd3 seems to be the overarching antagonist of iPS-cell reprogramming across all cell types tested. Its removal resolves these lengthy conflicts, enabling reprogramming factors to unilaterally reactivate stem-cell genes essentially in all cells.

The authors' work unites hitherto irreconcilable findings concerning cell-fate conversions. Reprogramming often yields partially reprogrammed 'pre-iPS cells'; these are an imperfect facsimile of bona fide iPS cells and are stably trapped in a paused state in which they fail to execute the terminal reprogramming events[7,11]. Remarkably, the maturation of these cells into fully fledged iPS cells might be driven by downregulation of reprogramming factors[11,12]. Why these essential reprogramming drivers should be deleterious for the last phase of reprogramming has been unclear. The present findings suggest that sustained overexpression of reprogramming factors in pre-iPS cells might continuously recruit Mbd3 to stem-cell genes, and that curtailing the levels of these factors may relieve Mbd3-mediated inhibition, allowing coactivators to triumph and driving pre-iPS cells to pluripotency.

Teleologically, it is difficult to rationalize why reprogramming factors recruit Mbd3 to hinder their own success. This may be a molecular heirloom of stem cells, in which pluripotency factors interact with Mbd3 to suppress stem-cell genes[9] in order to preconfigure differentiation[3]. Therefore, permanent Mbd3 ablation might deleteriously alter reprogrammed iPS cells by impeding their subsequent differentiation[9]. Technical approaches that transiently inactivate this protein during reprogramming may prove decisive. Whether Mbd3 removal can potentiate the direct conversion of skin cells into other differentiated cell lineages (such as brain or liver cells[1]) also remains a pertinent question.

The exceptional reprogramming efficiencies described by Rais *et al.* are evidence that cellular identity is a surprisingly malleable property that might be reforged — for example, to generate cell types of therapeutic value. Of equal importance is the unprecedented insight provided by this work into the molecular mechanisms that direct stem-cell reprogramming. Knowledge of how developmental decisions are made and revoked may also reciprocally illuminate our understanding of cancer biology, in which differentiated cells similarly relinquish their dedicated vocation and choose another path. ■

➲ NATURE.COM
For more on improved cell reprogramming, see:
go.nature.com/kimwwt

Kyle M. Loh *is in the Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA.* **Bing Lim** *is in the Stem Cell and Developmental Biology Group, Genome Institute of Singapore, 138672 Singapore, and at the Beth Israel Deaconess Medical Center, Boston, Massachusetts.*
e-mails: kyleloh@stanford.edu; limb1@gis.a-star.edu.sg

1. Graf, T. *Cell Stem Cell* **9,** 504–516 (2011).
2. Rais, Y. *et al. Nature* **502,** 65–70 (2013).
3. Loh, K. M. & Lim, B. *Cell Stem Cell* **8,** 363–369 (2011).
4. Loh, K. M. & Lim, B. *Nature* **488,** 599–600 (2012).
5. Doege, C. A. *et al. Nature* **488,** 652–655 (2012).
6. Luo, M. *et al. Stem Cells* **31,** 1278–1286 (2013).
7. Silva, J. *et al. PLoS Biol.* **6,** e253 (2008).
8. Soufi, A., Donahue, G. & Zaret, K. S. *Cell* **151,** 994–1004 (2012).
9. Reynolds, N. *et al. Cell Stem Cell* **10,** 583–594 (2012).
10. Hanna, J. *et al. Nature* **462,** 595–601 (2009).
11. Mikkelsen, T. S. *et al. Nature* **454,** 49–55 (2008).
12. Radzisheuskaya, A. *et al. Nature Cell Biol.* **15,** 579–590 (2013).

MATERIALS SCIENCE

Alloys with long memories

An alloy has been made that undergoes a remarkably reproducible phase transition over thousands of cycles. This finding could allow the development of practically useful materials that 'remember' their shape after deformation. SEE LETTER P.85

TOSHIHIRO OMORI & RYOSUKE KAINUMA

Shape-memory alloys are remarkable materials. If subjected to a deformation that would be irreversible in most metals, they revert to their original shape on heating. Such materials have applications in sensors, actuators (for example, switches that respond to heat) and medical devices such as the stents and guide-wires used in catheters. Because most applications of shape-memory alloys involve multiple deformation/re-formation cycles, the stability of an alloy's shape-memory properties over repeated use — known as its reversibility — is a pressing issue. On page 85 of this issue, Song *et al.*[1] report a great advance for the field: the development of a shape-memory alloy that has high reversibility.

When common metallic materials are subjected to an applied stress, deformation occurs. Small amounts of deformation are reversible, such that the material returns to its initial shape when the stress is removed; this deformation is measured in terms of a property called elastic strain. Any further deformation introduces defects in the material's crystal lattice, causing irreversible 'plastic' strain. However, shape-memory alloys behave differently. In these materials, a plastic strain introduced at low temperatures can be almost perfectly reversed when the material is heated above a critical temperature. Many shape-memory alloys are also superelastic — a large elastic strain of up to about 8% can be obtained above a critical temperature, giving the material a rubbery feel.

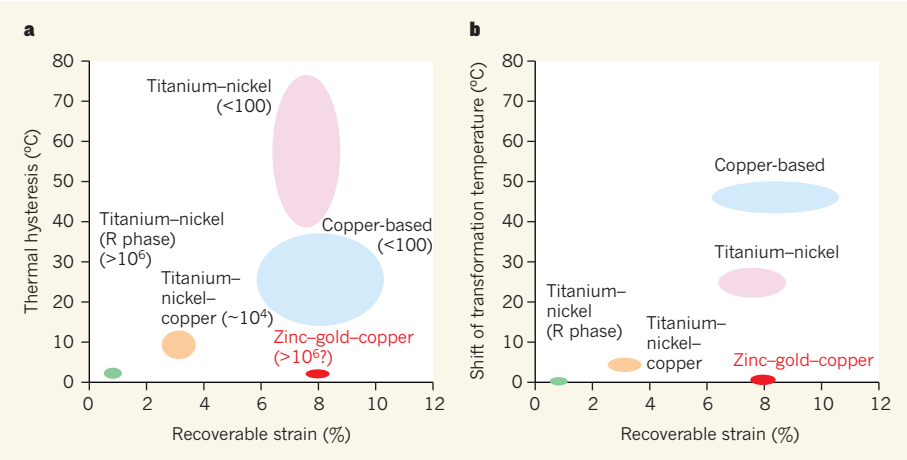


Figure 1 | Physical properties of shape-memory alloys. **a**, Shape-memory alloys usually show a thermal hysteresis in transformation temperature that is roughly proportional to the recoverable strain. The shaded regions indicate the ranges of thermal hysteresis and recoverable strain for different classes of shape-memory alloy. Song *et al.*[1] report a zinc–gold–copper alloy in which hysteresis is anomalously small with respect to the large recoverable strain. Values in parentheses[9,10] indicate the number of deformation/re-formation cycles for which the materials retain practically useful shape-memory properties; the value for the zinc–gold–copper alloy is an estimate based on the material's hysteresis. **b**, Here, the shaded regions indicate the ranges of recoverable strain and of transformation-temperature shift obtained during 10,000 heating–cooling cycles for shape-memory alloys. The transformation temperature of the zinc–gold–copper alloy hardly changes.

Both shape-memory and superelastic properties are associated with a solid-to-solid phase transition, known as a martensitic transformation, that does not involve the diffusion of atoms. Instead, it usually involves the distortion of a cubic lattice to form another lattice that has lower symmetry, such as a tetragonal structure. The parent phase and the product phase (the martensitic phase) coexist during the transition, which can be driven by temperature or by applied stress, and proceeds by migration of interfaces between the phases. Because friction opposes the movement of phase interfaces, hysteresis inevitably occurs — that is, differences arise between the temperature at which the martensitic transformation starts during cooling and the temperature at which the reverse transformation finishes upon heating. Similarly, for stress-induced transformations, differences develop between the stress that induces the transition and the stress at which the reverse transformation ends upon unloading.

The most successful and widely used shape-memory material is a titanium–nickel alloy in which roughly half the atoms are nickel. But this alloy exhibits a thermal hysteresis of about 50 °C and a stress hysteresis of 200 megapascals[2], both of which are undesirably large. These large hystereses are the result of low compatibility between the lattices of the parent and martensitic phases. The incompatibility causes elastic-strain fields close to the phase interfaces, which increase the friction that opposes interface migration. With each thermal transformation cycle (that is, a cycle of cooling and heating), the titanium–nickel alloy is damaged by the fields, and so the transformation temperature gradually shifts.

Researchers have been working to address these issues for a long time. It is now recognized that shape-memory alloys that undergo small shape changes can exhibit a smaller hysteresis and be more reversible than those that undergo larger shape changes (Fig. 1), because less damage is done to their crystal lattices during transformation. For example, the titanium–nickel alloy shows a large recoverable strain of about 8%, but its hysteresis is also large, as is the shift of its transformation temperature (25 °C after 10,000 thermal cycles for a typical alloy composition[3]). The addition of a third element such as copper to binary alloys is an effective way of reducing hysteresis and the shifts of transformation temperature, but the recoverable strain decreases as a result[4]. Titanium–nickel alloy can also undergo a martensitic transition different from the one that causes large recoverable strain, called the R-phase transformation. The material maintains its shape-memory properties for more than a million cycles of this transition, because the hysteresis is small and only a slight shift of transformation temperature occurs, but the recoverable strain is limited to only 1% (refs 5,6).

Song *et al.* have achieved a breakthrough: a material that exhibits not only a small hysteresis (and a highly repeatable transformation temperature), but also a large recoverable strain. They did this by realizing a strategy that had previously been proposed[7] for controlling hysteresis in martensitic transformations. The idea is that if a material satisfies specific crystallographic conditions — called cofactor conditions — then the lattice of the martensitic phase will be perfectly compatible with that of the parent phase.

During a normal martensitic transformation, a high density of certain lattice defects is usually generated in the martensitic phase, and a strained layer forms between the parent and martensitic phases. But if almost perfect cofactor conditions can be achieved, the generation of these defects and layers can be prevented regardless of the fraction of the material that exists as martensitic domains[8]. This vastly reduces the friction that opposes interface migration and so lowers hysteresis. The cofactor conditions can be calculated using the lattice parameters of the parent and martensitic phases, which depend on alloy composition.

In their work, Song and colleagues experimentally tuned the composition of zinc–gold–copper alloys until they found one that almost perfectly satisfies the cofactor conditions. The resulting material has a strikingly small thermal hysteresis of 2 °C and a transformation-temperature shift of less than 0.5 °C after 16,000 thermal cycles, yet has a large recoverable strain of 8% (comparable to that of the titanium–nickel alloy). Furthermore, the authors observed a highly unusual microstructure in the martensitic phase, called riverine morphology, which has never previously been reported for this type of transformation. Notably, the authors' concept for tuning the properties of martensitic materials may be applicable to many kinds of diffusionless transformation.

The authors proved the validity of their concept using simple thermal cycles. However, shape-memory properties (during cycles in which a material is subjected to applied stress and then heated) and superelasticity (during loading–unloading cycles) are more practically important, and so these properties should also be evaluated as soon as possible. The cost of the zinc–gold–copper alloy, and the difficulty of manufacturing it on a large scale, might also limit practical applications of the material. Nonetheless, by providing clear direction for the design of reliable shape-memory alloys, Song and co-workers' findings will create quite a stir in the field of materials science. ■

Toshihiro Omori *and* **Ryosuke Kainuma** *are in the Department of Materials Science, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan.*
e-mails: omori@material.tohoku.ac.jp; kainuma@material.tohoku.ac.jp

1. Song, Y., Chen, X., Dabade, V., Shield, T. W. & James, R. D. *Nature* **502,** 85–88 (2013).
2. Miyazaki, S. & Otsuka, K. *ISIJ Int.* **29,** 353–377 (1989).
3. Tadaki, T., Nakata, Y. & Shimizu, K. *Trans. Jpn Inst. Metals* **28,** 883–890 (1987).
4. Nam, T. H., Saburi, T. & Shimizu, K. *Mater. Trans.* **31,** 959–967 (1990).
5. Hwang, C. M. & Wayman, C. M. *Scripta Metall.* **17,** 381–384 (1983).
6. Miyazaki, S. & Otsuka, K. *Metall. Trans. A* **17,** 53–63 (1986).

7. James, R. D. & Zhang, Z. in *Magnetism and Structure in Functional Materials* (eds Planes, A., Mañosa, L. & Saxena, A.) 159–175 (Springer, 2005).
8. Delville, R. *et al. Phil. Mag.* **90,** 177–195 (2010).
9. Duerig, T. W., Melton, K. N., Stöckel, D. & Wayman, C. M. *Engineering Aspects of Shape Memory Alloys* (Butterworth-Heinemann, 1990).
10. Miyazaki, S., Sakuma, T. & Shibuya, T. *Application and Development of Shape Memory Alloy* (CMC, 2006).

ENZYMOLOGY

Modular biosynthesis branches out

Rings in biologically active molecules confer rigidity that helps the molecules to bind strongly and selectively to their targets. A ring-forming mechanism has been identified that involves a biochemically unusual reaction. SEE LETTER P.124

CRAIG A. TOWNSEND

Naturally occurring compounds are an important source of, and inspiration for, drugs used in human and animal medicine. Prominent among the engines of natural-product biosynthesis are modular proteins of enormous synthetic power and versatility. These proteins are programmed to select, chemically activate and assemble simple monomers in an ordered manner[1–3]. The assembly process usually involves reactions that link monomers 'head to tail', generating linear polymers. But in this issue, Bretschneider *et al.*[4] (page 124) characterize a process in which an enzyme known as a modular polyketide synthase (PKS) generates an orthogonal branch to an extending linear polymer, and uses it to forge a ring of six atoms fused to the polymer*. This behaviour imparts a second dimension to biosynthesis catalysed by these enzymes, and potentially provides synthetic biologists with a tool to modify and rigidify PKS products in a predictable manner.

The ability of carbonyl compounds (those that contain C=O groups) to react with each other is a cornerstone of carbon–carbon bond-forming reactions. PKSs, and the mechanistically allied fatty-acid synthase enzymes, exploit the intrinsic reactivity of small carbonyl-containing monomers to create the elaborate frameworks of polyketide natural products such as the antibiotic erythromycin, and rhizoxin, a compound with antitumour activity. When the portion of the rhizoxin intermediate shown in Figure 1 is bound to a β-ketoacyl synthase (KS) domain of a PKS, conventional chain extension would proceed by reaction of the intermediate with a malonyl monomer bound to an acyl-carrier-protein (ACP) domain of the enzyme (Fig. 1a, b). Organic chemists know this as a decarboxylative thio-Claisen reaction.

*This article and the paper under discussion[4] were published online on 18 September 2013.

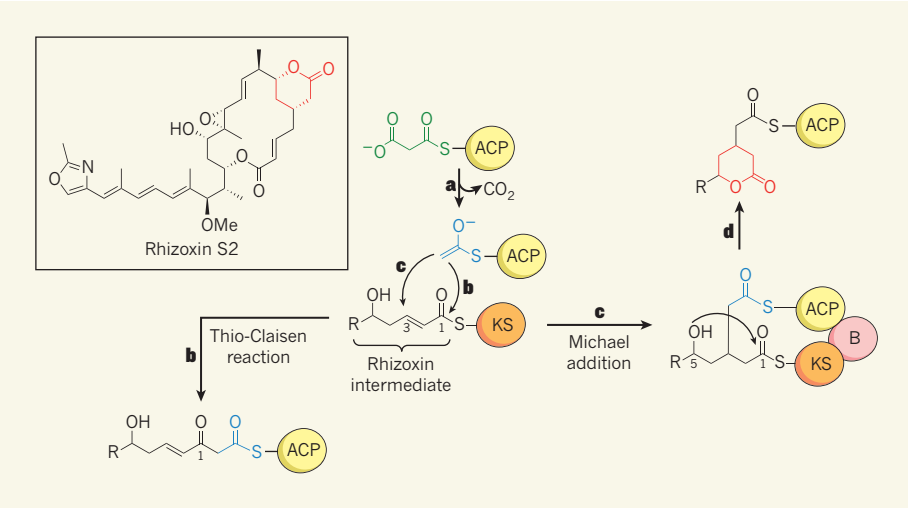


Figure 1 | Chain extension and branching mechanisms for modular polyketide synthases. Rhizoxin compounds, such as rhizoxin S2, are made by modular polyketide synthase (PKS) enzymes. The lactone ring of rhizoxins is shown in red. Me, methyl group. **a**, Most PKS reactions construct linear polymers. The process begins when an activated malonyl group (green) attached to an acyl-carrier-protein (ACP) domain of a PKS loses carbon dioxide to form an activated acetyl group (blue). **b**, In this example, a portion of a rhizoxin intermediate attached to the β-ketoacyl synthase (KS) is attacked by the activated acetyl group at the first carbon of the intermediate. A thio-Claisen reaction occurs, extending the molecular chain by two carbons. R represents the remainder of the molecular chain. **c**, Bretschneider *et al.*[4] report that, in the rhizoxin-forming PKS, KS and ACP domains flank a branching (B) domain. This domain directs the attack of the activated acetyl-ACP group to the intermediate's third carbon. The resulting Michael addition reaction creates a branch orthogonal to the main molecular chain. **d**, A hydroxyl (OH) group at carbon 5 of the intermediate subsequently attacks carbon 1, forming rhizoxin's lactone.

But in the modular PKS that makes rhizoxin, an unusual branching (B) domain is interposed between a KS and an ACP domain, forming an unprecedented biosynthetic mini-module. Bretschneider and colleagues' analysis of the B domain's amino-acid sequence revealed no similarity to the sequences of any known PKS catalytic domains. However, the authors' crystal structure of the KS–B di-domain revealed a protein-folding pattern much like that of dehydratase (DH) domains characteristic of fatty-acid synthases and PKSs[5].

In fact, the closest structural relative of the KS–B domain is a variant of the DH domain known as a product template (PT) domain, which is present in another family of PKSs[6,7]. PT domains, like DH domains, contain a histidine amino-acid residue that is crucial for catalytic activity. This residue has been replaced by a tyrosine residue in the B domain. Bretschneider and co-workers performed studies in which they replaced residues of the B and KS domains with other amino acids. Their results suggest that the B domain has no catalytic function, and that the unusual chain branching promoted by the rhizoxin-forming PKS takes place exclusively in the KS domain.

What is more, the authors found that, during branching, a reactive malonyl–ACP is delivered to the extending rhizoxin polyketide intermediate in an unusual way. Whether because of crowding by the B domain, or because of slight alterations to the architecture of the KS active site, the incoming malonyl unit adds at the third carbon of the intermediate, rather than at the first carbon (Fig. 1c). In other words, the malonyl unit undergoes a decarboxylative Michael addition reaction, rather than a thio-Claisen reaction.

In a series of excellent biochemical experiments, Bretschneider *et al.* established that a hydroxyl group (OH) on the fifth carbon atom of the resulting product reacts with a nearby group (a thioester connected to the KS domain) to generate a six-membered lactone ring linked to the downstream ACP domain (Fig. 1d). The intermediate formed in this process is then poised for further chain-extension reactions, and for a final macrocyclization reaction (a transformation in

which a macrocycle, a ring of nine atoms or more, is generated), yielding the structural core of rhizoxin. But although the chemical steps are clear, the role of the B domain and the processes that control the Michael reaction are still not fully understood. A related Michael addition process has been reported recently[8] for another specialized family of KS domains. In that reaction, two fatty-acid or polyketide chains combine using the same basic carbonyl chemistry as that observed by Bretschneider and colleagues, but the six-membered ring that forms is not a lactone.

Rhizoxin's lactone ring is essential for the molecule's ability to interfere with cell division, and therefore for its potential as an anticancer agent. More broadly, clinically and commercially important natural products and their derivatives also tend to contain ring structures[9]. Indeed, a catalytic mechanism has evolved whereby an internal group — typically containing an oxygen or nitrogen — in the final linear intermediates produced by PKSs and by other modular enzymes initiates a macrocyclization reaction, as in the biosynthesis of erythromycin and rhizoxin. So why is this?

The answer is that the conformational freedom of a macrocycle is much more restricted than that of its linear analogue. Such constraints have a major role in enabling the favourable, specific binding of molecules such as proteins and nucleic acids, and are a crucial consideration in fields such as drug design[10–12]. Smaller changes in structure, such as the formation of six-membered rings, can in turn propagate large changes in the properties of macrocycles[13], as Bretschneider and co-workers' paper shows.

The authors' findings expand the known uses of carbonyl chemistry in the biosynthetic assembly of complex natural products. The branching mechanism observed in rhizoxin is enabled by a small cassette of three enzymatic domains, which can now therefore be sought and identified in other PKS systems. This means that synthetic biologists can explore the use of this cassette in 'designer' biosynthetic systems, potentially allowing them to make biologically active molecules that take advantage of the additional, localized conformational rigidity provided by fused lactones. ■

Craig A. Townsend *is in the Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, USA.*
e-mail: ctownsend@jhu.edu

1. Staunton, J. & Weissman, K. J. *Nat. Prod. Rep.* **18,** 380–416 (2001).
2. Fischbach, M. A. & Walsh, C. T. *Chem. Rev.* **106,** 3468–3496 (2006).
3. Marahiel, M. A. & Essen, L.-O. *Meth. Enzymol.* **458,** 337–351 (2009).
4. Bretschneider, T. *et al. Nature* **502,** 124–128 (2013).
5. Labonte, J. W. & Townsend, C. A. *Chem. Rev.* **113,** 2182–2204 (2103).
6. Crawford, J. M. *et al. Nature* **461,** 1139–1143 (2009).
7. Crawford, J. M. & Townsend, C. A. *Nature Rev. Microbiol.* **8,** 879–889 (2010).
8. Fuchs, S. W. *et al. Angew. Chem. Int. Edn* **52,** 4108–4112 (2013).
9. Driggers, E. M., Hale, S. P., Lee, J. & Terrett, N. K. *Nature Rev. Drug Discov.* **7,** 608–624 (2008).
10. Leavitt, S. & Freire, E. *Curr. Opin. Struct. Biol.* **11,** 560–566 (2001).
11. Chang, C. A., Chen, W. & Gilson, M. K. *Proc. Natl Acad. Sci. USA* **104,** 1534–1539 (2007).
12. Diehl, C. *et al. J. Am. Chem. Soc.* **132,** 14577–14589 (2010).
13. Woodward, R. B. *et al. J. Am. Chem. Soc.* **103,** 3213–3215 (1981).

BIOLOGICAL TECHNIQUES

Chromosomes captured one by one

Data obtained from analysing chromosomal organization and interactions in individual cells unify previous results obtained by single-cell imaging and studies of population-averaged genomic interactions. **See Article p.59**

JOB DEKKER & LEONID MIRNY

Each human cell carries DNA molecules that, when combined, are more than two metres long. A fascinating problem in cell biology is, therefore, how these long molecules are organized inside the cell's nucleus. This question is not just of interest from a structural perspective. Understanding chromosome folding will also provide deeper insight into genomic processes, such as the regulation of gene expression and the maintenance of genome stability. On page 59 of this issue, Nagano *et al.*[1] describe a genomic approach to probing the three-dimensional arrangement of chromosomes within single cells*. The results reveal large cell-to-cell variation in chromosome structure and nuclear organization, but also show that common principles of organization derived previously from cell-population studies apply to individual cells.

*This article and the paper under discussion[1] were published online on 25 September 2013.

Chromosomes and individual genomic regions (loci) have been studied in single cells using microscopy[2]. Such work has shown that although chromosomes are not randomly organized, their structure varies between cells in a population. However, approaches at the molecular level, based on the chromosome conformation capture (3C) technique, have produced maps of averaged genome-wide interactions in chromatin (DNA–protein complexes)[3], allowing chromosome organization across large cell populations to be determined. These studies have led to the discovery that there are general principles of chromosome folding[4,5], but they give no information about cell-to-cell variability, nor how these folding principles are implemented at the single-cell level. Nagano *et al.* present a variant of the 3C technique (called single-cell Hi-C) that allows genome-wide measurement of chromatin interactions in individual cells.

In 3C, cells are chemically fixed to covalently link any pair of genomic loci that are in close spatial proximity. The chromatin is then digested with a restriction enzyme, and DNA ends are joined to form unique DNA-ligation products, each of which represents a contact between two loci in a cell in the population. The resulting ligation-product library can then be interrogated using deep-sequencing methods, to determine the genomic location of interacting loci (the Hi-C approach)[6]. Because 3C methods are typically performed on millions of cells, the measured chromatin-interaction frequencies reflect the probability with which loci are interacting across the population. However, it is difficult to use the data generated to discover which interactions occur together in single cells, how many cells display a particular pair-wise interaction, or to what extent nuclear organization varies in different cells.

Nagano and colleagues' single-cell Hi-C method adds a clever twist to the 3C procedure. It involves most of the steps of the classical 3C technique, but is performed inside the nuclei of permeabilized cells. Subsequent isolation of single nuclei allows all DNA-ligation products to be collected from each of the cells. Deep sequencing of these products yields a list of pair-wise chromatin interactions from that cell. The authors further applied statistical analyses to extract significant patterns of chromatin interaction that reflect various features of the genome's spatial organization.

Analysing individual immune cells called T helper cells, Nagano *et al.* note a striking level of cell-to-cell variability in interactions between and within chromosomes. Whereas in one cell a chromosome (more strictly, two homologous chromosomes) might interact with almost all others, in another cell the same

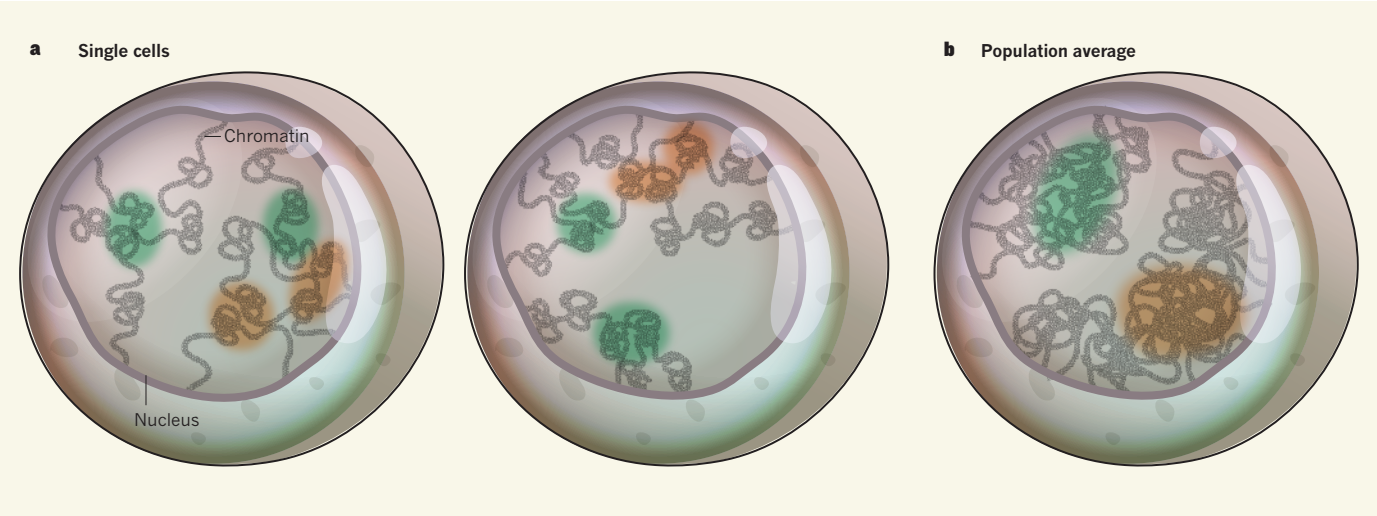


Figure 1| One versus many. **a**, Nagano *et al.*[1] report that, in individual cells, different domains of active (green) and inactive (orange) chromatin (DNA–protein complexes) interact with each other, with a greater tendency for interactions between domains of the same functional status. **b**, This tendency can be seen more clearly in interaction data averaged over a cell population. The advantage of single-cell analyses, however, is that they support the notion of domains as building blocks of chromosomal organization.

chromosome might interact with only a few others. In part, this is consistent with earlier microscopy observations[2] showing that each chromosome occupies a distinct 'chromosomal territory' and interacts with only a few neighbours, which differ from cell to cell.

How can the ability of a chromosome to interact with a dozen other chromosomes be reconciled with the formation of chromosomal territories? It could be that territory shape is highly irregular, with sub-chromosomal domains protruding from the main chromosomal mass. High-resolution microscopy should allow this hypothesis to be tested. It also remains to be seen whether the propensity of chromosomes to engage in many or only a few trans-chromosomal interactions reflects a state of individual chromosomes, or that of an individual cell — for example, its progression through the cell cycle or its response to signals.

Nagano *et al.* also report cell-to-cell variation in the internal organization of chromosomes. However, when disparate interaction patterns from individual cells are pooled, a common pattern emerges — the interactions 'clump' together. Such clumps correspond to interactions within previously identified topologically associating domains (TADs)[7–9], which were first discovered using population-averaged chromatin-interaction maps. TADs have been proposed to form the building blocks for modular assembly of larger-scale structures[10]. This modular architecture is also evident in trans-chromosomal interactions, which are mostly formed not by isolated, protruding genomic regions, but by entire domains. Nagano and co-workers' Hi-C data are too sparse to determine whether every TAD is present in every cell. But their statistical analysis suggests that cell-to-cell variability in chromosome organization is partly due to differential assembly of relatively reproducible, modular sub-chromosomal domains.

Although interactions between the structural domains differ among cells, they are not random, instead reflecting functional states of individual domains. Consistent with population-averaged data[6,11,12], in individual cells, domains enriched for active genomic regions tend to interact with other active domains; similarly, inactive domains tend to interact with other inactive domains (Fig. 1).

The information on interactions within a single chromosome can be used to build a three-dimensional model of that chromosome and so reproduce the observed interactions. Nagano *et al.* build three-dimensional models of the single-copy X chromosome (to eliminate ambiguity caused by homologous chromosomes) for several individual cells. The models show that domains marked by active chromatin are frequently located near the periphery of a chromosome, where they may interact with active domains on other chromosomes, in agreement with population-averaged data[11].

Population-based data have also suggested that the probability of interactions between a pair of loci on the same chromosome decreases approximately as the inverse of the genomic distance between the loci. Despite cell-to-cell variation in specific interactions, scaling of the interaction probability with genomic distance is surprisingly consistent between individual cells and agrees with the population-based scaling.

Questions remain. For example, how variable is chromatin organization within the modular chromosome domains? Are interactions between genes and their regulatory elements also stochastic across the cell population? What is the contribution of real-time chromatin dynamics to cell-to-cell variation, and is this different for local structures and higher-order conformations?

Answering some of these questions will require further technological developments, for instance to allow the capture of more than the currently possible 2.5% of all interactions in a cell. Answering others may require a combination of assays within the same cell — such as single-cell Hi-C with analysis of all RNA transcripts. This should help to uncover how cell-specific chromosome conformation relates to stochastic gene expression. Single-cell Hi-C promises to become an important approach in determining chromosomal organization within cells and its relevance to gene expression. ■

Job Dekker *is in the Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605-0103, USA.*
Leonid Mirny *is at the Institute for Medical Engineering and Science and in the Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*
e-mails: job.dekker@umassmed.edu; leonid@mit.edu

1. Nagano, T. *et al. Nature* **502,** 59–64 (2013).
2. Cremer, T. & Cremer, M. *Cold Spring Harb. Perspect. Biol.* **2,** a003889 (2010).
3. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295,** 1306–1311 (2002).
4. Bickmore, W. A. & van Steensel, B. *Cell* **152,** 1270–1284 (2013).
5. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. *Nature Rev. Genet.* **14,** 390–403 (2013).
6. Lieberman-Aiden, E. *et al. Science* **326,** 289–293 (2009).
7. Dixon, J. R. *et al. Nature* **485,** 376–380 (2012).
8. Nora, E. P. *et al. Nature* **485,** 381–385 (2012).
9. Sexton, T. *et al. Cell* **148,** 458–472 (2012).
10. Gibcus, J. H. & Dekker, J. *Mol. Cell* **49,** 773–782 (2013).
11. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. *Nature Biotechnol.* **30,** 90–98 (2012).
12. Yaffe, E. & Tanay, A. *Nature Genet.* **43,** 1059–1065 (2011).

doi:10.1038/nature12482

Supervolcanoes within an ancient volcanic province in Arabia Terra, Mars

Joseph R. Michalski[1,2] & Jacob E. Bleacher[3]

Several irregularly shaped craters located within Arabia Terra, Mars, represent a new type of highland volcanic construct and together constitute a previously unrecognized Martian igneous province. Similar to terrestrial supervolcanoes, these low-relief paterae possess a range of geomorphic features related to structural collapse, effusive volcanism and explosive eruptions. Extruded lavas contributed to the formation of enigmatic highland ridged plains in Arabia Terra. Outgassed sulphur and erupted fine-grained pyroclastics from these calderas probably fed the formation of altered, layered sedimentary rocks and fretted terrain found throughout the equatorial region. The discovery of a new type of volcanic construct in the Arabia volcanic province fundamentally changes the picture of ancient volcanism and climate evolution on Mars. Other eroded topographic basins in the ancient Martian highlands that have been dismissed as degraded impact craters should be reconsidered as possible volcanic constructs formed in an early phase of widespread, disseminated magmatism on Mars.

The source of fine-grained, layered deposits[1,2] detected throughout the equatorial region of Mars[3] remains unresolved, though the deposits are linked to global sedimentary processes, climate change, and habitability of the surface[4]. A volcanic origin has been suggested on the basis of the stratigraphy, morphology and erosional characteristics of the deposits[5]. The case for a volcanic source is further strengthened by the spectroscopic detection of sulphates in many of these deposits[6] and detailed analyses of such rocks at the Meridiani Planum landing site, which revealed materials altered under water-limited, acidic conditions that were probably governed by volcanic outgassing[7]. Yet, although very fine-grained ash can be dispersed globally from a large explosive eruption on Mars[5,8], the currently known volcanic centres are unlikely to have been the sources for thick, low-latitude layered deposits in Arabia Terra[9].

The lack of identifiable volcanic sources that could have produced possible volcanogenic sediments in Meridiani Planum or in Gale crater is not a unique problem. In fact, 70% of the crust was resurfaced by basaltic volcanism, with a significant fraction emplaced from as yet unrecognized sources[10]. Thus, undetected volcanic source regions must exist within the ancient crust of Mars. Therefore, the following questions arise: first, is ancient volcanism poorly understood because higher Noachian erosion rates[11] obliterated evidence for source regions? Second, are ancient volcanoes highland volcanoes of fundamentally different character from the well-recognized, massive, Hesperian shield volcanoes[12,13]? We suggest that the answer to the second question is yes; we propose a new category of ancient volcanic construct that has escaped detection until now.

Volcanism is the thread binding nearly every aspect of Mars's geological evolution. The crust of the planet was built through magmatism and effusive volcanism[14], although an early phase of explosive volcanism might have emplaced a significant amount of fragmented material across the ancient crust[15]. Volatiles outgassed[16] from volcanoes controlled atmospheric chemistry[17] and strongly affected climate[18–20] throughout Martian history. The geochemistry and habitability of Martian soils and sedimentary rocks are ultimately controlled by the global sulphur cycle, which is fundamentally linked to volcanism[21]. It is therefore critical to understand all styles and phases of Martian volcanism and how they have affected the Martian climate through time.

Evidence for volcanism in Arabia Terra

We present evidence for a new category of ancient volcanic construct on Mars, ancient supervolcanoes, which together could have produced vast amounts of lava and pyroclastic materials throughout Arabia Terra and beyond. The features, which we call 'plains-style caldera complexes', are characterized by the presence of collapse features, low topographic relief (lower than that of typical paterae), and association with plains-style lavas and friable, layered deposits. Taken together, the features, each with explosive outputs probably in excess of terrestrial supervolcanoes, constitute a previously unrecognized ancient volcanic province in Arabia Terra (Fig. 1).

The best example of a plains-style caldera complex is Eden patera, which is a large, irregularly shaped topographic depression (dimensions ~55 km northwest–southeast and 85 km southwest–northeast) located at 348.9° E, 33.6° N within Noachian–Hesperian ridged plains of probable volcanic origin. The complex, which reaches a maximum depth ~1.8 km below surrounding plains, includes at least three linked depressions (Fig. 2) bounded by arcuate scarps and associated with numerous faults and fractures. Although this feature has never been differentiated from impact craters in the region, it lacks any geological indicator of an impact origin, such as the presence of ejecta, an uplifted rim, nearly circular geometry or the presence of a central peak[22]. Its high ratio of depth to diameter is inconsistent with that of an ancient impact crater that has been modified by erosion[23]. We therefore rule out an impact origin for Eden patera.

We interpret Eden patera as a caldera complex on the basis of its similarity to terrestrial calderas[24] and its association with features that indicate formation by means of collapse and volcanism both within and outside the depression. The surrounding terrain comprises ridged plains typical of Hesperian basaltic volcanism on Mars[10]. Within the complex are fault-bounded blocks that display surfaces similar to the adjacent ridged plain lavas (Fig. 2a). These blocks are tilted towards the crater centre and are unrelated to headwall scarps that would

[1]Planetary Science Institute, Tucson, Arizona 85719, USA. [2]Department of Earth Sciences, Natural History Museum, London SW7 5BD, UK. [3]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

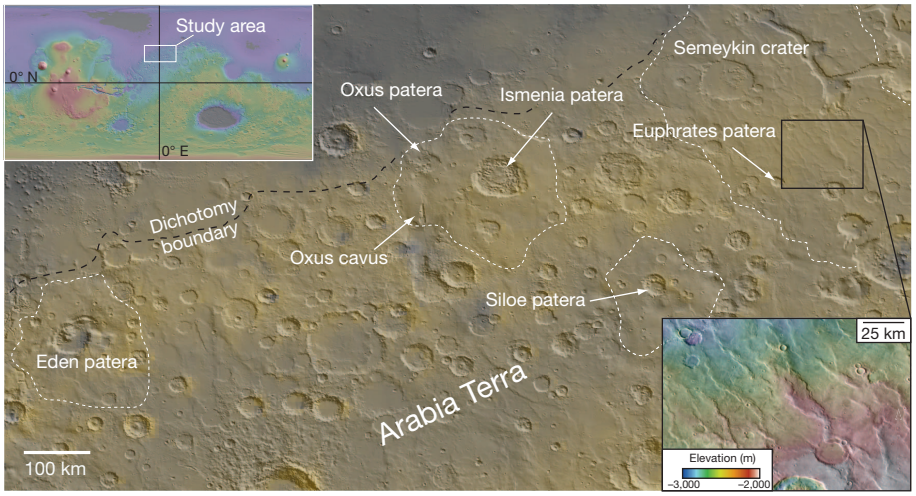


Figure 1 | Geographic context of the northern Arabia Terra region. The dusty nature of Arabia Terra is shown in false-colour TES-derived albedo data draped over MOLA hillshade data; bright colours correspond to dusty surfaces. Recently named geographic features discussed in the text are labelled.

suggest a process similar to landslides. Graben associated with the interior fault blocks may have originally been linked with circumferential graben outside the complex related to older collapses or progressive formation through 'piecemeal', multicyclic evolution[24]. We interpret a mound ~700 m high (11 km north–south and 23 km east–west) within the complex to be a graben-related vent (Fig. 2b). Two sets of nearly continuous terraces are found ~100 and 150 m above the caldera floor. These terraces are strikingly similar to the 'black ledge' described during the Kīlauea Iki eruption in 1959 (ref. 25), indicating high stands of a drained lava lake[26]. A small mound (1 km across) several hundred metres high and located between the two terraces shows surface cracks similar to a tumulus[27]. Although tumuli clefts form during inflation, we suggest that these cracks formed as the lava lake drained and the sinking lake crust was draped onto caldera wall rocks.

The presence of volcanic features and significant faulting consistent with collapse leads us to conclude that these linked depressions represent a large caldera complex formed in the Late Noachian to Early Hesperian. A lacustrine origin for the terraces is unlikely due to the

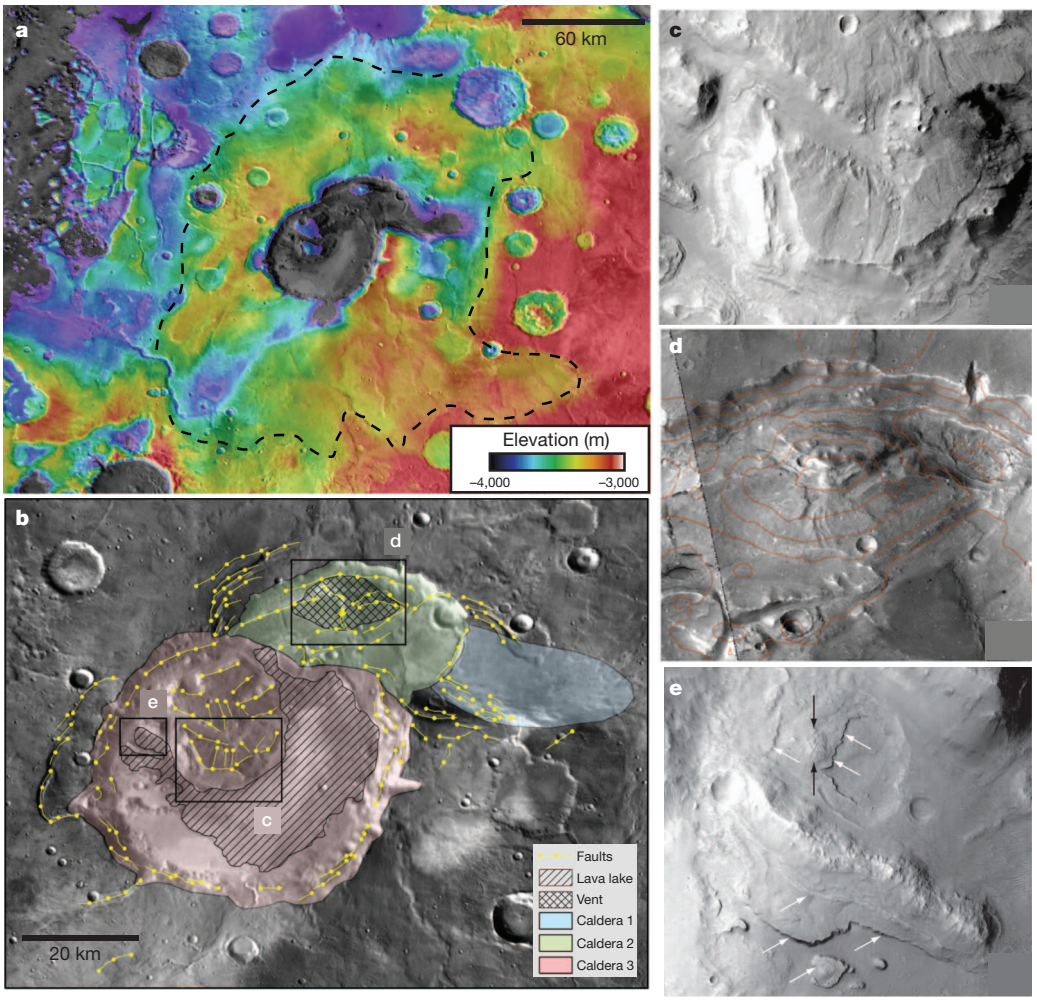


Figure 2 | The geology of Eden patera. a, MOLA topographic data are draped over THEMIS daytime infrared data, showing the morphology of Eden patera. **b**, Geological mapping reveals the presence of at least three calderas, indicated by coloured shading. **c–e**, Enlargements of the rectangles in **b**. The caldera contains evidence for fault blocks that preserve ridged plain lavas on the upper surface (**c**), a probable vent (**d**), and a series of terraces that mark lava high stands of a once active lava lake (white arrows) and cracked crust (black arrows) due to the draping of fragile crust onto pre-existing surfaces during lava lake drainage (**e**).

paucity of channels found in or around the depression that could be linked to aqueous surface processes. In addition, there is no apparent evidence for lacustrine sediments within the basin, and the depression is deeper than expected for a feature of this size that was partly filled by outside sediment. The sequential development of this feature (calderas 1–3 in Fig. 2) seems to have undergone a transition from surface sagging (caldera 1 in Fig. 2) to significant disruption of the crust and subsequent down-dropping of large surface blocks (calderas 2 and 3 in Fig. 2).

Several other features throughout the region have similar characteristics. Euphrates patera is an irregularly shaped depression that reaches 700 m in depth below the surrounding lava plains and contains several benches in the interior that might be explained by sequential episodes of collapse or lava-lake high stands (Fig. 3). The irregular, rhombohedral form of the depression might relate to shortening in the southwest–northeast direction. Fractured surface textures in the centre of the depression are morphologically similar to lava surfaces disrupted by collapse caused by the withdrawal of lava.

Other features in northern Arabia Terra contain evidence for collapse associated with volcanic activity. Siloe patera (6.6° E, 35.2° N) is a set of nested deep depressions that reach ~1,750 m below the surrounding plains (Fig. 3). Similar to Eden patera, the nested craters are characterized by steep-walled depressions linked by arcuate scarps and faults. The primary structure is linked to a subtle northeast–southwest-trending depression to the south that reaches ~700 m depth, which we interpret as evidence for sagging due to the migration of a magma body at depth. Although there is no evidence for impact ejecta around the structure, there is a single set of lobate flows emanating from the southwest portion of the depression rim, which may represent a single set of lava or pyroclastic flows reaching ~60 km from the rim. Irregular mounds of friable materials inside the nested craters are interpreted as pyroclastics from the volcano or as younger friable deposits of another origin.

Some other depressions in the region contain less well preserved evidence for volcanism, but in all cases they contain suites of features that are difficult to explain by other geological processes. Semeykin crater is a large scalloped depression surrounded by lava plains and friable deposits; it also contains mounds of friable materials in its interior and ridged plains along the exterior. A suite of features, Ismenia patera, Oxus patera and Oxus cavus, are located together near 0° E, 38.5° N. The two paterae have scalloped, breached rims composed of layered materials. Oxus cavus is an elongated depression within a broad mound 200–300 m high adjacent to a deep depression indicative of sagging or collapse. Although none of these structures individually contains as many pieces of evidence to clearly point to volcanism as are seen at Eden patera, all of the features contain some evidence for structural collapse, which is most likely to have occurred through magmatic activity (although other hypotheses are considered below).

Eden patera and Euphrates patera represent the strongest evidence for large calderas in Arabia Terra, on the basis of their association with features that are diagnostic of surface disruption and collapse coupled with evidence for effusive and explosive volcanism. Some of the other features with fewer diagnostic features might not all represent caldera formation, or they could have experienced a range of processes responsible for the current morphology. Nonetheless, the region does show strong evidence that several large depressions did not form as impact craters and are most easily explained as volcanic calderas.

The roles of ice and impact

Some depressions throughout Arabia Terra have previously been interpreted as thermokarst features[28,29]. There is no doubt that geological surfaces in and around the Arabia Terra region have been modified by ice[30], but we argue that it is unlikely that ice removal could have created the collapse features themselves. Scalloped depressions in the Utopia Planitia region of Mars bear a striking resemblance in size, shape and morphology to thermokarst features found on Earth[31,32]; both terrestrial and Martian types form depressions on the order of metres to tens of metres in depth[33,34] (Fig. 4). Thus, those well-accepted thermokarst features are orders of magnitude smaller than the collapse features discussed here, whereas the proposed volcanic

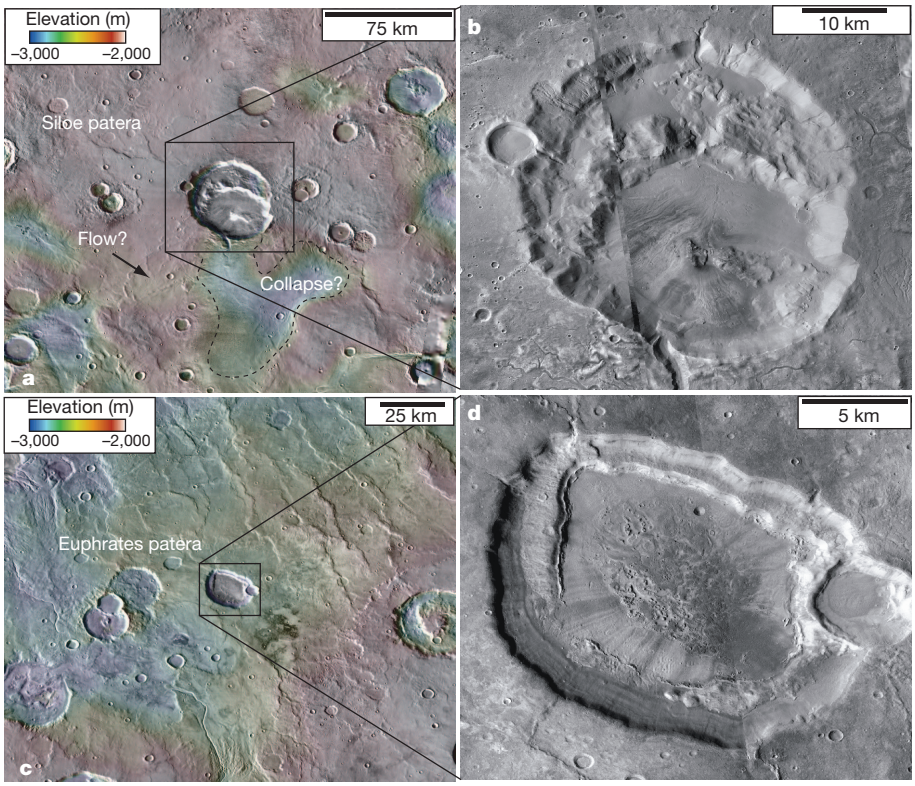


Figure 3 | The geology of Siloe and Euphrates paterae. MOLA data draped over CTX images show the morphologies of Siloe patera (**a**; rectangle enlarged in **b**) and Euphrates patera (**c**; rectangle enlarged in **d**).

structures in Arabia Terra are of the same scale and morphology as terrestrial supervolcanoes[35] (Fig. 4). If these proposed volcanic structures are in fact the result of thermokarst, then they are a new type of thermokarst beyond any that has been definitively recognized previously.

In addition, the large volume of the collapse features is a strong constraint on the possible origins. If they formed by collapse associated with the removal of subsurface ice, it necessarily implies that a significant volume of ice was removed from each location, quickly enough to cause the high strain rates required for faulting. However, none of the features is associated with outflow channels, which are typically cited as evidence for the rapid removal of surface or near-surface ice. Furthermore, the amount of ice that could have existed below such depressions can be constrained from quantitative models of Martian subsurface porosity[36]. For example, if Eden patera had been formed by the removal of subsurface ice, it would have been necessary for all of the available void space to be entirely saturated with ice to a depth of ~10 km (see Supplementary Information). We therefore conclude that, although ice and thermokarst processes could have been involved in the modification of the collapse features, it is unlikely to explain the origin of the collapse or the presence of the large depressions.

It is also possible that the depressions in Arabia Terra represent degraded impact craters. However, none of the features described above contain evidence for impact geology, such as the presence of ejecta, raised crater rims, central peaks or inverted stratigraphy. It is possible that erosion has removed such evidence, but the proposed calderas are found adjacent to ancient impact craters of similar size (and possibly similar age) that have preserved clear evidence for impact origins (Fig. 5). Furthermore, impact craters that have been degraded by erosion[37] have much lower depth-to-diameter ratios than those measured in the proposed calderas (Fig. 5). The relations between depth and diameter among the calderas are only consistent with depth-to-diameter ratios of impact craters that are only partly modified; such craters have preserved at least some critical aspects of impact geology.

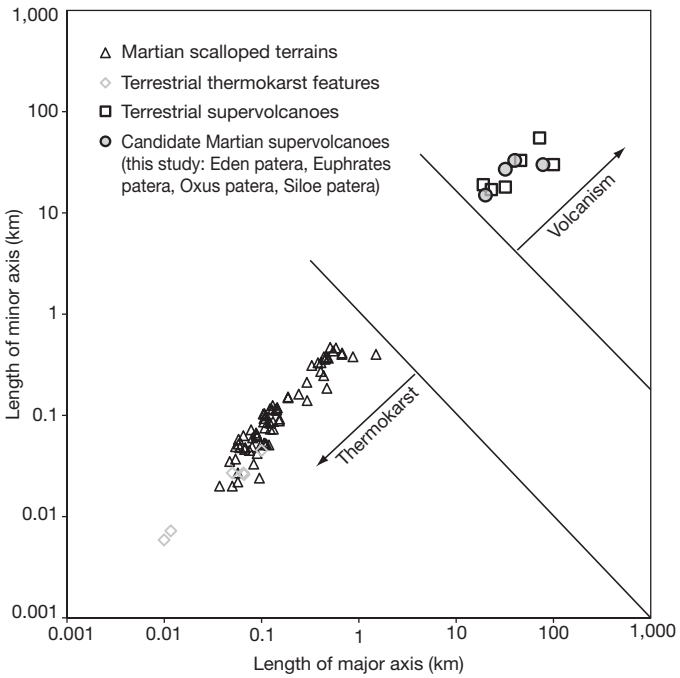


Figure 4 | Comparison of thermokarst features, terrestrial supervolcanoes and the putative supervolcanoes on Mars. A plot of the dimensions of typical terrestrial and Martian thermokarst features shows that they have roughly similar sizes[32,34]. The proposed calderas in Arabia Terra have similar dimensions to those of terrestrial supervolcanoes, which together are orders of magnitude larger than known thermokarst features.

A new category of Martian volcanic construct

Taken together, these features constitute a new category of Martian volcano that can be described as plains-style caldera complexes, of which Eden patera is the type example. Eden patera is not associated with a major edifice. Each of the Martian low-slope paterae recognized previously[12,13] shows a major edifice related to repeated volcanic deposition of explosive and effusive products. Thus, Eden patera seems to be a new class of Martian volcanic feature, formed through a combination of magma withdraw, subsurface magma migration (caldera 1) and/or major explosive episodes that would have distributed ash regionally or globally such that they did not accumulate near the vent (calderas 2 and 3). These geomorphic features are most analogous to those of a terrestrial supervolcano.

On Earth a supervolcano is defined as a volcano that can produce at least 1,000 km^3 of volcanic materials in an eruption. On Mars it is generally not possible to link a single volcanic deposit to a particular eruption event. However, erupted volumes can be constrained from the volume of void space observed in the caldera itself, if that collapse

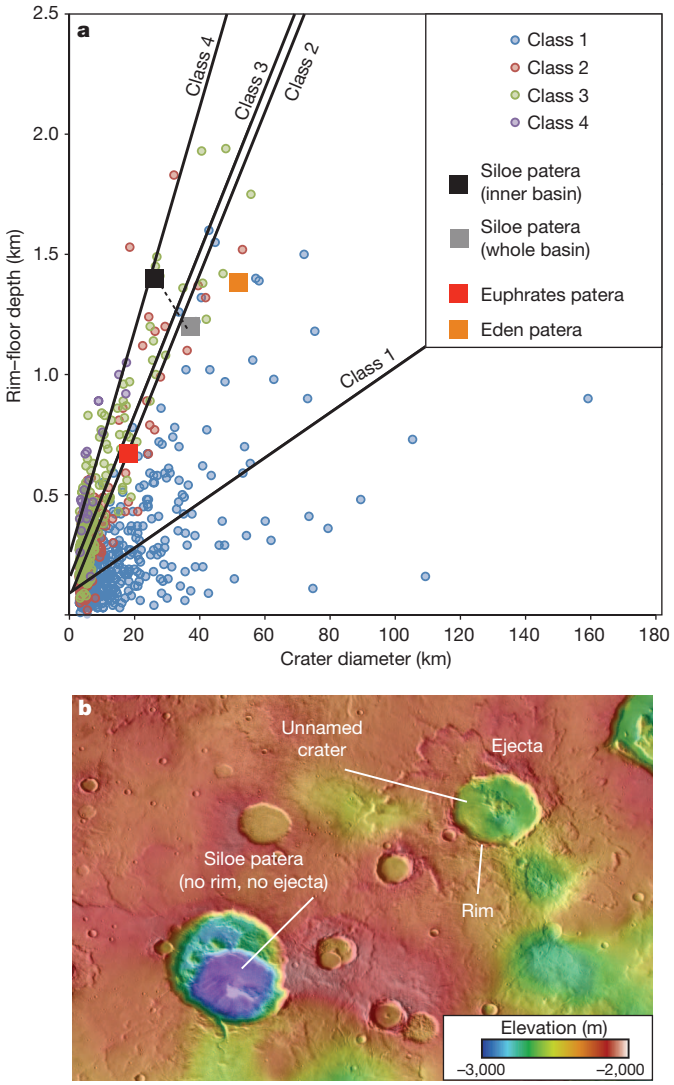


Figure 5 | Comparison of the depth-to-diameter ratios of possible Martian supervolcanoes with those of known impact craters. **a**, Plot of crater measurements for all of the craters within the area of Fig. 1 with diameters of 1 km or more that have previously been categorized according to their level of preservation[37]. Class 1 craters are the most degraded and class 4 are the least degraded (essentially pristine). The proposed supervolcanoes plot along trendlines associated with moderately modified craters that preserved impact morphologies. **b**, However, the calderas clearly do not contain morphological evidence for impact processes as seen in adjacent craters of similar size.

is assumed to have occurred as a result of the removal of magma during eruptive events. Focusing on a subset of these features including Eden patera, Oxus cavus, Semeyken crater and Ismenia patera, the average depression volume is more than 3,300 km^3. This volume at each site could have been produced by the removal of a comparable amount of dense-rock-equivalent material. Assuming an average density of 2,800 kg m^{-3} of the magma and an estimated density of 2,000 kg m^{-3} for erupted lava or 1,300 kg m^{-3} for tephra, it is possible to estimate the amount of erupted material from each source. The average minimum erupted volume could be more than 4,600–7,200 km^3 for each of these caldera complexes. Although this estimate cannot be linked to a single eruption event, nor can we differentiate void space created by explosive ejection from that created by magmatic subsidence, these features are unlike other known Martian volcanoes and it is likely that they fall in the category of terrestrial supervolcano, on the basis of both geomorphology and eruptive potential.

The question remains: Why would large calderas associated with explosive volcanism occur in northern Arabia Terra? One possibility is that volatile-rich crust was subducted beneath Arabia Terra during an ancient episode of plate tectonics[38]. However, although the presence of northwest–southeast-trending scarps related to thrust faulting in northern Arabia Terra related to southwest–northeast compression might seem consistent with such an interpretation, the estimated displacement on such faults is small and does not support the model of an active plate margin[28,39]. It is more likely that the dichotomy boundary evolved as a result of crustal thinning associated with endogenic processes[39]. The crust within Arabia Terra is relatively thin and more similar to thicknesses modelled for the northern lowlands than for the southern highlands[40]. Even so, we consider an origin due to subduction to be an open question that merits further consideration.

We suggest that a combination of regional extension and thermal erosion of the lower crust in the Late Noachian to Early Hesperian led to a rapid ascent of magma in the northern Arabia Terra region. It is not necessary that the magmas were of higher viscosity (more silicic) or had higher volatile content than other Martian magmas. The lower gravity and atmospheric pressure on Mars lead to bubble nucleation at greater depths and greater gas expansion in comparison with Earth[41]. As a result, pyroclastic eruptions would be more commonly associated with basaltic volcanoes on Mars than on Earth, particularly if the magma rapidly ascended and erupted and was not stored in degassing magma chambers for long periods, as is thought to occur at younger, large shield volcanoes[42]. In fact, it is possible that explosive volcanism was more prevalent globally on early Mars because the ancient crust was thinner, leading to less devolatilization of magmas during ascent. The result may have been the deposition of vast quantities of tephra early in Mars's history.

Links to global geology

Explosive eruptions of fine-grained materials might be linked to the formation of fretted terrains that also occur in northern Arabia Terra, the origin of which represents one of the major outstanding mysteries in Mars science[29] (Fig. 6). Youthful ice-related processes may have modified the fretted terrains, but the sediment of which they are composed was probably deposited in the Noachian to Early Hesperian[43]. These voluminous, fine-grained sediments may be tephra deposits from explosive volcanic activity in northern Arabia Terra. In fact, layered terrains throughout Arabia Terra might consist of tephra deposits, but previous work has suggested that the source region was the Tharsis province[5]. A much simpler explanation is that plains-style calderas produced these sediments locally[44] (Fig. 6).

Our understanding of volcanism[45] on Mars continues to evolve as numerous, small (tens of kilometres in diameter) and dispersed volcanic centres are recognized throughout the Tharsis region[46–49], and degraded, ancient volcanic centres are recognized in the southern highlands[12,13]. Major volcanic constructs are now recognized in several distinct provinces throughout the Martian surface, although with a paucity of features previously identified between Tharsis and Syrtis Major (Fig. 4). The features identified here constitute a major volcanic province in Arabia Terra, which fills a void in a large fraction of the surface where volcanoes are expected to have occurred but have never been recognized.

The origin of altered, fine-grained, layered, clay-bearing and sulphate-bearing sediments throughout the equatorial region of Mars has yet to be explained. A local volcanic source could explain the presence of clastic materials composing these deposits and could serve as a significant source of volcanogenic sulphur that might have led to acidic alteration at the surface and strongly perturbed the Martian climate, sending it into periods of significant warming[18] or substantial cooling[19]. We suggest that fine-grained deposits at the Meridiani Planum and Gale crater landing sites, as well as friable deposits in the equatorial region of Mars, might ultimately have originated from volcanic sources in Arabia Terra. Further mapping of plains-style caldera complexes might reveal additional ancient volcanic source regions distributed throughout the Martian highlands. Deciphering the nature of an early

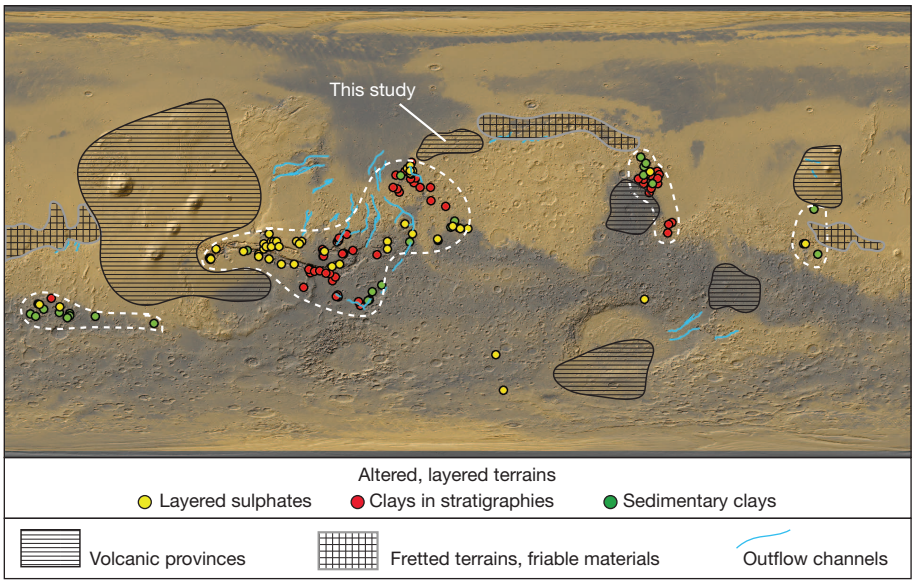


Figure 6 | Links to global geology. The distribution of major volcanic provinces on Mars in relation to friable and fretted terrain, layered sulphates[17] and layered clay-bearing terrains[50].

phase of widespread, disseminated, explosive volcanism will be critical to revealing the climate history and past habitability of Mars.

METHODS SUMMARY

The primary data sets used to evaluate the geomorphology of topographic depressions in the Arabia Terra region were gridded elevation data from the Mars Orbiter Laser Altimeter (MOLA) and a global mosaic of daytime infrared images from the Thermal Emission Imaging System (THEMIS). Additional data products included high-resolution images and digital topographic data from the High Resolution Stereo Camera (HRSC) aboard the Mars Express spacecraft, high-resolution images from the High Resolution Imaging Science Experiment (HiRISE) and Mars Context Imager (CTX) aboard the Mars Reconnaissance Orbiter, and the Mars Orbiter Camera (MOC) aboard the Mars Global Surveyor spacecraft. These data products are available within the publicly available Java Mission-planning and Analysis for Remote Sensing (JMARS) software produced by Arizona State University (available at http://jmars.mars.asu.edu). Image-based geological mapping was performed after geo-registering these data products within a geographic information system (GIS). Data from the Thermal Emission Spectrometer (TES) were used to evaluate dust cover and albedo.

Full Methods and any associated references are available in the online version of the paper.

Received 8 May; accepted 15 July 2013.

1. Malin, M. C. & Edgett, K. S. Sedimentary rocks of early Mars. *Science* **290,** 1927–1937 (2000).
2. Edgett, K. S. & Malin, M. C. Martian sedimentary rock stratigraphy: outcrops and interbedded craters of northwest Sinus Meridiani and southwest Arabia Terra. *Geophys. Res. Lett.* **29**(24), 2179, http://dx.doi.org/10.1029/2002gl016515 (2002).
3. Hynek, B. M. Implications for hydrologic processes on Mars from extensive bedrock outcrops throughout Terra Meridiani. *Nature* **431,** 156–159 (2004).
4. Bibring, J. P. *et al.* Global mineralogical and aqueous mars history derived from OMEGA/Mars express data. *Science* **312,** 400–404 (2006).
5. Hynek, B. M., Phillips, R. J. & Arvidson, R. E. Explosive volcanism in the Tharsis region: global evidence in the Martian geologic record. *J. Geophys. Res. Planets* **108,** 5111, http://dx.doi.org/10.1029/2003je002062 (2003).
6. Gendrin, A. *et al.* Sulfates in martian layered terrains: the OMEGA/Mars Express view. *Science* **307,** 1587–1591 (2005).
7. McCollom, T. M. & Hynek, B. M. A volcanic environment for bedrock diagenesis at Meridiani Planum on Mars. *Nature* **438,** 1129–1131 (2005).
8. Wilson, L. & Head, J. W. Explosive volcanic eruptions on Mars: tephra and accretionary lapilli formation, dispersal and recognition in the geologic record. *J. Volcanol. Geotherm. Res.* **163,** 83–97 (2007).
9. Kerber, L., Head, J., Madeleine, J. B., Forget, F. & Wilson, L. The dispersal of pyroclasts from ancient explosive volcanoes on Mars: implications for the friable layered deposits. *Icarus* **219,** 358–381 (2012).
10. Greeley, R. & Spudis, P. Volcanism on Mars. *Rev. Geophys.* **19**(1), 13–41 (1981).
11. Golombek, M. P. *et al.* Erosion rates at the Mars Exploration Rover landing sites and long-term climate change on Mars. *J. Geophys. Res. Planets* **111,** E12S10, http://dx.doi.org/10.1029/2006je002754 (2006).
12. Williams, D. A. *et al.* The Circum-Hellas Volcanic Province, Mars: overview. *Planet. Space Sci.* **57,** 895–916 (2009).
13. Xiao, L. *et al.* Ancient volcanism and its implications for thermal evolution of Mars. *Earth Planet. Sci. Lett.* **323–324,** 9–18 (2012).
14. Zuber, M. T. The crust and mantle of Mars. *Nature* **412,** 220–227 (2001).
15. Bandfield, J. L., Edwards, C. S., Montgomery, D. R. & Brand, B. D. The dual nature of the martian crust: young lavas and old clastic materials. *Icarus* **222,** 188–199 (2013).
16. Lammer, H. *et al.* Outgassing history and escape of the Martian atmosphere and water inventory. *Space Sci. Rev.* **174,** 113–154 (2012).
17. Gaillard, F., Michalski, J., Berger, G., McLennan, S. M. & Scaillet, B. Geochemical reservoirs and timing of sulfur cycling on Mars. *Space Sci. Rev.* **174,** 251–300 (2013).
18. Halevy, I., Zuber, M. T. & Schrag, D. P. A sulfur dioxide climate feedback on early Mars. *Science* **318,** 1903–1907 (2007).
19. Tian, F. *et al.* Photochemical and climate consequences of sulfur outgassing on early Mars. *Earth Planet. Sci. Lett.* **295,** 412–418 (2010).
20. Johnson, S. S., Mischna, M. A., Grove, T. L. & Zuber, M. T. Sulfur-induced greenhouse warming on early Mars. *J. Geophys. Res.* **113,** E08005, http://dx.doi.org/10.1029/2007JE002962 (2008).
21. King, P. L. & McLennan, S. M. Sulfur on Mars. *Elements* **6,** 107–112 (2010).
22. French, B. M. *Traces of Catastrophe: A Handbook of Shock-metamorphic Effects in Terrestrial Meteorite Impact Structures* (LPI Contribution no. 954, Lunar and Planetary Institute, 1998).
23. Malin, M. & Dzurisin, D. Landform degradation on Mercury, the Moon, and Mars: evidence from crater depth/diameter relationships. *J. Geophys. Res.* **82,** 376–388 (1977).
24. Acocella, V. Caldera types: how end-members relate to evolutionary stages of collapse. *Geophys. Res. Lett.* **33,** L18314, http://dx.doi.org/10.1029/2006gl027434 (2006).
25. Richter, D. H., Eaton, J. P., Murata, K. J., Ault, W. U. & Krivoy, K. L. *Chronological Narrative of the 1959–1960 Eruption of Kilauea Volcano, Hawaii* (US Geological Survey Professional Paper 539-E, 1970).
26. Stovall, W. K., Houghton, B. F., Harris, A. J. L. & Swanson, D. A. Features of lava lake filling and draining and their implications for eruption dynamics. *Bull. Volcanol.* **71,** 767–780 (2009).
27. Walker, G. P. L. Structure and origin by injection of lava under surface crust of tumuli, 'lava rises', 'lava-rise pits', and 'lava-inflation clefts' in Hawaii. *Bull. Volcanol.* **53,** 546–558 (1991).
28. McGill, G. E. Crustal history of north central Arabia Terra, Mars. *J. Geophys. Res. Planets* **105,** 6945–6959 (2000).
29. Sharp, R. P. Mars: Fretted and chaotic terrains. *J. Geophys. Res.* **78,** 4073–4083 (1973).
30. Head, J. W., Mustard, J. F., Kreslavsky, M. A., Milliken, R. E. & Marchant, D. R. Recent ice ages on Mars. *Nature* **426,** 797–802 (2003).
31. Niu, F. J., Lin, Z. J., Liu, H. & Lu, J. H. Characteristics of thermokarst lakes and their influence on permafrost in Qinghai-Tibet Plateau. *Geomorphology* **132,** 222–233 (2011).
32. Bouchier, A. *Response to Permafrost Failures on Hillslopes in the Brooks Range, Alaska*. MS thesis, Colorado School of Mines (2008).
33. Soare, R. J., Osinki, G. R. & Roehm, C. L. Thermokarst lakes and ponds on Mars in the very recent (late Amazonian) past. *Earth Planet. Sci. Lett.* **272,** 382–393 (2008).
34. Sejourne, A. *et al.* Scalloped depressions and small-sized polygons in western Utopia Planitia, Mars: A new formation hypothesis. *Planet. Space Sci.* **59,** 412–422 (2011).
35. Miller, C. F. & Wark, D. A. Supervolcanoes and their explosive supereruptions. *Elements* **4,** 11–15 (2008).
36. Clifford, S. M. *et al.* Depth of the Martian cryosphere: Revised estimates and implications for the existence and detection of subpermafrost groundwater. *J. Geophys. Res. Planets* **115,** E07001, http://dx.doi.org/10.1029/2009je003462 (2010).
37. Robbins, S. J. & Hynek, B. M. A new global database of Mars impact craters ≥1 km: 1. Database creation, properties, and parameters. *J. Geophys. Res.* **117,** E05004 http://dx.doi.org/10.1029/2011je003966 (2012).
38. Sleep, N. Martian plate tectonics. *J. Geophys. Res.* **99,** 5639–5655 (1994).
39. Watters, T. R. Thrust faults along the dichotomy boundary in the eastern hemisphere of Mars. *J. Geophys. Res.* **108** (E6) 5054 (2003).
40. Neumann, G. A. *et al.* Crustal structure of Mars from gravity and topography. *J. Geophys. Res.* **109,** E08002, http://dx.doi.org/10.1029/2004JE002262 (2004).
41. Wilson, L. & Head, J. W. Mars: review and analysis of volcanic eruption theory and relationships to observed landforms. *Rev. Geophys.* **32,** 221–263 (1994).
42. Wilson, L., Scott, E. D. & Head, J. W. Evidence for episodicity in the magma supply to the large Tharsis volcanoes. *J. Geophys. Res.* **106**(E1), 1423–1433 (2001).
43. Irwin, R. P., Watters, T. R., Howard, A. D. & Zimbelman, J. R. Sedimentary resurfacing and fretted terrain development along the crustal dichotomy boundary, Aeolis Mensae, Mars. *J. Geophys. Res. Planets* **109,** E09011, http://dx.doi.org/10.1029/2004je002248 (2004).
44. Kerber, L., Michalski, J., Bleacher, J. & Forget, F. in *44th Lunar and Planetary Science Cconference, Lunar and Planetary Institute, Houston, Texas, USA*, abstract 2290 (2013).
45. Hodges, C. A. & Moore, H. J. *Atlas of Volcanic Landforms on Mars* (US Geological Survey Professional Paper 1534, 1994).
46. Bleacher, J. E., Greeley, R., Williams, D. A., Cave, S. R. & Neukum, G. Trends in effusive style at the Tharsis Montes, Mars, and implications for the development of the Tharsis province. *J. Geophys. Res. Planets* **112,** E09005, http://dx.doi.org/10.1029/2006je002873 (2007).
47. Hauber, E., Bleacher, J., Gwinner, K., Williams, D. & Greeley, R. The topography and morphology of low shields and associated landforms of plains volcanism in the Tharsis region of Mars. *J. Volcanol. Geotherm. Res.* **185,** 69–95 (2009).
48. Bleacher, J. E. *et al.* Spatial and alignment analyses for a field of small volcanic vents south of Pavonis Mons and implications for the Tharsis province, Mars. *J. Volcanol. Geotherm. Res.* **185,** 96–102 (2009).
49. Richardson, J. A., Bleacher, J. E. & Glaze, L. S. The volcanic history of Syria Planum, Mars. *J. Volcanol. Geotherm. Res.* **252,** 1–13 (2013).
50. Ehlmann, B. L. *et al.* Subsurface water and clay mineral formation during the early history of Mars. *Nature* **479,** 53–60 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. Frey, B. Hynek, S. Wright, J. Zimbelman and L. Tornabene for discussions that improved the quality of the manuscript. Funding was provided by the NASA Mars Data Analysis programme.

Author Contributions J.R.M. performed the initial observations, processed image and topographic data and wrote most of the manuscript. J.E.B. wrote portions of the manuscript, performed geological mapping and processed imaging and topographic data. Both authors synthesized the results, developed the ideas and edited the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R.M. (michalski@psi.edu).

METHODS

Identification of volcanic features. Most well-recognized volcanic edifices on Mars occur as central vent structures within topographically elevated terrain built through sustained volcanism around the vent. Pavonis Mons (Supplementary Fig. 1) is an example of typical shield-style volcanism on Mars. Note that Pavonis Mons contains evidence for collapse and crustal sagging owing to removal or migration of magma. The central caldera is a steep-sided, nearly circular crater that formed during the latest stage of volcanic activity. However, that caldera is nested within a larger set of ring-fractures that suggest more extensive collapse or additional collapse events. Complex calderas are common on the Earth and Mars, and occur as a result of collapse associated with magma withdrawal, owing to migration of magma at depth, removal of magma during eruptions, or both. Tyrrhenus Mons (Supplementary Fig. 1b) is an ancient volcano of different character on Mars. It also is defined by a topographic rise with ring fractures. However, the flanks of Tyrhennus Mons have a much lower profile than Pavonis Mons and are composed of fingering, eroded, layered materials thought to indicate the presence of pyroclastic materials. Tyrrhena patera (the main caldera) might be the final location of the central vent, but the caldera is breached and eroded, and there is evidence for secondary calderas on the volcano. The plains-style caldera complexes that we have identified in northern Arabia Terra bear characteristics similar to each of these volcanoes, yet some other characteristics that are fundamentally different. Most notably, the calderas in Arabia Terra do not occur on topographically elevated volcanic constructs, which is probably one reason why they have never previously been identified as volcanoes despite abundant evidence for volcanic processes.

The International Astronomical Union (IAU) formally named six features (five paterae and one cavus) located in northern Arabia Terra in 2012. These features have not been discussed by their proper names in previous literature. As discussed in the main text, these features, as well as Semeykin crater (which was previously named) have morpohological characteristics that are inconsistent with impact origins. They are not the only depressions in Arabia Terra with enigmatic origins, but they are the subset of features on which this paper is focused.

Supplementary Fig. 2 shows the morphology of all seven features discussed in the text. Of these, Eden patera and Euphrates patera bear the strongest evidence for ancient volcanism. The others probably formed through collapse, although the link to volcanism is less clear in the other cases.

Calculation of volumes. One of the goals of this work is to constrain the amount of collapse that occurred at each of the putative caldera complexes. Estimates of collapse volume are important for placing minimum constraints on the amount of magma involved in ancient igneous processes at each site, and for testing alternative hypotheses for the origins of these features (for example pseudokarst, described below).

To estimate the amount of collapse that has occurred, we mapped the features in a GIS environment and used MOLA elevation data to calculate the volume of each depression. The volume calculations are straightforward but depend on several assumptions. First we describe the technical process, and then the assumptions.

For each site, gridded MOLA data were contoured and draped onto MOLA hillshade and THEMIS daytime infrared data. The contoured data helped to delineate the maximum topographic level of the depression at each feature. We then converted gridded MOLA elevation data to triangulated irregular networks (TINs) at each site. The TINs provide a combined quantitative measure of surface elevation and area. Then, for each site, we fitted a plane to the maximum allowable elevation corresponding to the closest approximation to a closed depression. We then calculated the volume of void space beneath the plane, within the caldera at each site.

Examples of the volume calculations are shown in Supplementary Fig. 3. Note that the fit of an elevation plane to each site is imperfect. One assumption we make is that topography has not changed since the formation of the depressions. This is clearly not so, but it is a limitation on our approach. There is clear evidence that the entire region has been tilted towards the north since the formation of these features. In addition, several of the calderas described in this work show evidence that they were breached, which means that there is not an obvious closed depression at most structures. Therefore, delineation of a single closed depression grossly underestimates the actual volume of the structure because the calderas are typically breached at some elevation along the rim. Last, younger impact craters have been superimposed on each site, which further complicates the effort to define a single elevation contour related exclusively to the caldera collapse itself. Given these challenges, we have made every effort to perform the volume calculations with the most conservative approach possible, to avoid overestimating the volume of each depression. We have therefore chosen elevations that in each case are below the rim of the depression, to provide the best estimate of a closed depression with the knowledge that this decision results in an underestimate of the total caldera volume.

There are errors associated with these analyses, both in the direction of artificially increasing the volume estimates and in the direction of artificially decreasing them. One of the major errors resulting in underestimation of the volume calculations is related to the fill deposits within the depressions themselves. Those materials were probably sourced from the caldera in each case, but their topographic setting now is still considered part of the underlying terrain. In other words, there is no way to identify the true caldera floor because friable fill deposits bury the floor in most cases. We are calculating volumes of the void space that exists above modern topographic depression in each case. Our calculations therefore actually correspond to the volume of the caldera that has not been filled by friable materials, lavas or colluvial deposits.

There are two sources of error that lead to overestimation of volumes. The first is related to the erosional breaching of rims of the depression. In fitting a plane to the best estimate of the closed depression, there is still some additional volume added by calculating void space above the plains surrounding the breached depression. However, we made every effort to avoid this bias as much as possible, and the errors that did occur are likely to have been small. Another bias includes the calculation of void space within superimposed impact craters that have interior depressions rivalling the depth of the caldera itself (see Supplementary Fig. 4). However, these errors are again extremely small and do not change the calculated volumes appreciably.

Could the depressions have formed by pseudokarst? Mars is in many ways a periglacial planet. Permafrost is likely to be (and to have been) much more widespread and geologically important at the global scale on Mars than on Earth. Catastrophically melted subsurface ice has been postulated as a likely source for water that carved immense outflow channels on the surface. It has also implicated in the formation of terrains bearing periglacial features such as fields of pitted terrain, as seen in some parts of the Elysium basin. The possibility that the collapse features described in this work could have formed from the removal of subsurface ice there bears consideration.

To test this hypothesis we used the volume calculations described above to constrain how much ice must have been removed to produce the collapse by removal of ice from the subsurface. Models describing the amount and distribution of subsurface ice on Mars have been produced[36]. These calculations include models of subsurface porosity as a function of depth. Using those models of porosity, we can then calculate the amount of pore space that could potentially have been filled with ice beneath a given feature. In other words, is there enough pore space available that, even if entirely filled with subsurface ice, would result in the collapse volume of the depression if all of that ice were removed?

The best test case is Eden patera. Here, ~4,000 km^3 of void space exists. If that space was created by means of collapse that was related to removal of ice, it stands to reason that the ice must have been present essentially beneath the feature itself. If the ice was widely distributed in area, its removal would probably have produced multiple small collapse pits (if any at all) or regional subsidence. We therefore focus on the area of the depression itself. In the case of Eden patera, this area is roughly equal to 5,000 km^2.

Supplementary Figure 4 shows the decay of porosity with depth on Mars and the cumulative volume of void space beneath an area of 5,000 km^2 beneath Eden patera. Pore space decays to near zero by a depth of ~10 km. If all of the void space to this depth were completely filled with ice, it would result in a total volume of ~4,000 km^3—roughly equal to the volume of collapse at Eden patera. Therefore, the calculations, to first order, suggest that the volume of collapse at Eden patera could potentially be explained theoretically by the removal of subsurface ice. However, we suggest that the calculations present a compelling case that ice was not solely responsible for the formation of the collapse at Eden patera because they imply that all of the void space became filled with ice to a great depth and then all of that ice was somehow removed from the subsurface without leaving any traces of fluvial features (that is, outflow channels) that could be related to catastrophic melting.

These volume estimates provide some constraints on the amount of material that was erupted from plains-style caldera complexes in the northern Arabia Terra region. The volumes of the depressions represent, in the strictest sense, the amount of void space produced by a combination of structural collapse and eruption of lavas and/or pyroclastics. Structural collapse could occur as a result of withdrawal of magma, or migration of a magma chamber at depth, and the voids therefore do not necessarily relate directly to erupted volumes. However, explosive eruptions often continue to fragment magma within the volcano's conduit, and the final caldera volume can also be an underestimate of an eruption's total volume. These calculations therefore provide some guidance on the scale of the eruptive potential of the Arabia volcanic province.

Assuming that the void space within calderas relates directly to the removal of magma during eruptions, we can produce some simple scaling calculations to estimate how much material may have been erupted. By assuming a dense rock equivalent (DSE) of caldera volume equal to a typical mafic magma with density of 2,800 kg m^{-3}, we can then scale the DSE for a lava density of 2,000 kg m^{-3} or a tephra of density 1,000-1,500 kg m^{-3}. Using these scaling factors and the volume calculations described above, we calculated the estimated minimum erupted volumes described in the text.

doi:10.1038/nature12535

Genomic organization of human transcription initiation complexes

Bryan J. Venters[1]† & B. Franklin Pugh[1]

The human genome is pervasively transcribed, yet only a small fraction is coding. Here we address whether this non-coding transcription arises at promoters, and detail the interactions of initiation factors TATA box binding protein (TBP), transcription factor IIB (TFIIB) and RNA polymerase (Pol) II. Using ChIP-exo (chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing), we identify approximately 160,000 transcription initiation complexes across the human K562 genome, and more in other cancer genomes. Only about 5% associate with messenger RNA genes. The remainder associates with non-polyadenylated non-coding transcription. Regardless, Pol II moves into a transcriptionally paused state, and TBP and TFIIB remain at the promoter. Remarkably, the vast majority of locations contain the four core promoter elements—upstream TFIIB recognition element (BRE_u), TATA, downstream TFIIB recognition element (BRE_d), and initiator element (INR)—in constrained positions. All but the INR also reside at Pol III promoters, where TBP makes similar contacts. This comprehensive and high-resolution genome-wide detection of the initiation machinery produces a consolidated view of transcription initiation events from yeast to humans at Pol II/III TATA-containing/TATA-less coding and non-coding genes.

The classic model for assembling the minimal core transcription machinery at mRNA promoters starts with the recruitment of TBP to the TATA box core promoter element (CPE)[1]. Next is the docking of TFIIB, which straddles TBP and locks onto flanking TFIIB-recognition elements (BRE_u and BRE_d)[2,3]. Together with TFIIF, TFIIB then engages Pol II in its active site to help set the transcription start site (TSS) at INR[4–6]. The recruitment of the transcription machinery has long been thought to be an important rate-limiting step in gene expression[7]. Concepts in transcription initiation by all three RNA polymerases (I, II and III) have been guided by this basic theme[8].

Clashing with this seemingly simplified view is that the TATA box has been identified at only ~10% of human promoters[9,10], with most genes ostensibly being classified as 'TATA-less' in all three RNA polymerase systems. The other CPEs are apparently equally rare. A second complication of the classic view, particular to multicellular eukaryotes, is that the general transcription factors may be largely pre-assembled at promoters. There, Pol II is in a transcriptionally engaged but paused state, approximately 30–50 base pairs (bp) downstream from the TSS[11–13]. A third complication is that transcription of genomes is not restricted to coding genes, but seems to be quite pervasive, without clear evidence of being coupled to definable promoters[14]. These complications, together, paint a seemingly complex picture of eukaryotic transcription initiation.

Towards reconciling simplistic models against complex data, we recently developed the ChIP-exo assay to map sites of protein–DNA interactions at near single-base resolution[15]. We discovered in yeast that so-called TATA-less promoters actually possess degenerate versions of the TATA box, and that most yeast promoters assemble the transcription machinery fundamentally in accord with the classic model, although a deep dichotomy between the TATA/SAGA/stress-induced genes and TATA-less/TFIID/housekeeping genes remains. This led us to consider whether similar simplicity was true in humans, albeit with the additional complications of paused polymerase and pervasive non-coding transcription.

TBP/TFIIB separation from paused Pol II

Using ChIP-exo, we detected 159,117 TFIIB locations (peak pairs) in K562 cells, of which 36% were associated with ENCODE-defined transcriptional domains[17] (Extended Data Fig. 1a). Remarkably, half were associated with heterochromatic regions, which are generally thought to be devoid of stable RNA production. However, heterochromatic transcription may be more pervasive than previously thought.

We assigned a TBP/TFIIB location to more than 50% of all annotated protein-coding K562-expressed genes (Extended Data Fig. 1b), thereby providing independent validation. Seemingly expressed genes that lacked a TBP–TFIIB location may have arisen from several sources including rare but stable mRNAs, detection noise, and antisense transcription arising from a more distal promoter. TBP/TFIIB/Pol II occupancy and mRNA levels were correlated (Extended Data Fig. 1c), as expected of recruitment being at least partially rate-limiting in gene expression.

We initially focused on all 8,364 K562 TFIIB locations near the TSS of 6,511 coding RNAs as defined by RefSeq[18]. Figure 1a provides one example of the raw tag distribution and the identified CPEs concentrated ~25 bp upstream of the *RPS12* ribosomal protein gene TSS. When individual genes were examined (Fig. 1b), or averaged across all 6,511 genes (Fig. 1c), two regions of high TFIIB/TBP/Pol II occupancy were observed. The major rightward peaks corresponded to primary promoter transcription initiated complexes (Fig. 1c, top panel). Those in the leftward direction matched divergent TSSs[19–22], although the resulting RNA was less abundant than expected from TFIIB/TBP/Pol II occupancy levels (Fig. 1c, bottom versus top panel; note that secondary TSS represents only 24% of the total TSS signal). This may result from RNA instability, as seen in yeast. The clear spatial separation of complexes indicates that divergent transcripts arise from distinct initiation complexes, most (78%) of which were in CpG islands. On average, two complexes were detected per CpG island[23], regardless of island length, with the centre of the island being enriched ~100 bp downstream of the primary TSS (Extended Data Fig. 2a, b). Complexes

[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. †Present address: Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA.

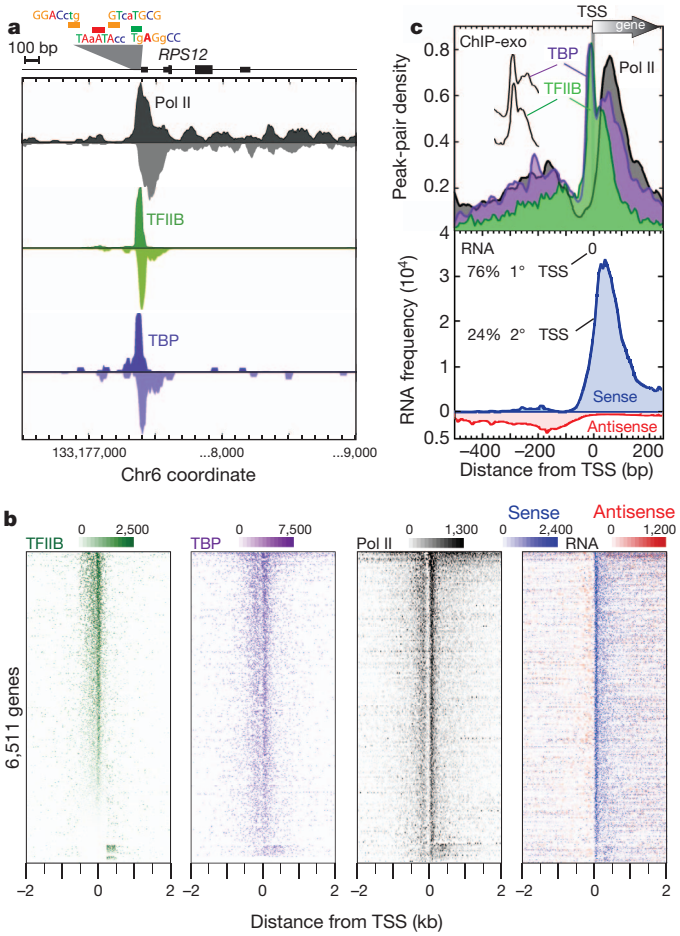


Figure 1 | Transcription machinery organization at human mRNA promoters. **a**, Smoothed distribution of strand-separated ChIP-exo tag 5′ ends at the *RPS12* gene. Core promoter elements (CPEs) are shown with lower casing denoting mismatches to the consensus. **b**, Peak-pair distribution of RNA at RefSeq genes (rows). Rows are linked, and sorted by TFIIB occupancy. **c**, Top, averaged ChIP-exo patterns around the closest (primary) RefSeq TSS. The 'spikes' of TBP and TFIIB are indiscernible (vertically offset in inset). Bottom, distribution of secondary polyadenylated RNA[38], with traces separated by sense (blue) and antisense (red, inverted trace) orientations relative to the corresponding mRNA TSS.

tended to be separated by 70–180 bp (Extended Data Fig. 2c, red), and had largely uncorrelated occupancies (Extended Data Fig. 2c, black), which suggests that they are generally regulated independently.

For the vast majority of transcription units, Pol II crosslinked 50 bp downstream of the primary TSS (Fig. 1b, c), where it is expected to pause after initiating transcription[13]. Pol II was most depleted over the core promoter, indicating that it does not stably reside there in proliferating K562 cells. Therefore, when Pol II enters the core promoter, it rapidly initiates transcription and then moves into a paused state ~50 bp downstream, thereby preventing any new polymerase from detectably engaging the core promoter.

The crosslinking pattern of human TFIIB was of particular interest because TFIIB in budding yeast crosslinks broadly across the relatively stable single-stranded DNA region within the Pol II active site at core promoters[16], in accord with crystallographic models of 'open' complexes[24]. Remarkably, human TFIIB maintained its contact within this region, despite the absence of polymerase (Fig. 1c, top). Mechanistically, this might occur via TFIIB contacts with BRE_d (ref. 3 and see below), which are absent in budding yeast. The coincidence of TBP and TFIIB crosslinking at the BRE_d suggests that TBP may be predominantly crosslinking to TFIIB there, rather than directly to the TATA element.

BRE_u, TATA, BRE_d and INR are common

We looked for CPEs (illustrated in Fig. 2a) within the narrow intervals defined by 8,364 mRNA TSS-proximal TFIIB locations. Remarkably, and consistent with yeast[16], nearly 85% of them had a sequence with 0–3 mismatches to the TATA-box consensus (TATAWAWR)[25] (Fig. 2b, c). Less than 3% had a perfect match to the consensus. Deviations from the TATA-box consensus inversely correlated with TFIIB and TBP occupancy levels (Extended Data Fig. 3a), indicating that the TATA element sequence quality contributes to their occupancy level, consistent with previous observations[26] on their *in vivo* functionality.

Several controls put the false positive rate for TATA elements at ~20% (Fig. 2c). First, 10,000 randomly generated sequences having the same human genome sequence bias found that only 16% were called by chance. Second, a scrambled version of the motif (having 0–3 mismatches) was identified only 20% of the time, and had no positional relationship with TFIIB/TBP (not shown). Third, coordinates having a single isolated tag were used to generate an essentially random set of false-positive locations, and the analysis repeated. TATA elements (0–3 mismatches) were identified only 20% of the time. Fourth, whereas control sequences were distributed randomly across the query space, the distribution of TATA elements was not random. Instead it displayed a tight peak 20 bp upstream of the TFIIB and TBP locations (Fig. 2d and data not shown).

TFIIB in complex with TBP makes sequence-specific contacts with BRE_u and BRE_d, which flank the TATA box[2,3] and are upstream of the INR (Fig. 2a). However, these elements are essentially non-existent in yeast, and ill-defined across mammalian genomes. Using the identified TATA elements as a reference point, we searched upstream for BRE_u and downstream for BRE_d and INR. Notably, in nearly every

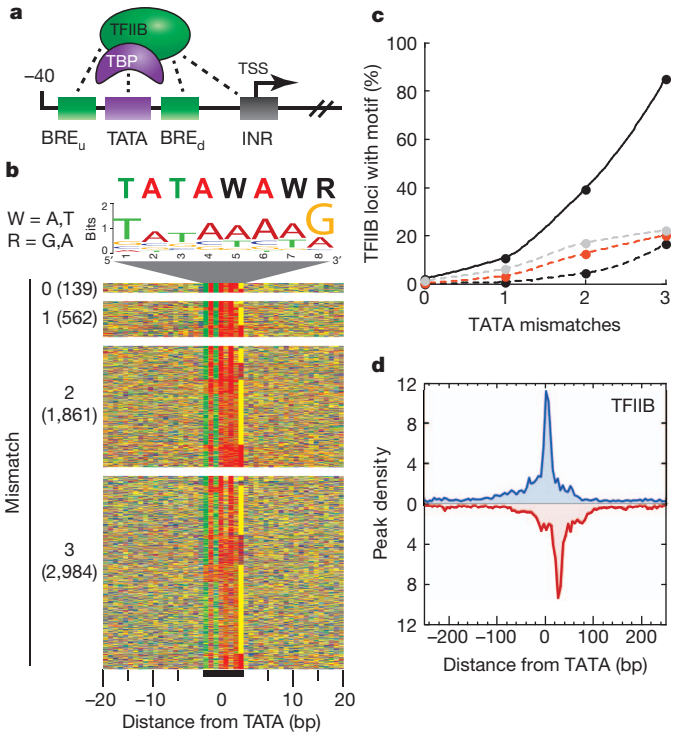


Figure 2 | TATA elements at most mRNA genes. **a**, Core promoter schematic. **b**, Nucleotide distribution for TATA elements with 0–3 mismatches (panels) to the consensus, and sorted by ascending *P* value. Colours are reflected in the logo colour. **c**, Cumulative percentage of TFIIB locations having a TATAWAWR sequence with 0–3 mismatches (solid line). Controls include a randomized sequence (60% GC, dashed black line), a scrambled consensus (dashed red line), and 8,364 locations represented by a single background tag (dashed grey line). **d**, Distance of strand-specific TFIIB peaks (exonuclease stop sites) from TATA element midpoints. Opposite-strand peaks are in red and inverted.

instance a sequence with three or less mismatches to the literature-derived consensus for BRE_u (SSRCGCC)[2], BRE_d (RTDKKKK)[3] and INR (YYANWYY)[27] was found (Fig. 3a–c). Remarkably, sequences within BRE_d and INR seemed to co-vary. For example, the BRE_d consensus tended towards either GTKGGGG or ATKTTTT, rather than an equal mixture of all possible combinations (Fig. 3b), making them less degenerate than the consensus would suggest. Similarly, the INR consensus tended towards either CCANWCC or TTANWTT (Fig. 3c). Sequence bifurcation was not observed with TATA or BRE_u elements. Given the strong bias towards either strong (G/C) or weak (A/T) base pairing, this sequence dimorphism may reflect selection for distinct thermodynamic stabilities towards helix melting, which is an essential first step in initiation at these elements. Consistent with this, A/T-rich BRE_d and INR elements had substantially higher crosslinking levels of TFIIB than their G/C-rich counterparts (not shown). However, this may not explain the strand bias of the sequences.

Similar to our TATA analysis, the TFIIB peak density was tightly focused at a fixed distance from each CPE (Fig. 3d), and were rarely found in randomized controls (Fig. 3e), thereby validating them. TFIIB peak pairs were centred over BRE_d, suggesting that the primary crosslinking point is at or near the BRE_d. Unlike the TATA element, the BRE and INR elements deviated relatively little from their consensus (compare Figs 2c and 3e), and such deviations did not correlate with TBP and TFIIB occupancy levels (not shown). Thus, BRE and INR sequence variability may regulate occupancy of the basal initiation complex to a lesser extent than TATA. Within their search space, the locations of each CPE peaked at previously defined canonical positions (Fig. 3f and Extended Data Fig. 3b), thereby providing cross-validation and a core promoter consensus: SSRCGCCTATAWAWRNRTDKKKK$(N)_{13}$YYANWYY. The tolerance for mismatches in these elements seems to be 2-3-2-1, respectively.

150,000 non-coding initiation complexes

We next examined the remaining 150,753 putative TFIIB locations that were far (>500 bp) from a protein-coding gene (Supplementary Data 1). At a 20% false discovery rate per element, we identified at least three of the four CPEs at 97% of all non-mRNA TFIIB locations (Extended Data Fig. 4a). Deviations from the consensus were no more than at mRNA genes (average of 5 deviations across 28 positions within the four CPEs). TBP, TFIIB and Pol II peaked at the same canonical distances from each motif as found at mRNA promoters (Extended Data Fig. 4b, c). They were also embedded in a similar chromatin environment as mRNA promoters (Fig. 4a, b), but displayed comparatively lower TFIIB occupancy (Extended Data Fig. 4d).

Remarkably, TBP/TFIIB/Pol II complexes were linked to the production of non-polyadenylated RNA (87% had them) rather than polyadenylated transcripts (Fig. 4c and Extended Data Fig. 5), which is in agreement with the finding of enhancer RNAs[28]. Their locations mapped precisely to the location of TFIIB. Non-polyadenylated transcript levels also correlated more strongly with 'non-coding' TFIIB occupancy than did polyadenylated levels (Fig. 4d), further validating the link. Taken together, we conclude that the vast majority of all 159,117 TFIIB locations (non-coding plus coding) detected in K562 cells represent bona fide and fundamentally identical core promoter initiation complexes, of which ~5% produce mRNA and ~95% produce RNA that is non-polyadenylated and non-coding.

Restricted motif spacing in promoters

We searched for an overall core promoter consensus (SSRCGCCTATAWAWRNRTDKKKK$(N)_{13}$YYANWYY) and CPE-spacing variants within 60 bp of all TFIIB locations, and plotted their distribution relative to TFIIB (Extended Data Fig. 6). Remarkably, the consensus spacing defined in Fig. 3f displayed the strongest positional relationship with TFIIB (Fig. 5a). For example, a consensus having the spacing between BRE_u and TATA reduced by 1 bp displayed almost no positional relationship with TFIIB, as would be expected of a random/non-functional sequence.

There was very little or no tolerance for variable spacing between CPEs, which reflects structural constraints of the initiation complex[5]. Surprisingly, proper spacing was accompanied by greater sequence deviations within individual CPEs (thick versus thin black line in Fig. 5a), whereas small spacing deviations were accompanied by stronger elements (Fig. 5b). In short, core promoters may be weak by design, through a compensatory balance of sequence and spacing deviations from the consensus. This allows for greater dependence on transcriptional activators, but also provides for a specified basal output.

We conducted ChIP-exo mapping of TFIIB locations across four ENCODE cancer cell lines: HeLa S3, HepG2 and MCF7 in addition to K562 (cervical, liver, breast and blood, respectively). We detected TFIIB at 9,074 mRNA genes in at least one cell line, and at 1,691 genes in all lines (group 1 in Extended Data Fig. 7). Cluster analysis suggested that although TFIIB occupancy levels varied from gene to gene, most were relatively constant at individual genes across cell lines.

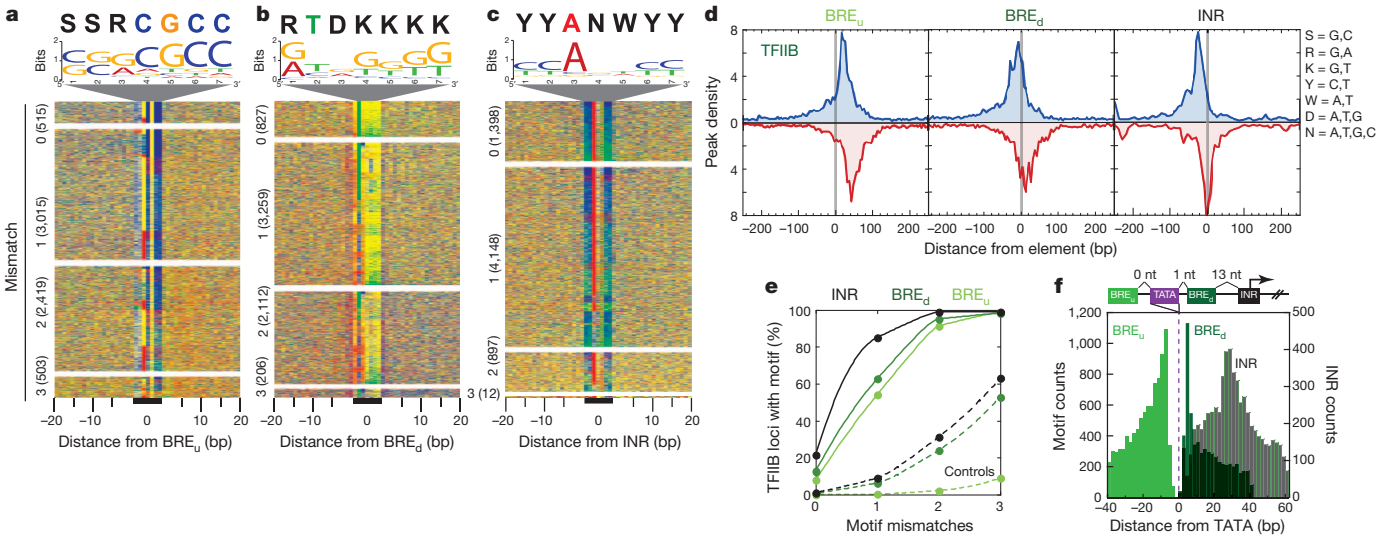


Figure 3 | BRE and INR at most mRNA genes. **a–c**, Nucleotide distribution for BRE_u, BRE_d and INR, vertically separated by 0–3 mismatches to the consensus, and sorted by ascending *P* value within panels. **d**, Distance of strand-specific TFIIB peaks from BRE_u, BRE_d and INR. Opposite-strand peaks are in red and inverted. **e**, Cumulative percentage of genes with 0–3 mismatches to each motif in **a–c**. Controls were randomized sequences (60% GC, dashed lines). **f**, Distribution of CPEs relative to TATA box midpoints.

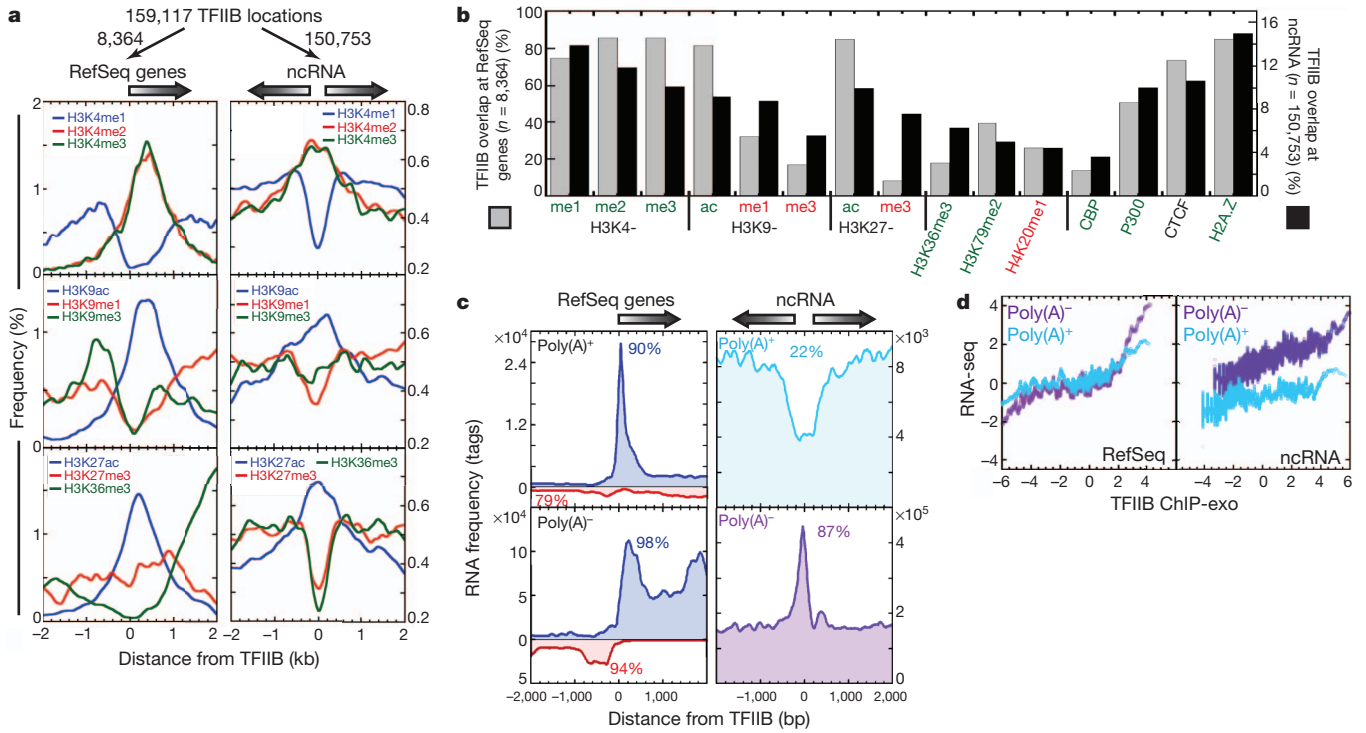


Figure 4 | Non-coding TFIIB locations have chromatin marks and non-polyadenylated RNA. **a**, Distribution of chromatin marks around TFIIB at RefSeq genes (left) and ncRNA (right). **b**, TFIIB locations that overlap with chromatin marks and epigenetic regulators[39]. **c**, Distribution of polyadenylated[38] and non-polyadenylated[40] RNA-seq tags around TFIIB >500 bp from a RefSeq TSS. Percentages reflect TFIIB having an RNA tag <2 kb away. Left panels include sense (blue) and antisense (red and inverted) strands for RefSeq genes, which was not applied to ncRNA (right panels). **d**, 100-gene moving average of polyadenylated and non-polyadenylated RNA levels versus TFIIB occupancy at mRNA and ncRNA genes (left and right panels, respectively) on a median-centred $\log_2$ scale.

About one-third displayed noticeable cell-type specificity (for example, group 3 in Extended Data Fig. 7). For non-coding initiation complexes, we focused on those present in two or more cell types, and found 100,349 such locations (376,074 locations were found in at least one cell type). Non-coding complexes seemed to have more cell-type specificity and were bimodally distributed at high and low occupancy levels. This heterogeneity may reflect more numerous and diverse roles for the resulting non-coding transcription and/or RNA in cell-type specific physiology compared to proteins.

tRNA genes have TATA and BRE

With some exception[29], transfer RNA genes have been classically defined as TATA-less, in which TFIIIC recognizes specific sequences downstream of the TSS, then recruits TFIIIB to a region immediately upstream of the TSS that lacks apparent sequence specificity[30,31]. Pol III then binds to form an initiation complex. TFIIIB contains TBP (and BRF, a factor related to TFIIB), and thus it has been enigmatic as to how TBP in TFIIIB engages the upstream region without a TATA box.

Remarkably, TBP crosslinked ~21 bp upstream of 386 tRNA genes (Fig. 6a, left), as seen at Pol II promoters. In nearly every instance we found a TATA element (Fig. 6a, middle) that was ~18 bp further upstream (Fig. 6b). Similar to TBP crosslinking through TFIIB, we suspect that TBP crosslinks through BRF. Indeed, the peaks of BRF and TBP crosslinking are coincident at Pol III genes in mice[32]. As with Pol II promoters, we found a BRE_d centred between each TBP peak pair (Fig. 6a, right panel) and a BRE_u immediately upstream of TATA (not shown). Enrichment of these elements, but not the Pol-II-specific INR[33], were statistically significant (Fig. 6c). Thus, TBP in complex

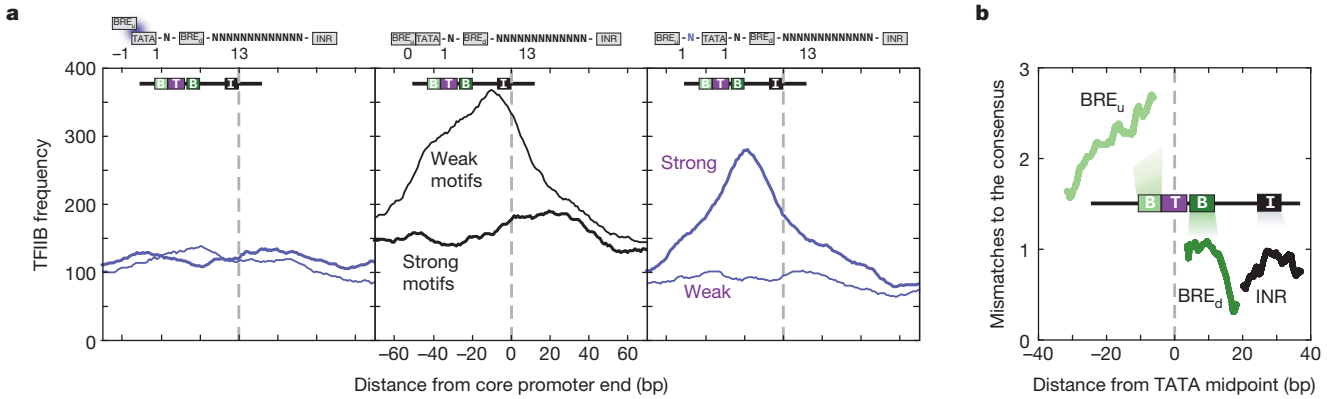


Figure 5 | Restricted spacing of CPEs. **a**, Candidate core promoter enrichment at varying distances from all 159,117 TFIIB locations, for spacing variants having motifs with weak and strong *P* values. **b**, Moving average of mismatches to the indicated motif consensus as a function of distance from TATA.

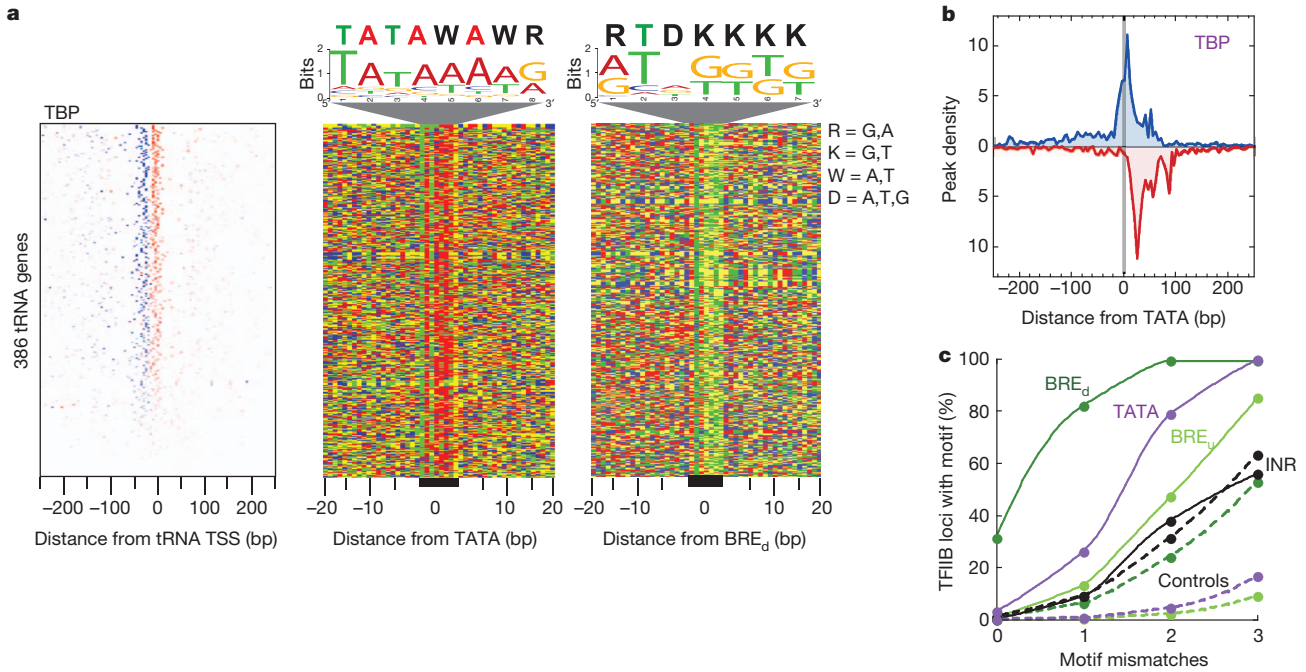


Figure 6 | TATA and BRE elements at most tRNA genes. **a**, Left, TBP peak density separated by forward and reverse strand orientation (blue and red colours, respectively) relative to each tRNA TSS. Corresponding sequences are shown in the right two panels (provided in Supplementary Data 4). **b**, Average distribution of TBP peaks around all identified tRNA TATA elements. **c**, Cumulative percentage of tRNA genes with the indicated promoter element having 0–3 mismatches to the consensus. Dashed lines represent calculations for an equivalent number of randomized sequences for the colour-linked solid traces.

with a TFIIB family member engages a set of BRE$_u$–TATA–BRE$_d$ CPEs similarly in Pol II and III systems.

Consolidated genomic view of initiation

Genome-wide mapping of the general transcription machinery at near single-base resolution offers a consolidated model of certain transcription initiation events from yeast to humans, Pol II to Pol III, TATA-containing to TATA-less, and mRNA to non-coding RNA (ncRNA). In general, a TFIIB/BRF family member is recruited to all coding or non-coding core promoters via a TBP family member and spatially-constrained CPEs. Sequence-specific (BRE$_d$) contact with the DNA a few base pairs downstream of TATA might 'bookmark' the site of DNA melting for a rapidly departing Pol II or III. Yeast Pol II is relatively slow to depart, and so it produces equivalent TFIIB-open promoter contacts/crosslinking in the absence of BRE$_d$. Pol II then scans downstream several base pairs, where it encounters an INR that allows for productive transcription, which subsequently pauses 30–50 bp further downstream. In yeast, where an INR and pausing appear absent, a nucleosome border may help to set the start site of productive transcription.

Although core promoters are seemingly long (~43 bp in human) for sequence-specific binding, they are designed to be inherently low in specificity, presumably to keep basal transcription low and to maintain high dependence on transcriptional activators. Appropriate specificity is achieved via a blend of degeneracy in motif sequence and spacing. Broad clusters of TSSs at mammalian genes[4] can therefore be explained in terms of clusters of weak core promoters, many of which may fall below bioinformatic detection.

The discovery that transcription of the human genome is vastly more pervasive than what produces coding mRNA raises the question as to whether Pol II initiates transcription promiscuously through random collisions with chromatin as biological noise or whether it arises specifically from canonical Pol II initiation complexes in a regulated manner. Our discovery of ~150,000 non-coding promoter initiation complexes in human K562 cells and more in other cell lines suggests that pervasive non-coding transcription is promoter-specific, regulated, and not much different from coding transcription, except that it remains nuclear and non-polyadenylated. An important next question is the extent to which transcription factors regulate production of ncRNA.

We detected promoter transcription initiation complexes at 25% of all ~24,000 human coding genes, and found that there were 18-fold more non-coding complexes than coding. We therefore estimate that the human genome potentially contains as many as 500,000 promoter initiation complexes, corresponding to an average of about one every 3 kilobases (kb) in the non-repetitive portion of the human genome. This number may vary more or less depending on what constitutes a meaningful transcription initiation event. The finding that these initiation complexes are largely limited to locations having well-defined core promoters and measured TSSs indicates that they are functional and specific, but it remains to be determined to what end. Their massive numbers would seem to provide an origin for the so-called dark matter RNA of the genome[34], and could house a substantial portion of the missing heritability[35].

METHODS SUMMARY

Human cells were grown, treated with formaldehyde, and processed through the ChIP-exo assay as described in the Methods and elsewhere[36]. Sequence tags were normalized to input, peaks were called and paired, and pairs with >4 tags retained. FIMO[37] was used to find literature-defined motifs within pre-defined distances from TFIIB peak-pairs or TATA elements. Illumina sequencing statistics, TFIIB locations, and Position-Specific Percentage Matrix (PSPM) tables are presented in Extended Data Table 1 and Supplementary Data 1 and 2, respectively.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 March; accepted 9 August 2013.

Published online 18 September; corrected online 2 October 2013 (see full-text HTML version for details).

1. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56,** 549–561 (1989).
2. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription:

sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **12,** 34–44 (1998).
3. Deng, W. & Roberts, S. G. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **19,** 2418–2423 (2005).
4. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **339,** 225–229 (2010).
5. Kostrewa, D. *et al.* RNA polymerase II–TFIIB structure and mechanism of transcription initiation. *Nature* **462,** 323–330 (2009).
6. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495,** 481–486 (2013).
7. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386,** 569–577 (1997).
8. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45,** 439–446 (2012).
9. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (2005).
10. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
11. Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced *hsp70* gene in *Drosophila melanogaster* cells. *Mol. Cell. Biol.* **6,** 3984–3989 (1986).
12. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).
13. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 (2013).
14. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
15. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–1419 (2011).
16. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483,** 295–301 (2012).
17. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
18. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65 (2007).
19. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322,** 1855–1857 (2008).
20. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).
21. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319,** 1791–1792 (2008).
22. Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* **22,** 2399–2408 (2012).
23. Rozenberg, J. M. *et al.* All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC Genomics* **9,** 67 (2008).
24. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II–TFIIB complex. *Nature* **493,** 437–440 (2013).
25. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116,** 699–709 (2004).
26. Singer, V. L., Wobbe, C. R. & Struhl, K. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev.* **4,** 636–645 (1990).
27. Smale, S. T. & Baltimore, D. The ''initiator'' as a transcription control element. *Cell* **57,** 103–113 (1989).
28. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
29. Hamada, M., Huang, Y., Lowe, T. M. & Maraia, R. J. Widespread use of TATA elements in the core promoters for RNA polymerases III, II, and I in fission yeast. *Mol. Cell. Biol.* **21,** 6870–6881 (2001).
30. Geiduschek, E. P. & Tocchini-Valentini, G. P. Transcription by RNA polymerase III. *Annu. Rev. Biochem.* **57,** 873–914 (1988).
31. White, R. J. & Jackson, S. P. Mechanism of TATA-binding protein recruitment to a TATA-less class III promoter. *Cell* **71,** 1041–1053 (1992).
32. Carrière, L. *et al.* Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. *Nucleic Acids Res.* **40,** 270–283 (2012).
33. Verrijzer, C. P., Chen, J. L., Yokomori, K. & Tjian, R. Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. *Cell* **81,** 1115–1125 (1995).
34. Kapranov, P. & St Laurent, G. Dark matter RNA: existence, function, and controversy. *Front. Genet.* **3,** 60 (2012).
35. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Rev. Genet.* **11,** 446–450 (2010).
36. Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* **Chapter 21,** Unit 21.24 (2012).
37. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).
38. Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20,** 413–427 (2010).
39. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
40. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Reja, S. Mahony, P. Albert and Y. Li for bioinformatic assistance, and M. Cousar and K.-Y. Chan-Salis for experimental support. This work was supported by National Institutes of Health grant GM059055.

Author Contributions B.J.V. performed the experiments and conducted data analyses. B.J.V. and B.F.P. conceived the experiments, analyses and co-wrote the manuscript.

Author Information Sequencing data have been deposited at the NCBI Sequence Read Archive under accession number SRA067908. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.F.P. (bfp2@psu.edu).

METHODS

Cell culture. Human chronic myelogenous leukaemia cells (K562, ATCC) were maintained between 1×10^5–1×10^6 cells ml^{-1} in DMEM media supplemented with 10% bovine calf serum at 37 °C with 5% CO_2. Human adenocarcinoma cells from the cervix (HeLa S3, ATCC), liver (HepG2, ATCC) and breast (MCF7, ATCC) were grown in a similar manner as K562 cells except that they were maintained between 25% and 90% confluence. Cells were washed in PBS (1× PBS, 8 mM Na_2HPO_4, 2 mM KH_2PO_4, 150 mM NaCl and 2.7 mM KCl) before incubation with formaldehyde in a final concentration of 1% for 10 min. Cells were lysed (10 mM Tris, pH 8, 10 mM NaCl, 0.5% NP40, and complete protease inhibitor cocktail (CPI; Roche)), and then the nuclei lysed (50 mM Tris, pH 8, 10 mM EDTA, 0.32% SDS, and CPI). Purified chromatin was resuspended in immunoprecipitation dilution buffer (40 mM Tris, pH 8.0, 7 mM EDTA, 56 mM NaCl, 0.4% Triton X-100, 0.2% SDS, and CPI) and sonicated with a Bioruptor (Diagenode) to obtain fragments with a size range between 100 and 500 bp.

ChIP-exo and antibodies. With the following modifications, ChIP-exo was carried out as previously described[36] with chromatin extracted from 10 million cells, ProteinG MagSepharose resin (GE Healthcare), and 3 μg of TFIIB (Santa Cruz Biotech, sc-225), TBP (Santa Cruz Biotech, sc-204) or Pol II (Santa Cruz Biotech, sc-899, directed against the N terminus of the Pol II large subunit encoded by POL2RA).

Alignment to genome, peak calling and data access. Libraries were sequenced on an Illumina HiSeq sequencer. The entire length of the sequenced tags was aligned to the human hg18 reference genome using BWA[41] with default parameters. Raw sequencing data are available at the NCBI Sequence Read Archive (accession SRA067908). The resulting sequence read distribution was used to identify peaks on the forward (W) and reverse (C) strand separately using the peak calling algorithm in GeneTrack (sigma = 20, exclusion zone = 40 bp)[42]. For strand-specific and strand-merged plots, sequencing tags were normalization to input. All 11,458 locations that were present in the ENCODE designated blacklist were removed from the analysis. Peaks were paired if they were 0–80 bp in the 3′ direction from each other and on opposite strands. Because patterns described here were evident among individual biological replicates, and replicates were well correlated, we merged all tags from biological replicate data sets to make final peak-pair calls. Peak pairs were considered to be TFIIB if they had a tag count of >4 in the merged data sets. A total of 159,117 locations met these criteria. Peak-pair matches across cell lines required that their midpoints be within 80 bp of each other.

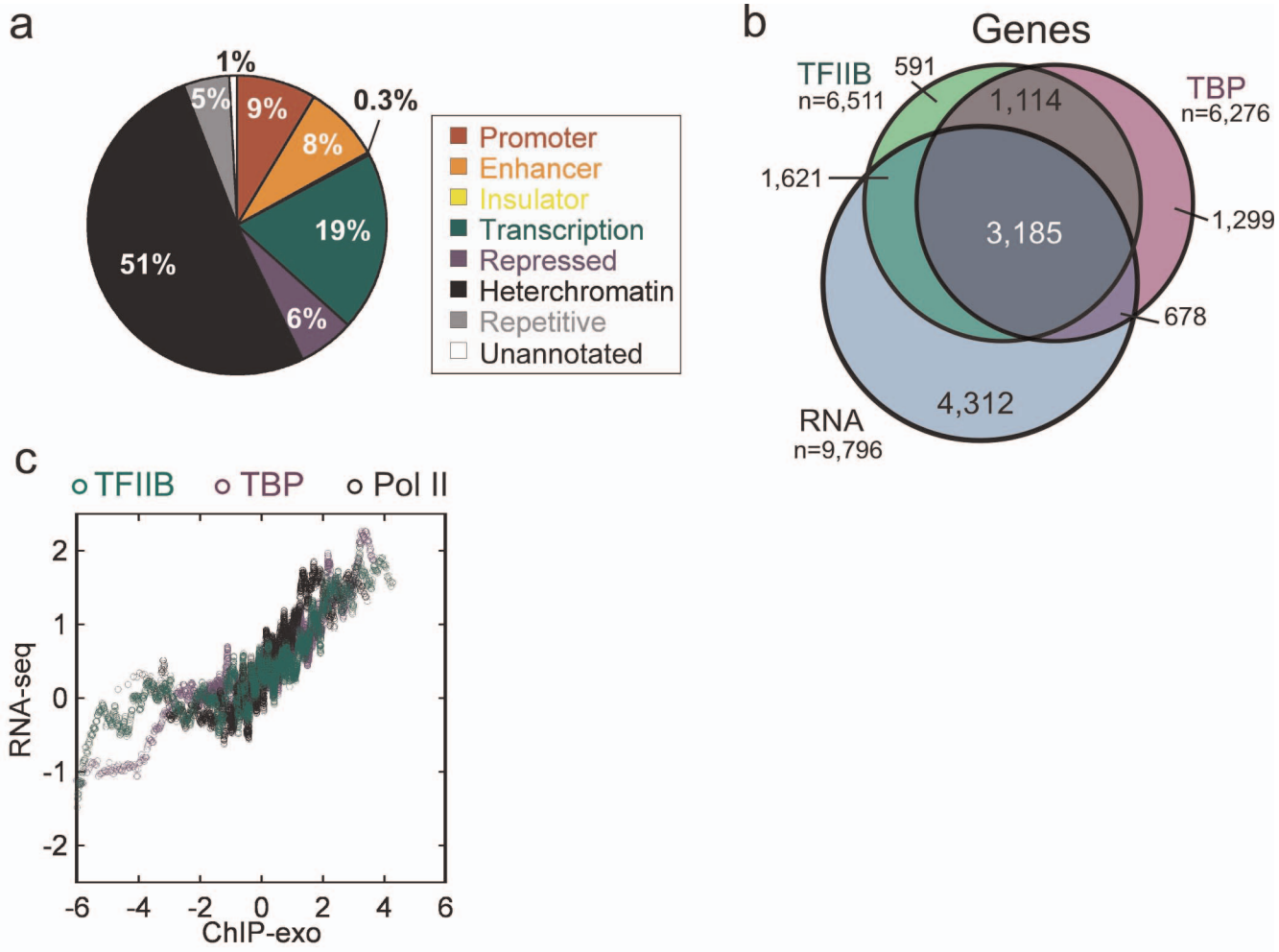
NCBI-curated RefSeq TSSs (n = 26,987)[18] comprising 23,181 non-redundant mRNA genes were considered. Assignment of TFIIB (8,364 peak pairs) and TBP (7,642 peak pairs) to the nearest RefSeq TSS in Supplementary Data 1 required that they be within 500 bp of the TSS, yielding 6,511 non-redundant mRNA genes. Importantly, using a more stringent interval only marginally changed these numbers and did not alter our conclusions. If a gene had >1 TSS, then the TSS nearest to the bound location (peak-pair midpoint) was used as the primary TSS, and other nearby TSSs were considered secondary (Fig. 1c, bottom).

Motif analysis. At each of these 6,511 promoters, using the MEME suite of tools[37], we searched for TATA elements within 80 bp of the midpoint of TFIIB-bound locations on the sense strand, first by searching for the consensus TATAWAWR (Supplementary Data 2), then sequentially for one to three mismatches to the consensus, if an element was not found. In rare cases in which multiple elements were found, we chose the one closest to the TFIIB peak. This rule had no qualitative effect on the data because such events were rare and choosing the furthest element gave the same result (not shown). Moreover, peak motif detection for BRE_u, TATA and INR was not centred over TFIIB, indicating that this distance criterion was not driving the observed motif enrichment at TFIIB locations. Using a similar strategy, we searched for candidate BRE_u element (Supplementary Data 2) within 40 bp upstream of the 5,546 identified TATA elements, and searched for candidate BRE_d and INR elements (Supplementary Data 2) within 40 bp and 60 bp downstream of the 5,546 TATA elements, respectively. At Pol III promoters, candidate BRE_d elements were required to be within 20 bp of a TBP peak-pair midpoint, and in the same orientation as the TATA element.

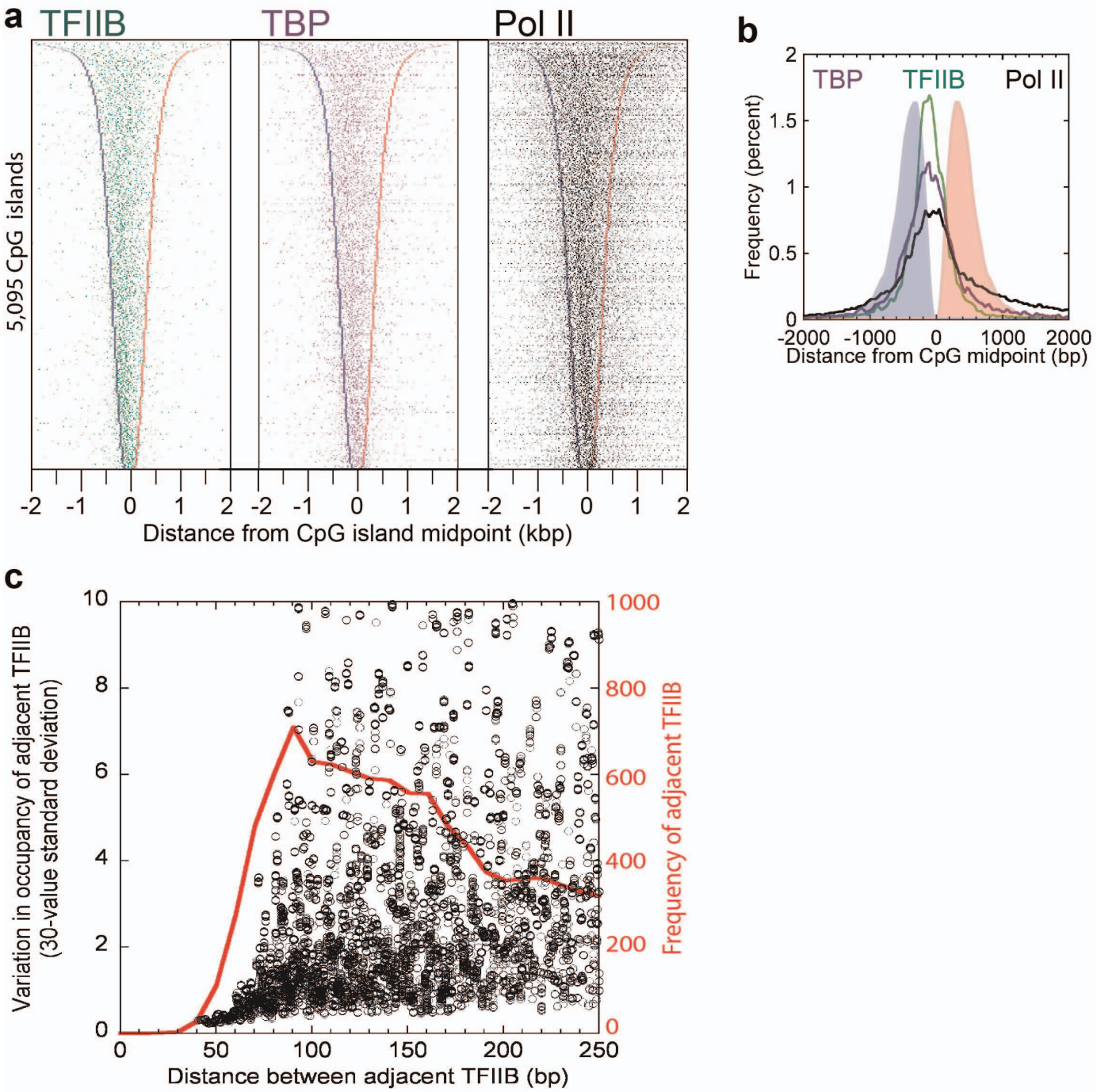
Our searches infrequently picked up multiple motif instances within the search window. Where this did occur, we chose the motif with the best match to the published consensus (not the closest to TFIIB). In the situation where we obtained more than one motif with the same number of mismatches, we chose the one closest to TFIIB. Third, when we discard these multiple occurrences, the results qualitatively did not change. Fourth, the peak locations that we obtained for BRE_u, TATA and INR were not centred over TFIIB. Instead they peaked at the canonical location that had been established in the literature. This provided independent validation.

Using a core promoter (CP) PSPM matrix derived from individual CPE logos and spacing from Figs 2 and 3 (Supplementary Data 2), FIMO[37] was used to find 39–47-bp CP sequences within 120 bp of a TFIIB peak pair, and had either a P value $< 3 \times 10^{-4}$ or between 3×10^{-4} and 10^{-3}. Only the strongest CP in each group was considered. Distances between the two (TFIIB peak-pair midpoint to the CP 3′ end) were then calculated. Their frequency distribution was then plotted as a 11-bp moving average.

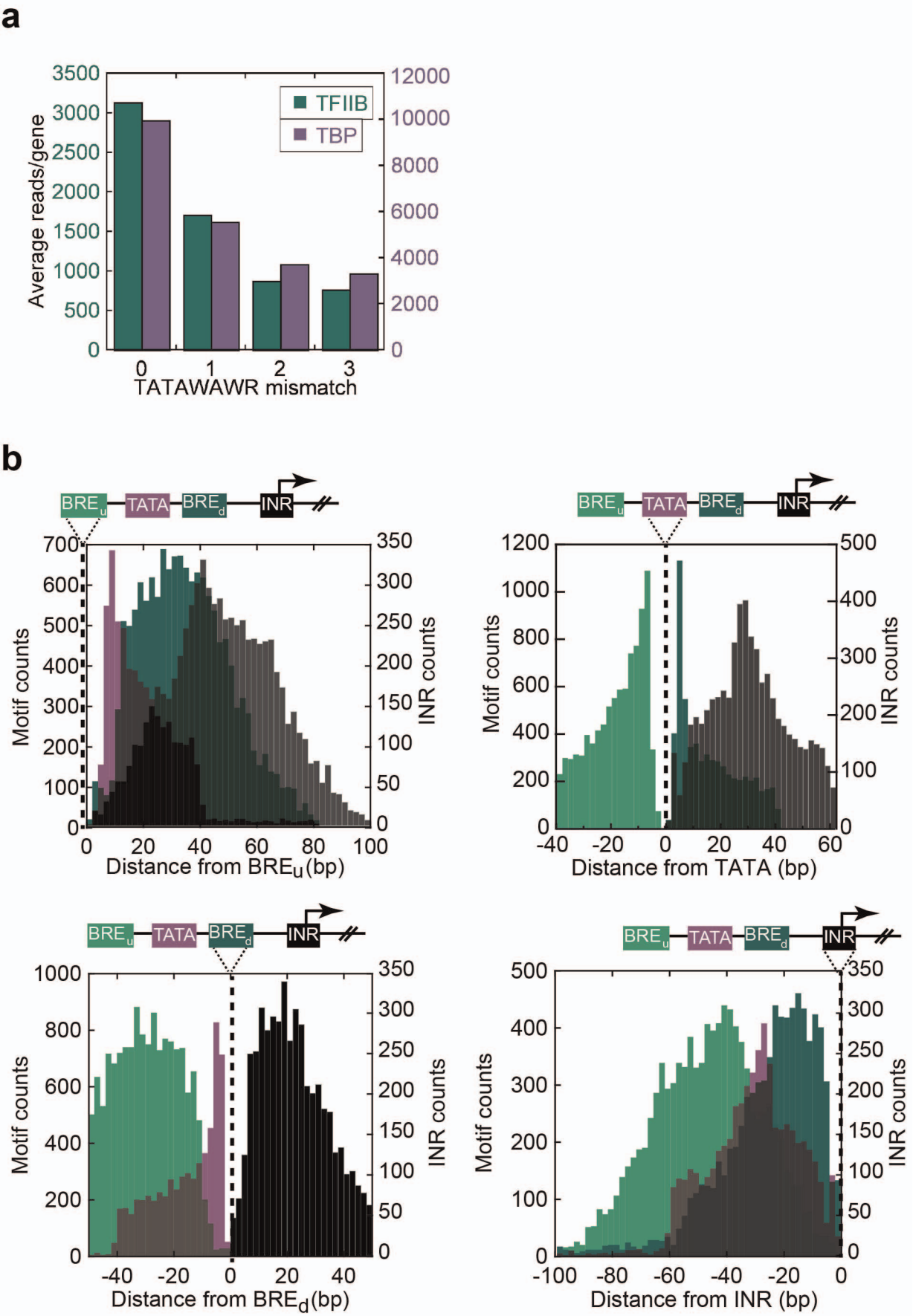
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
42. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack–a genomic data processing and visualization framework. *Bioinformatics* **24,** 1305–1306 (2008).



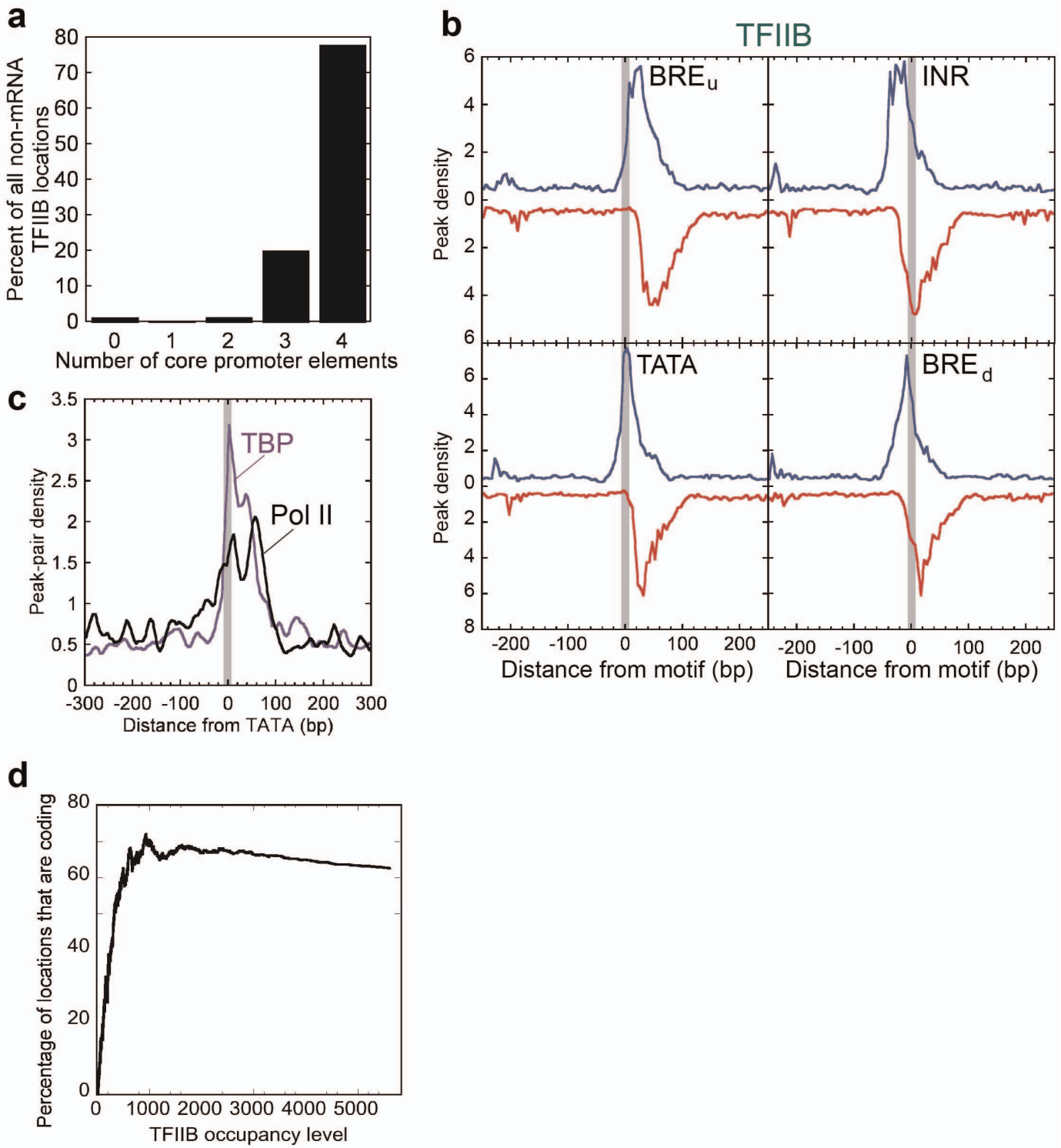
Extended Data Figure 1 | Validation of ChIP-exo data and association with ENCODE annotated regions. **a**, Pie chart of all 159,117 TFIIB-bound locations in K562 cells parsed into ENCODE-annotated regions. **b**, Venn overlap among mRNA genes having TBP or TFIIB locations (<500 bp from its TSS) and genes with measured polyadenylated mRNA levels detected by RNA-seq[38]. Data thresholding may contribute to non-overlapping sets. **c**, Moving average (100-gene) of mRNA levels versus TFIIB/TBP/Pol II occupancy levels on a median-centred $\log_2$ scale.



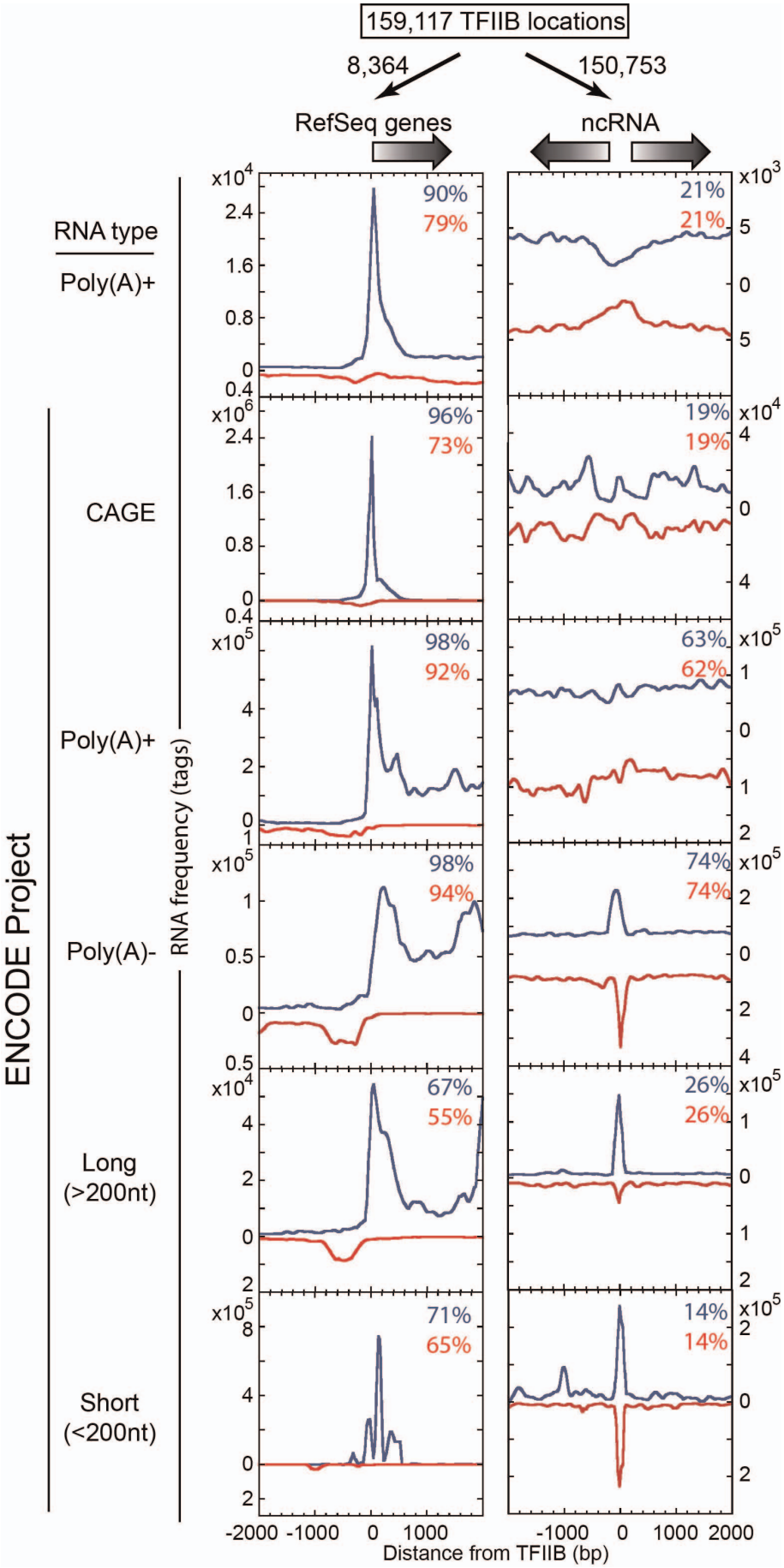
Extended Data Figure 2 | Distribution of TFIIB/TBP/Pol II in CpG islands that overlap mRNA TSSs. **a**, Peak-pair distribution for TFIIB, TBP and Pol II at the 5,095 CpG islands that overlap with the mRNA TSSs from Fig. 1b (78% overlap), and with the direction of transcription to the right. Rows are linked, and sorted by CpG island length. CpG island borders are indicated by blue and red bars, respectively. **b**, Shown is the averaged data from **a**. **c**, All 159,117 TFIIB locations were sorted by location, and inter-TFIIB distances calculated (red trace). Data were then sorted by distance, and the standard deviation of adjacent TFIIB occupancy ratios was calculated on a sliding window of 30 values. Peak calling parameters preclude detection of two separate TFIIB locations approximately <40 bp apart. Those that were 40–70 bp apart were correlated, whereas those >70 bp apart were less correlated or uncorrelated.



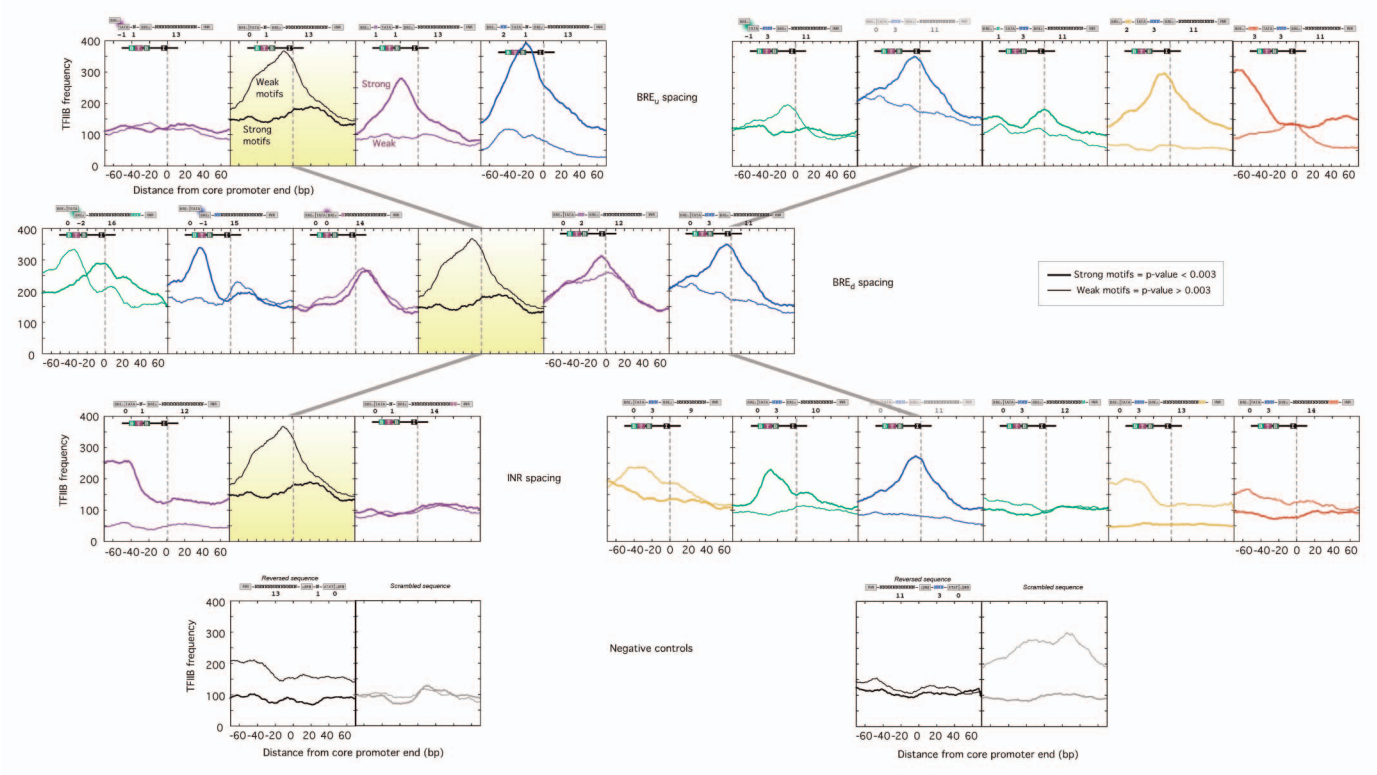
Extended Data Figure 3 | Properties of CPEs associated with RefSeq genes. **a**, Average TFIIB and TBP occupancy parsed by the number of mismatches to the TATA consensus. **b**, Distribution of each candidate CPE relative to each other.



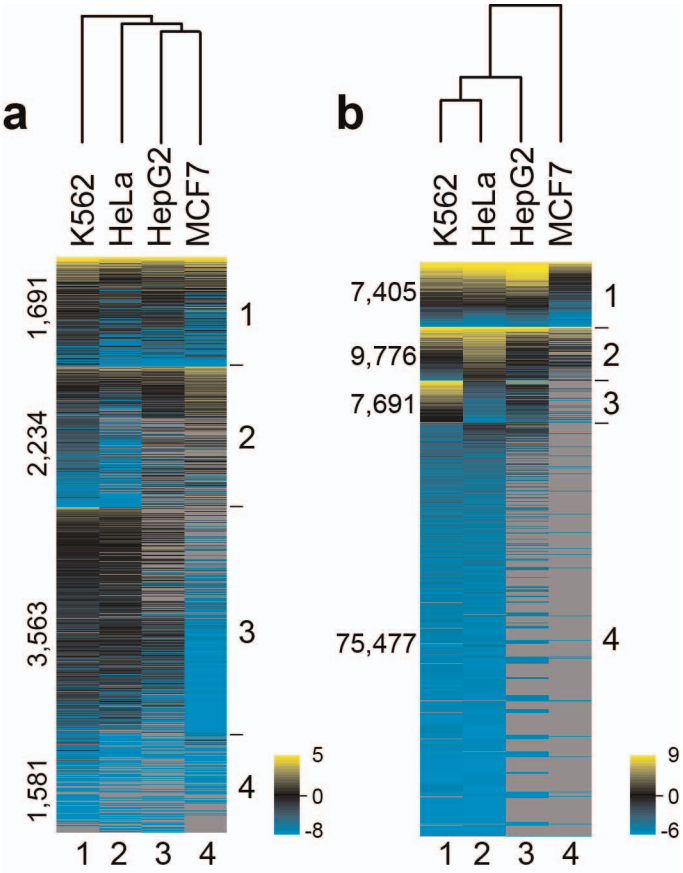
Extended Data Figure 4 | CPEs at non-coding loci bound by TFIIB. **a**, Bar graph showing the percentage of all 150,754 putative 'non-coding' TFIIB binding locations (>500 bp from an annotated RefSeq TSS) that have the indicated number of CPEs. **b**, Distribution of ChIP-exo peaks on each strand relative to the indicated CPE, for 150,754 putative non-coding TFIIB locations. Opposite strand traces (red) are inverted. **c**, Distribution of TBP (purple) and Pol II (black) peak-pair midpoints relative to the TATA motif midpoint derived from the 150,754 TFIIB putative non-coding locations. **d**, TFIIB occupancy versus percentage of locations that code for proteins. All 159,117 TFIIB locations were sorted by occupancy level, and the percentage of locations linked to an annotated RefSeq feature was plotted as a moving average.



Extended Data Figure 5 | Enrichment of different RNA fractions at 159,117 TFIIB locations throughout the human genome. Frequency distribution RNA 5′ ends for poly(A)$^+$ (ref. 38) (top) and ENCODE project RNA fractions[40] as indicated to the far left. Traces in the left panels are separated by sense (blue) and antisense (red, inverted) orientations relative to the corresponding mRNA TSS, which is directed to the right. Because the TSS orientation is not known for the poly(A)$^-$ ncRNA loci, positive and negative strand tags were plotted relative to the TFIIB midpoint. The percentage of putative TFIIB locations that exist within 2 kb of an RNA tag are indicated in the top right corner of each plot.



Extended Data Figure 6 | TFIIB core promoter distances. Candidate CP at varying distances from all 159,117 TFIIB locations, for the indicated spacing variants (not all possible combinations were tested). Digits within spacing variant schematic reflect the base-pair spacing (N) between elements. CPE with high *P* values (less correlated to the PSPM matrix) have thin lines, whereas low/strong *P* values ($<3 \times 10^{-4}$) have thick lines.



Extended Data Figure 7 | Promoter complexes across cancer cell lines. **a**, **b**, Occupancy levels for TFIIB linked to coding genes (**a**) and non-coding regions (**b**) in the indicated cell type were normalized by column. The colour scales represent the range of average-centred, $\log_2$-transformed values within each respective column. Detection in all four cell types defines group 1. Groups 2–4 were parsed by *k*-means clustering. Rows were sorted within groups based on TFIIB occupancy averaged across the four cell types (yellow, black, cyan and grey denote high, medium, low and zero occupancy, respectively). For clarity in **b**, TFIIB locations that were detected in only one cell line were excluded from clustering. Columns were hierarchically clustered. The MCF7 data set had 20–30% of the coverage of other cell lines (reported in Supplementary Data 3), which probably accounts for an excessive number of zero-occupancy loci (grey).

Extended Data Table 1 | Statistics of Illumina sequencing

Factor	Antibody	Cell Line	Total Reads	Uniquely Mapped Reads	Unique Mapping Rate
Input	none	K562	126,007,656	104,591,819	83%
Input	none	K562	109,745,112	91,160,835	83%
		Totals:	235,752,768	195,752,654	
TBP	sc-204	K562	97,896,951	60,581,579	62%
TBP	sc-204	K562	181,420,753	132,655,896	73%
TBP	sc-204	K562	200,167,837	115,213,419	58%
		Totals:	479,485,541	308,450,894	
TFIIB	sc-225	K562	64,473,390	43,727,825	68%
TFIIB	sc-225	K562	129,513,614	80,930,721	62%
		Totals:	193,987,004	124,658,546	
Pol II	sc-899	K562	40,833,504	31,260,456	77%
Pol II	sc-899	K562	119,799,682	88,431,598	74%
		Totals:	160,633,186	119,692,054	
TFIIB	sc-225	HeLa-S3	62,249,055	41,815,431	67%
TFIIB	sc-225	HeLa-S3	185,240,056	123,002,393	66%
		Totals:	247,489,111	164,817,824	
TFIIB	sc-225	HepG2	78,313,847	50,505,201	64%
TFIIB	sc-225	HepG2	264,530,278	172,112,282	65%
		Totals:	342,844,125	222,617,483	
TFIIB	sc-225	MCF7	25,615,261	14,780,271	58%
TFIIB	sc-225	MCF7	120,958,757	28,600,410	24%
		Totals:	146,574,018	43,380,681	

Summary of uniquely mapped sequencing reads for each biological replicate.

doi:10.1038/nature12593

Single-cell Hi-C reveals cell-to-cell variability in chromosome structure

Takashi Nagano[1]*, Yaniv Lubling[2]*, Tim J. Stevens[3]*, Stefan Schoenfelder[1], Eitan Yaffe[2], Wendy Dean[4], Ernest D. Laue[3], Amos Tanay[2] & Peter Fraser[1]

Large-scale chromosome structure and spatial nuclear arrangement have been linked to control of gene expression and DNA replication and repair. Genomic techniques based on chromosome conformation capture (3C) assess contacts for millions of loci simultaneously, but do so by averaging chromosome conformations from millions of nuclei. Here we introduce single-cell Hi-C, combined with genome-wide statistical analysis and structural modelling of single-copy X chromosomes, to show that individual chromosomes maintain domain organization at the megabase scale, but show variable cell-to-cell chromosome structures at larger scales. Despite this structural stochasticity, localization of active gene domains to boundaries of chromosome territories is a hallmark of chromosomal conformation. Single-cell Hi-C data bridge current gaps between genomics and microscopy studies of chromosomes, demonstrating how modular organization underlies dynamic chromosome structure, and how this structure is probabilistically linked with genome activity patterns.

Chromosome conformation capture[1] (3C) and derivative methods (4C, 5C and Hi-C)[2–6] have enabled the detection of chromosome organization in the three-dimensional space of the nucleus. These methods assess millions of cells and are increasingly used to calculate conformations of a range of genomic regions, from individual loci to whole genomes[3,7–11]. However, fluorescence *in situ* hybridization (FISH) analyses show that genotypically and phenotypically identical cells have non-random, but highly variable genome and chromosome conformations[4,12,13], probably owing to the dynamic and stochastic nature of chromosomal structures[14–16]. Therefore, although 3C-based analyses can be used to estimate an average conformation, it cannot be assumed to represent one simple and recurrent chromosomal structure. To move from probabilistic chromosome conformations averaged from millions of cells towards determination of chromosome and genome structure in individual cells, we developed single-cell Hi-C, which has the power to detect thousands of simultaneous chromatin contacts in a single cell.

Single-cell Hi-C

We modified the conventional or 'ensemble' Hi-C protocol[3] to create a method to determine the contacts in an individual nucleus (Fig. 1a and Supplementary Information). We used male mouse spleenic $CD4^+$ T cells, differentiated *in vitro* to T helper (T_H1) cells to produce a population of cells (>95% $CD4^+$), of which 69% have 2*n* genome content, reflecting mature cell withdrawal from the cell cycle. Chromatin crosslinking, restriction enzyme (BglII or DpnII) digestion, biotin fill-in and ligation were performed in nuclei (Fig. 1a and Extended Data Fig. 1a) as opposed to ensemble Hi-C, in which ligation is performed after nuclear lysis and dilution of chromatin complexes[3]. We then selected individual nuclei under the microscope, placed them in individual tubes, reversed crosslinks, and purified biotinylated Hi-C ligation junctions on streptavidin-coated beads. The captured ligation products were then digested with a second restriction enzyme (AluI) to fragment the DNA, and ligated to customized Illumina adapters with unique 3-bp (base pair) identification tags. Single-cell Hi-C libraries were then PCR amplified, size selected and characterized by multiplexed, paired-end sequencing.

De-multiplexed single-cell Hi-C libraries were next filtered thoroughly to systematically remove several sources of noise (Extended Data Fig. 1b–f and Supplementary Information). Hi-C in male diploid cells can theoretically give rise to at most two ligation products per autosomal restriction fragment end, and one product per fragment end from the single X chromosome. Using BglII, the total number of distinct mappable fragment-end pairs per single cell cannot therefore exceed 1,201,870 (Extended Data Fig. 1g and Supplementary Information). In practice, deep sequencing of the single-cell Hi-C libraries demonstrated that following stringent filtering our current scheme allows recovery of up to 2.5% of this theoretical potential, and has identified at least 1,000 distinct Hi-C pairings in half (37 out of 74) of the cells. Deep sequencing confirmed saturation of the libraries' complexity, and allowed elimination of spurious flow cell read pairings and additional biases (Extended Data Tables 1–3). On the basis of additional quality metrics we selected ten single-cell data sets, containing 11,159–30,671 distinct fragment-end pairs for subsequent in-depth analysis (Extended Data Fig. 1h–l). Visualization of the single-cell maps suggested that despite their inherent sparseness, they clearly reflect hallmarks of chromosomal organization, including frequent *cis*-contacts along the matrix diagonal and notably, highly clustered *trans*-chromosomal contacts between specific chromosomes (Fig. 1b).

Single-cell and ensemble Hi-C similarity

We used the same population of $CD4^+$ T_H1 cells to generate an ensemble Hi-C library. Sequencing and analysis[17] of 190 million read pairs produced a contact map representing the mean contact enrichments within approximately 10 million nuclei. The probability of observing a contact between two chromosomal elements decays with linear distance following a power law regime for distances larger than 100 kilobases (kb)[3,18]. We found similar regimes for the ensemble, individual

[1]Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK. [2]Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel. [3]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK. [4]Epigenetics Programme, The Babraham Institute, Cambridge CB22 3AT, UK.
*These authors contributed equally to this work.

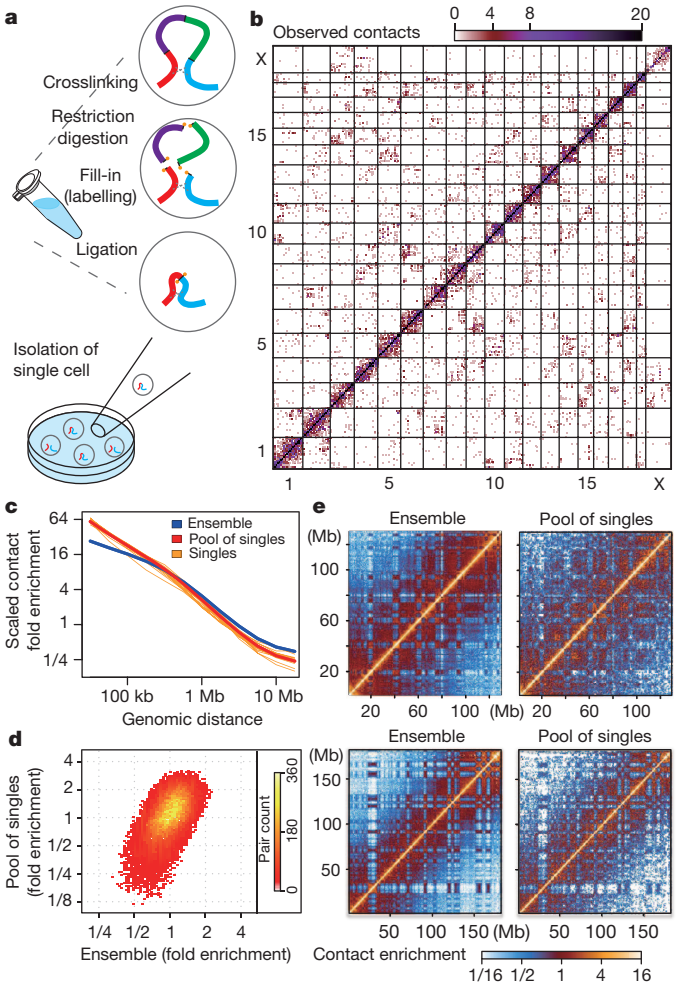


Figure 1 | Single-cell and ensemble Hi-C. **a**, Single-cell Hi-C method. **b**, Single-cell Hi-C heatmap (cell 5), coverage for 10-Mb bins. **c**, Contact enrichment versus genomic distance, from ensemble Hi-C, pool of 60 single cells and 10 individual cells, scaled to normalize sequencing depths. **d**, Normalizing by the trends in **c**, intra-chromosomal contact enrichments for 1-Mb square bins, comparing ensemble and pooled single-cell Hi-C (Spearman correlation = 0.56). **e**, Intra-chromosomal contact enrichment maps of ensemble and pooled single-cell Hi-C, for chromosome 10 (top) and chromosome 2 (bottom), using variable bin sizes.

cells and a pool of 60 single cells (Fig. 1c). Moreover, after normalizing the matrices given this canonical trend, comparison of intra-chromosomal interaction intensities for the pool and ensemble, by global correlation analysis of contact enrichment values at 1-megabase (Mb) resolution generates a highly significant correspondence (Fig. 1d). This is emphasized by the high similarity observed in comparisons of individual chromosomes from ensemble and pooled Hi-C maps (Fig. 1e). In summary, despite different experimental procedures and sparse nature of the single-cell matrices, the pooled matrix retains the most prominent properties of the ensemble map, confirming the validity of the approach and prompting us to explore further the similarities and differences among the individual cell chromosomal conformations.

Intra- and interdomain contacts

A key architectural feature of ensemble Hi-C data sets is their topological domain structure[18–20]. As expected, 1,403 domains were identified in the T_H1 cell ensemble Hi-C map[18] (Supplementary Table 1 and Supplementary Information). We used the ensemble domains to ask whether the same domain structure can be observed at the single-cell level. Visual inspection of the domain structure overlaid on individual intra-chromosomal contact maps (Fig. 2a), and global statistical analysis of the ratios between intra- and interdomain contact intensities in individual cells (approximately twofold enrichment on scales of 100 Kb to 1 Mb; Fig. 2b and Extended Data Fig. 2a), both supported the idea that domains are observed consistently in the single-cell maps. To test whether domain structures are variable between individual cells, we estimated the distributions of intradomain contact enrichments across cells and compared it to the distributions derived from reshuffled maps. We reasoned that cell-to-cell variation in intradomain contact intensities would result in an increase of the variance of this distribution compared to the expected variance resulting from sampling contacts in uniformly (shuffled) intradomain contacts. However, the data (Fig. 2c) showed that the distributions for the intradomain enrichments in real cells are not more varied than expected (Kolmogorov–Smirnov $P < 0.52$). A similar observation was derived by comparison of the correlations between intradomain contact enrichments for pairs of real and pairs of reshuffled maps (Extended Data Fig. 2b). Although this analysis cannot quantify variability in the high-resolution internal structure of domains, the data suggest that domain intactness is generally conserved at the single-cell level.

Visual comparison of whole-chromosome contact maps (Fig. 2d) suggested that unlike intradomain interactions, interdomain contacts within single-cell chromosomes are structured non-uniformly. The maps showed large-scale structures as indicated, for example, by specific insulation points separating chromosomes into two or more mega domains in a cell-specific fashion. To rule out the possibility that this can be explained by sparse sampling of contacts in each single-cell map we again used reshuffled controls. In each map (real or randomized) we quantified the frequency of loci that strongly polarize the matrix into two weakly connected submatrices (using an insulation score; Supplementary Methods). We confirmed that single-cell maps indeed show many more such loci than reshuffled maps (Fig. 2e and Extended Data Fig. 2c). The reshuffled controls made by mixing contacts from different single-cell maps, are in fact similar to sparse versions of the ensemble map, which do not show specific structure at the intradomain level. Along similar lines, the correlation in contact intensities between domains on the same chromosome in pairs of single-cell maps is lower compared to reshuffled controls (Fig. 2f). Taken together, these data show that domains form a robust and recurrent conformational basis that is evident in each of the single cells. However, interdomain contacts are highly variable between individual cells, suggesting large-scale differences in higher-order chromosome folding that are obscured in ensemble maps, averaged over millions of such structures[21].

Three-dimensional modelling of X chromosomes

To determine whether the single-cell Hi-C data are consistent with unique chromosome conformations we developed a modelling approach to reconstruct the conformations of the single-copy, male X chromosome. We used intra-chromosomal contacts as distance restraints and calculated structural models using a simulated annealing protocol to condense a particle-on-a-string representation of individual chromosomes from random initial conformations (Supplementary Information), to produce both fine-scale and low-resolution models, with backbone particles representing either 50 or 500 kb of the chromosome, respectively. For fine-scale calculations, each intra-chromosomal contact restrained its precise position on the chromosome, whereas low-resolution calculations combined contacts into larger bins. Tests of our simulation protocol demonstrated that restraint density was the most important parameter for modelling (Extended Data Fig. 3a, b). Hence, from the ten high-quality single-cell data sets, we selected six with the largest numbers of intra-chromosomal X contacts, plus one with a lower number of contacts (cell 9) for contrast.

Repeat calculations starting from random positions generated 200 X-chromosome models for each cell at both scales. The fine-scale models displayed very low numbers of restraint violations (Extended Data Fig. 3c). We introduced an estimated average unit DNA distance

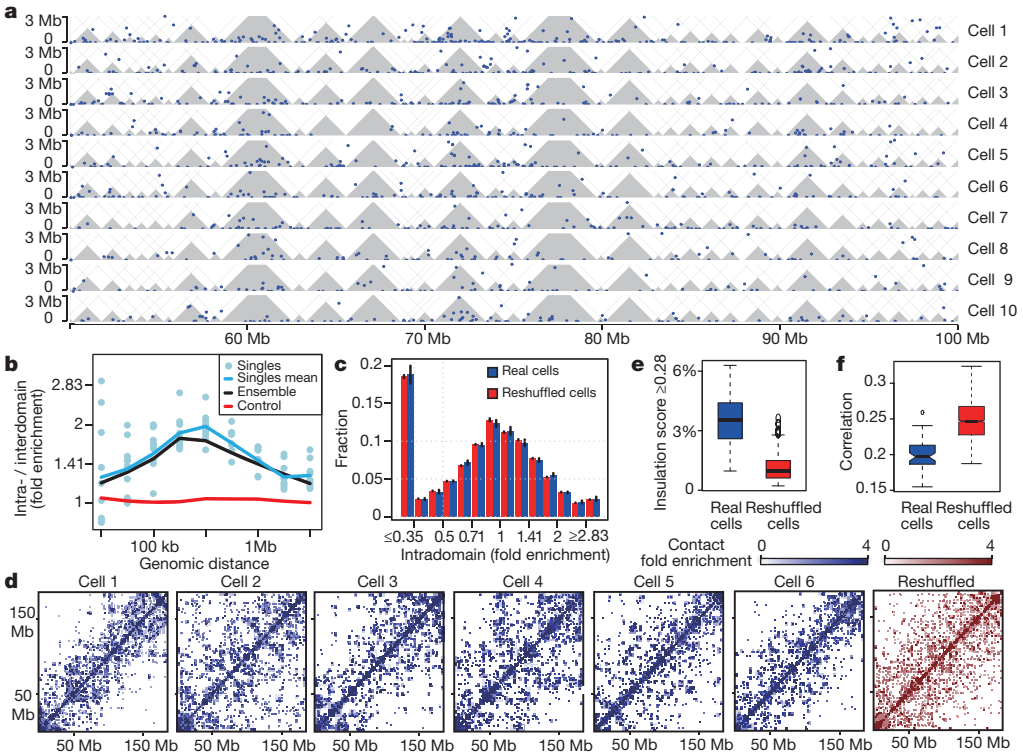


Figure 2 | Conserved intradomain, but not interdomain structure in single cells. **a**, Intra-chromosomal contact maps of 50-Mb region of chromosome 2 from ten single cells. Individual contacts up to 3-Mb distance are shown as blue dots. Domains inferred from the ensemble Hi-C maps in grey. **b**, Ratios between intradomain and interdomain contact enrichments over genomic distance. Control is combined trend of 10 single cells calculated by repeatedly shifting the domains randomly. **c**, Distribution of intradomain contact enrichments per domain from 9 cells (where BglII was used) and reshuffled data sets (black bars, s.e.m.). **d**, Maps of interdomain contact intensities for chromosome 2 from individual cells and reshuffled controls using variable bin sizes. **e**, Distribution of percentage of loci with high insulation scores in single versus reshuffled cells. **f**, For all pairs of single cells, the correlations between interdomain contact numbers of all pairs of domains within the same chromosome were computed. Shown are the distributions of these correlations in the real and reshuffled cells.

length[22] to approximate packaging of chromatin fibres (~0.15 μm per 50 kb) (Supplementary Information). This resulted in models with a mean X-chromosome territory diameter of 4.3 μm (range 3.3–5.9 μm), in good agreement with X-chromosome paint FISH in T_H1 cells (Fig. 3a; mean diameter 3.7 μm) and chromosome territory sizes in live cells[23]. We confirmed that the restrained points in a single cell are indeed close in the structures calculated from them (Extended Data Fig. 3c, d). Interestingly, the single-cell distance matrix demonstrates how the network of contacts in a model imparts further structural information beyond the directly observed contacts (Extended Data Fig. 3d).

Comparison of the low-resolution models demonstrated convergence towards a single conformation for each single-cell data set (Fig. 3b and Extended Data Fig. 3e). For fine-scale models, hierarchical clustering revealed four or five that were most representative of the data (Fig. 3c). In all cases models from a single cell were significantly more similar to each other than to models from different cells (Extended Data Fig. 4a, b).

Highlighting four regions of the X chromosome showed large-scale conformational differences between cells (Fig. 3d), supporting the finding of highly variable interdomain contacts. Models created by shuffling Hi-C contacts, or combining contacts from two cells resulted in structures smaller and more compact than observed chromosome territories (Extended Data Fig. 4c, d) with many restraints stretched towards or exceeding their upper bounds (Extended Data Fig. 3c). These results reaffirm that the variation in single-cell contacts is not the result of partial sampling of a single underlying structure.

We next asked whether despite their cell-to-cell variability, X-chromosome structures share common folding properties that could be tested in real cells. One such important property, which is often consistent within a cell population, and with multiple potential functional implications, is localization within the chromosomal territory relative to its surface. To predict loci with consistent positions within their chromosome territory we calculated the structural density along the X chromosome (Supplementary Information) and identified regions with consistently high or low structural density (Fig. 3e). We chose five such regions (P1–P5) with predicted positions near the surface (P1, P2, P5; low structural density) or inside (P3, P4; high structural density) the model X-chromosome territories using the 1,200 models from the six cells (Extended Data Fig. 4e). We then performed double label DNA FISH with X-chromosome paints and P1–P5 bacterial artificial chromosome (BAC) probes (Fig. 3f) to test directly these predictions. The distances between DNA FISH signals and edge of the chromosome territory in over one hundred T_H1 cells showed that probes P1, P2 and P5 were indeed found predominantly outside or towards the edge of the chromosome territory, whereas signals for probes P3 and P4 were found at internal positions (Fig. 3g). These data show that despite highly variable interdomain structure of the X-chromosomal territory, some of its key organizational properties are robustly observed across the cell population.

Domains at the interface

Data from *trans*-chromosomal contacts were overlaid on the X-chromosome models, and this showed that *trans*-chromosomal contacting regions are strongly enriched towards the inferred surface of the models (Fig. 3h), providing further validation. These observations prompted us to explore further the structural characteristics of interfaces between chromosomal territories, and the relationships between such interfaces and the domain structure of the territory itself. We found that *trans*-chromosomal contact enrichments of domains vary across cells (Fig. 4a), showing a significant difference between the mean contact enrichment per domain in the real and reshuffled maps ($P < 1.2 \times 10^{-9}$, Kolmogorov–Smirnov test). The higher variance of the distribution for the real data suggests that some domains are more likely to contact elements on other chromosomes. Previous work has suggested that active genomic regions on the sub-domain scale often loop out of their chromosome territories[24], which may imply less defined local domain structures and disassociation from their chromosome territory. However, our analysis shows that *trans*-contacting domains retain domain organization, as demonstrated by the intradomain contact probabilities within them (Fig. 4b and Extended Data Fig. 5a, b). Conversely, *trans*-contacting domains show slightly reduced

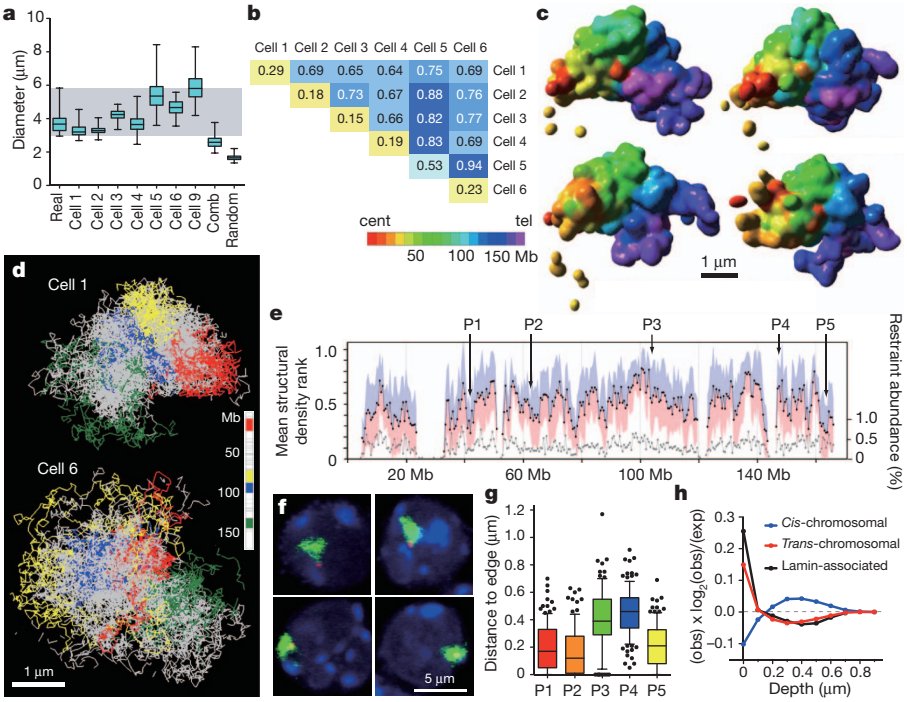


Figure 3 | Structural modelling of X chromosomes. **a**, Distribution of longest diameter of X-chromosome paint DNA FISH signals in 62 male T_H1 cells (real), 200 structural models calculated for each single cell (cell 1 to cell 9), 200 structures from combined data set (cell 1 and cell 2; comb) and 200 structures from 20 randomized cell 1 data sets (random; 10 calculations per data set). Whiskers denote minimum and maximum. **b**, Average coordinate root-mean-square deviation (r.m.s.d.) values in microns comparing 200 low-resolution structural models for each cell and between cells. **c**, Four surface-rendered models of the X chromosome from cell 1, which are most representative of the data based on hierarchical clustering of pair-wise r.m.s.d. values (Supplementary Information). Scale bar, 1 μm. **d**, Structural ensembles of the four most representative fine-scale models for cell 1 and cell 6, with four large regions coloured. Scale bar, 1 μm. **e**, Mean structural density rank for 500-kb regions (black) from 6 × 200 fine-scale models from cell 1 to cell 6. Standard deviation is shown in blue and pink. Abundance of intra-chromosomal restraints is shown in grey, right axis. DNA FISH probes (P1–P5) are indicated. **f**, DNA FISH on T_H1 cells. X-chromosome paint (green) and specific locus signals (red). **g**, Distribution of DNA FISH distance measurements between signal centres for probes P1–P5 and edge of the X-chromosome territory in T_H1 cells (n = 114, 113, 105, 115 and 108 for P1–P5, respectively). Whiskers denote 10 and 90 percentiles. **h**, Enrichment of *cis*-, *trans*-contacts and lamin-B1-associated domains at various depths on the chromosome models relative to null hypothesis of random positions.

contact intensity to other domains on the same chromosome (Fig. 4c and Extended Data Fig. 5c, d), consistent with localization on the interfaces of their territories rather than dissociation from them.

Analyses of ensemble Hi-C data have previously shown that active marks correlate with enrichment of *trans*-chromosomal contacts[3,17]. Using the single-cell maps combined with annotation of domains based on their enrichment for histone H3 lysine 4 tri-methylation (H3K4me3) hotspots[25] (Fig. 4d), we tested whether this correlation is the result of low frequency re-localization of active domains to other chromosome territories (looping out), or from frequent localization of active domains on territory interfaces. As shown in Fig. 4e, domains with high *trans*- to *cis*-chromosomal contact ratios (excluding intradomain) are highly correlated with H3K4me3 enrichment in all cells. However, the data show that domains (including active ones) retain their association with the territory in almost all cases. Very few domains with strong *trans*-contacts were found to lack association with their own territory (Fig. 4e; upper left points in graphs). Some of this lack of perfect territory re-localization can be explained by having two copies of each autosomal domain, but the overall reduction with territory association for *trans*-contacting domains is much smaller than the 50% expected by this explanation (reduction estimated at 15–20% and 10% for contacts across 1–5 Mb and 10 Mb, respectively, Fig. 4c). Comparison of active domain localization shows that different active domains are highly *trans*-contacting in each cell (Extended Data Fig. 5e). Together, these data show that preferential localization of active domains to territory interfaces is a hallmark of chromosome organization in all cells. Active domains maintain their intradomain organization, and only partially lose intra-chromosomal contacts with other domains. Our data are consistent with the concept that chromosomal territories are maintained robustly despite the *trans*-chromosomal contacts between active domains.

Interestingly, domains associated with lamin B1 (ref. 26), which are thought to be primarily inactive regions, are also found towards the surface of the models (Fig. 3h). However, these domains are highly anti-correlated with H3K4me3 domains (Spearman's correlation = −0.73) and typically depleted of *trans*-chromosomal contacts (Extended Data Fig. 5f–i). Superposition of H3K4me3, lamin-B1-enriched domains and *trans*-chromosomal contacts on the X-chromosome models illustrates spatial partitioning of the active, *trans*-contacting regions from those that are lamin-associated, although both types of domains tend towards the surface of the chromosome territory, supporting the above descriptions of differential positioning of domains (Extended Data Fig. 5j and Supplementary Videos 1 and 2).

Ensemble Hi-C maps generate a highly complex view of chromosomal contacts, including low-intensity contacts between all possible chromosomal pairings[3,8,17,19]. In contrast, years of single-cell analyses by microscopy have suggested that individual cells have much simpler and discrete chromosome structures involving a limited number of interfaces between spatially constrained chromosomal territories[27,28]. Our single-cell maps bridge the gap between the genomic and imaging techniques, showing cell-specific clusters of *trans*-chromosomal contacts associating some pairs of chromosomes, and a lack of contacts between other chromosome pairs (Fig. 5a, blue). Such organization is completely lacking in reshuffled maps (Fig. 5a, red) confirming it is not a consequence of sparse contact sampling (Extended Data Fig. 6a, b). *Trans*-chromosomal contact clusters bring pairs of domains together, as shown by comparing the enrichment in *trans*-contacts between pairs of elements connecting the same two domains and pairs connecting

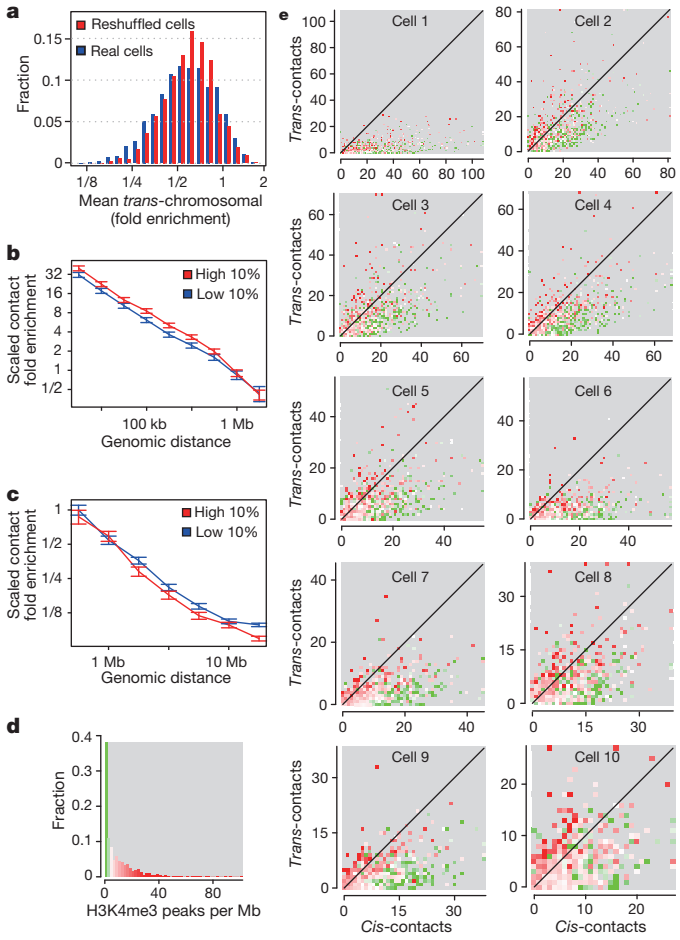


Figure 4 | Active domains localize to territory interfaces. **a**, Distribution of *trans*-chromosomal contact enrichments of each domain averaged across real and reshuffled cells. Reshuffling maintains the number of *cis*- and *trans*-contacts within each cell and chromosome. **b**, Intradomain contact enrichment over genomic distance for high versus low *trans*-chromosomal contacting domains selected independently in each cell, with 95% confidence intervals. **c**, Same sets as in **b** but plotting the enrichment of interdomain contacts. **d**, Distribution of H3K4me3 peak density in domains (number of peaks divided by size), colour-coded according to density. **e**, Domains plotted according to number of *trans*- and *cis*-chromosomal (excluding intradomain) contacts, colour coded for H3K4me3 density as in **d**.

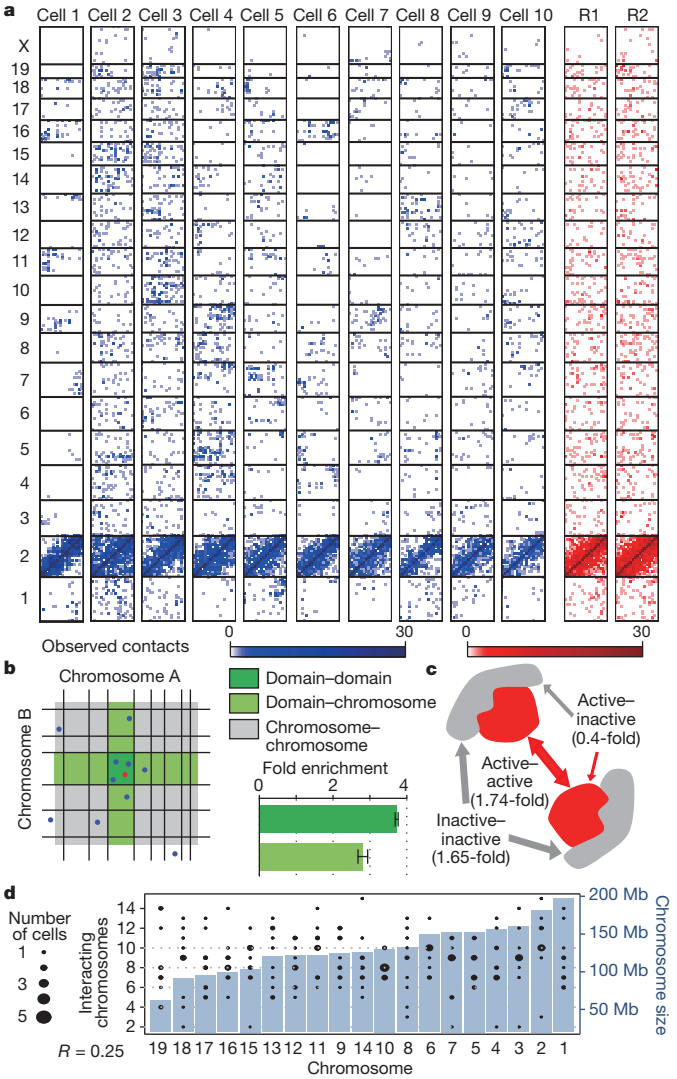


Figure 5 | Chromosomal interfaces. **a**, All *trans*-chromosomal contacts formed by chromosome 2 in real cells (blue) and reshuffled cells (red). **b**, Schematic diagram of a chromosomal interface between linearly adjacent domains, their borders marked in black on two chromosomes, A and B. We considered each of the two contacting fragments of every *trans*-chromosomal contact and classified every nearby *trans*-chromosomal contact as domain–domain, domain–chromosome and chromosome–chromosome, the latter being used as background for normalization (Supplementary Information). The contact under consideration is shown in red, and nearby contacts are shown in blue. Fold enrichments shown for each group type (error bars, standard deviation). **c**, *Trans*-chromosomal contacts are highly significantly enriched between active domains (H3K4me3 enriched) or between inactive domains, but not mixed interaction (chi-squared test; $P = 5.8 \times 10^{-18}$; even after taking account of the generally higher connectivity of active domains). **d**, Bar graph depicting mouse autosomes ordered by size with number of interacting chromosomes per single cell (black circles depict the distribution over individual cells). Mean number of interacting chromosomes changes modestly (30%) with chromosome size, suggesting a highly organized territory structure with surface that is not scaling with chromosome length.

one domain with two different domains (Fig. 5b). Such synergistic contacting preferentially brings together pairs of active domains, with interaction between active and inactive domains being underrepresented (Fig. 5c and Extended Data Fig. 6c). Although inactive domains are depleted as a group from *trans*-chromosomal interactions (Fig. 4e), inactive domains that engage in *trans*-contacts are more likely to interact with other inactive domains. Interestingly, analysis of interacting pairs of domains suggests that the number of chromosomes contacting each chromosome is relatively constant (less than 30% difference) despite the greater than threefold change in chromosome size, the total number of *trans*-chromosomal contacts in the map, or a number of other factors (Fig. 5d and Extended Data Fig. 7a–e). We note that even though the total number of chromosome–chromosome interfaces per single cell is bounded, the detailed interface between chromosome pairs can involve multiple domain–domain contacts reflecting higher-order organization (Extended Data Fig. 7f).

Overall, these results indicate that each chromosome contacts a discrete and fairly constant number of other chromosomes in a single cell, with little dependency on the chromosome size. At the single-cell level both the microscopic and genomic observations therefore indicate highly defined territory structures, which may harbour much of the chromosome within the territory, and expose a limited, relatively constant surface area engaged in chromosome-to-chromosome interfaces. As these interfaces are highly variable among different cells, their averaging by ensemble Hi-C contributes towards the relatively uniform *trans*-chromosomal contact matrices previously reported.

We have presented a new experimental strategy to create Hi-C contact maps from single cells. The approach allows for characterization of thousands of simultaneous contacts occurring in individual

cells, and provides unique insights into Hi-C technology and three-dimensional chromosomal architecture (Supplementary Videos 3 and 4). Single-cell contact maps reflect conservation of domain structure that was recently characterized[18–20], but show that interdomain and *trans*-chromosomal contact structure is highly variable between individual cells. Genome-wide statistical analysis and reconstruction of the single-copy X-chromosome models gave us the opportunity to quantify key features of chromosomal architecture. For example, active domains tend to locate on the boundaries of their chromosomal territories in the majority of nuclei, while maintaining associations with other domains on the same chromosome. Our results do not exclude chromosome territory intermingling[29], but argue against domains becoming completely immersed in other territories. Coupled with previous observations of small and large-scale chromatin mobility[30–32] a highly dynamic view of chromosomal organization emerges, where territories are continuously being remodelled, while maintaining some key local (domain) and global (depth from surface) organizational features.

METHODS SUMMARY

T_H1 cells from male mice were fixed and subjected to modified Hi-C, in which nuclei were maintained through restriction-enzyme digestion, biotin fill-in labelling and ligation. Single nuclei were isolated and processed to prepare single-cell Hi-C libraries for paired-end sequencing.

Sequences were mapped to the mouse genome, and abnormal read pairs were discarded. Read pairs that occurred only once (without duplication) in the library sequencing were removed. We chose 10 single-cell data sets for further in-depth analyses based on several quality criteria (see Supplementary Information). To validate the single-cell Hi-C procedure, we pooled the single-cell Hi-C data sets and compared them to an ensemble Hi-C data set prepared from approximately 10 million cells essentially as described[3]. We created reshuffled data sets by randomly redistributing contacts of the analysed single cells to create the same number of cells with the same number of contacts in each cell as a control to analyse statistically the variation among single-cell data sets.

We reconstructed three-dimensional X-chromosome structure models using restrained molecular dynamics calculations employing a simulated annealing protocol. A combination of unambiguous distance restraints from the X intra-chromosomal contacts in the single-cell Hi-C data set and anti-distance restraints between regions that were found not to contact each other in the ensemble Hi-C data set was used. To assess the precision and accuracy of the structure generation process we used the protocol to generate synthetic Hilbert curve structures, and explored the impact of varying the number of restraints. For pair-wise comparison of the structures, we calculated the root-mean-square deviation (r.m.s.d.). To compare the X-chromosome models to X-chromosome structure *in vivo*, we selected five loci with consistently high or low structural density in the models, and compared distances between the loci and the X-chromosome territory surface in cells (DNA FISH).

Full description of the methods can be found in the Supplementary Information.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 December 2012; accepted 27 August 2013.

Published online 25 September 2013.

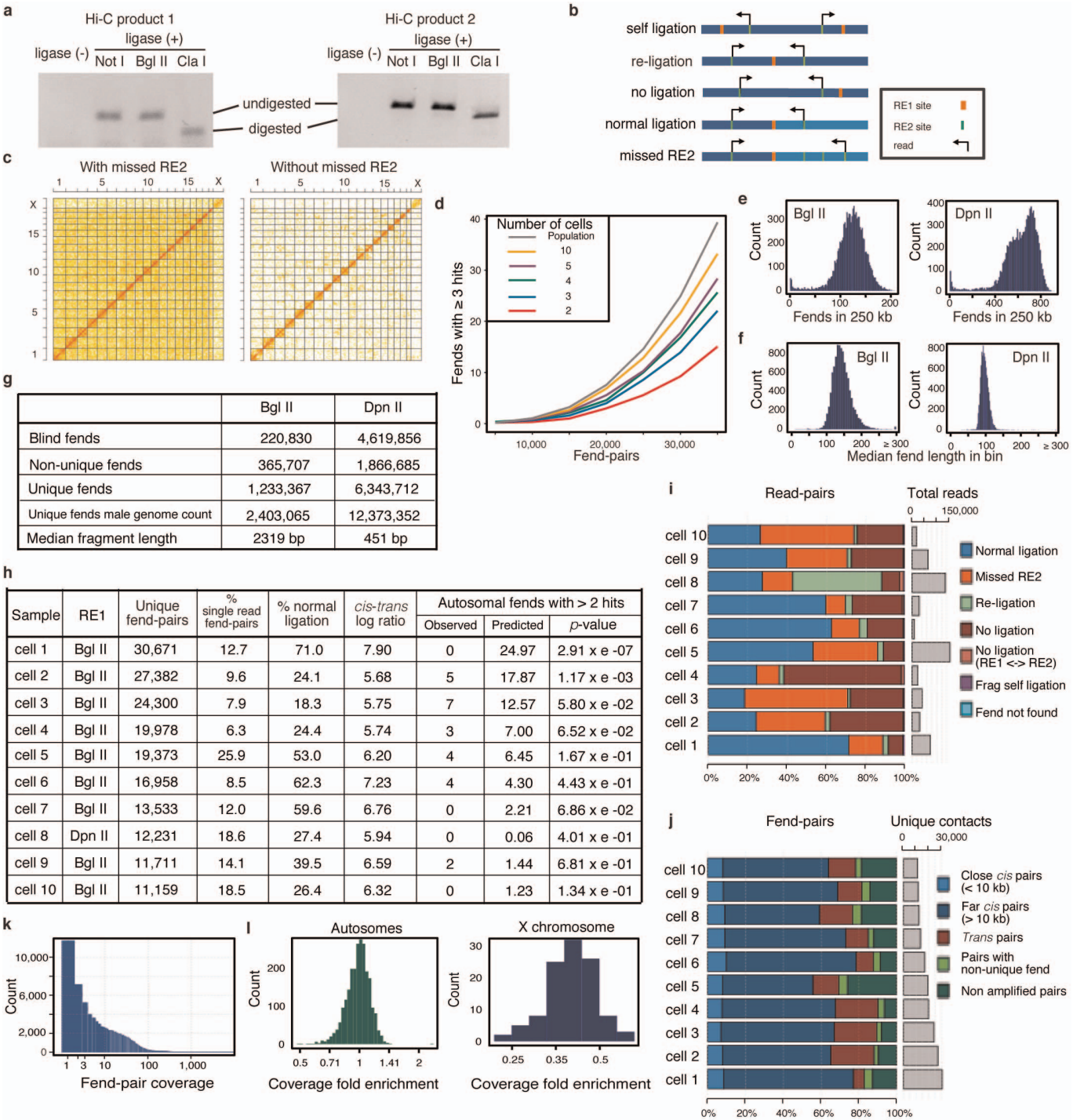
1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).
2. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
3. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
4. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genet.* **42,** 53–61 (2010).
5. 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).
6. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genet.* **38,** 1341–1347 (2006).
7. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465,** 363–367 (2010).
8. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnol.* **30,** 90–98 (2012).
9. Marti-Renom, M. A. & Mirny, L. A. Bridging the resolution gap in structural modeling of 3D genome organization. *PLOS Comput. Biol.* **7,** e1002125 (2011).
10. Tanizawa, H. *et al.* Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* **38,** 8164–8177 (2010).
11. van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9,** 969–972 (2012).
12. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genet.* **36,** 1065–1071 (2004).
13. Rapkin, L. M., Anchel, D. R., Li, R. & Bazett-Jones, D. P. A view of the chromatin landscape. *Micron* **43,** 150–158 (2012).
14. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447,** 413–417 (2007).
15. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Rev. Genet.* **8,** 104–115 (2007).
16. Osborne, C. S. *et al.* Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol.* **5,** e192 (2007).
17. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genet.* **43,** 1059–1065 (2011).
18. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148,** 458–472 (2012).
19. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
20. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385 (2012).
21. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49,** 773–782 (2013).
22. Jhunjhunwala, S. *et al.* The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133,** 265–279 (2008).
23. Müller, I., Boyle, S., Singer, R. H., Bickmore, W. A. & Chubb, J. R. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. *PLoS ONE* **5,** e11560 (2010).
24. Heard, E. & Bickmore, W. The ins and outs of gene regulation and chromosome territory organisation. *Curr. Opin. Cell Biol.* **19,** 311–316 (2007).
25. Deaton, A. M. *et al.* Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* **21,** 1074–1086 (2011).
26. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38,** 603–613 (2010).
27. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* **2,** 292–301 (2001).
28. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128,** 787–800 (2007).
29. Branco, M. R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4,** e138 (2006).
30. Chuang, C. H. *et al.* Long-range directional movement of an interphase chromosome site. *Curr. Biol.* **16,** 825–831 (2006).
31. Chubb, J. R., Boyle, S., Perry, P. & Bickmore, W. A. Chromatin motion is constrained by association with nuclear compartments in human cells. *Curr. Biol.* **12,** 439–445 (2002).
32. Dundr, M. *et al.* Actin-dependent intranuclear repositioning of an active gene locus *in vivo*. *J. Cell Biol.* **179,** 1095–1103 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank I. Clay, S. Wingett, K. Tabbada, D. Bolland, S. Walker, S. Andrews, M. Spivakov, N. Cope, L. Harewood and W. Boucher for assistance. This work was supported by the Medical Research Council, the Biotechnology and Biological Sciences Research Council (to P.F.), the MODHEP project, the Israel Science Foundation (to A.T.) and the Wellcome Trust (to E.D.L.).

Author Contributions T.N. and P.F. devised the single-cell Hi-C method. T.N. performed single-cell Hi-C and DNA FISH experiments. S.S. carried out ensemble Hi-C experiments. W.D. microscopically isolated single cells. Y.L., E.Y. and A.T. processed and statistically analysed the sequence data. T.J.S. and E.D.L. developed the approach to structural modelling and analysed X-chromosome structures. T.J.S. wrote the software for three-dimensional modelling, analysis and visualisation of chromosome structures. T.N., Y.L., T.J.S., E.D.L., A.T. and P.F. contributed to writing the manuscript, with inputs from all other authors.

Author Information Data deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE48262. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.F. (peter.fraser@babraham.ac.uk) for the single-cell Hi-C method, A.T. (amos.tanay@weizmann.ac.il) for the statistical analysis, or E.D.L. (e.d.laue@bioc.cam.ac.uk) for the structural modelling.



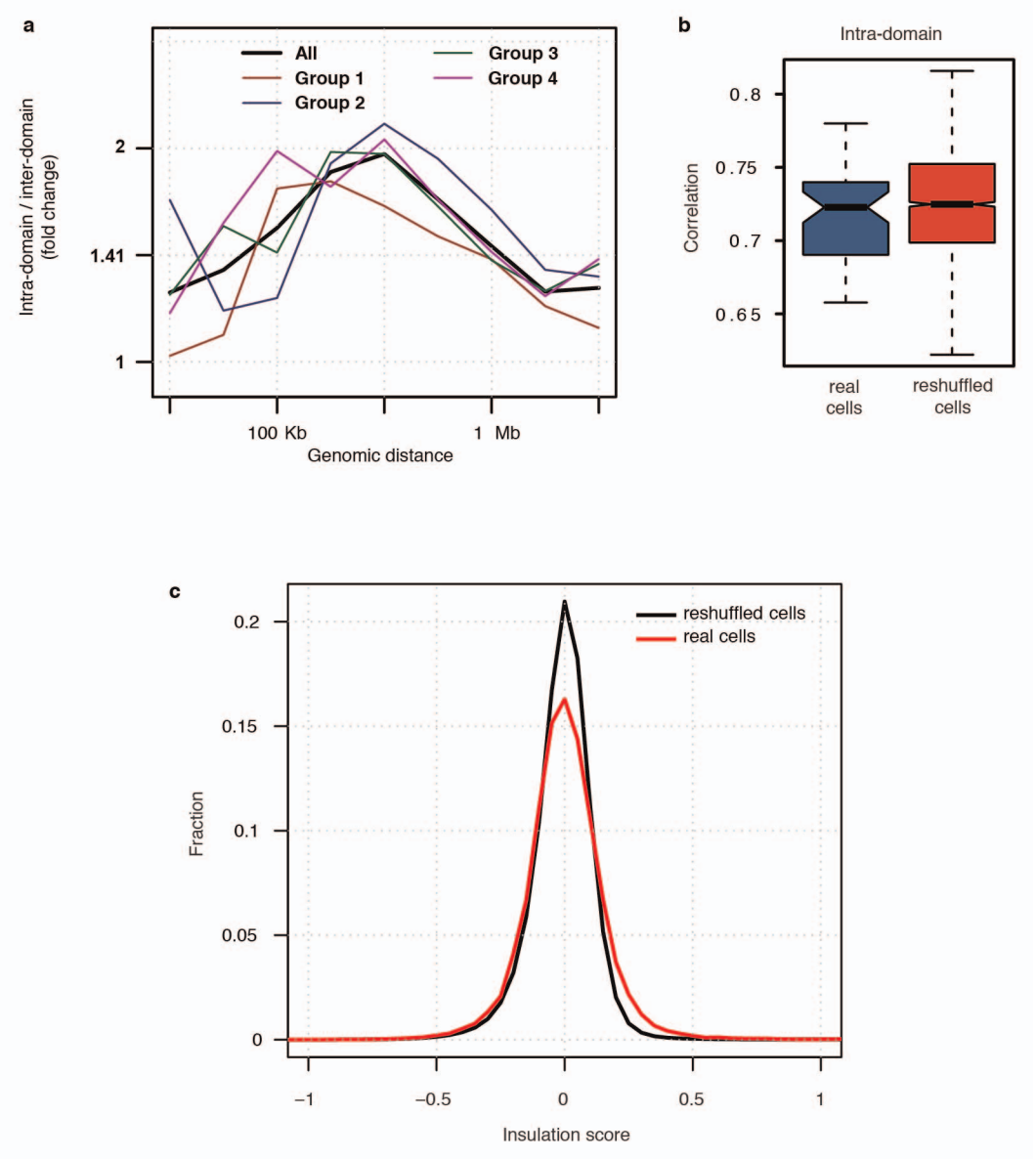
g

	Bgl II	Dpn II
Blind fends	220,830	4,619,856
Non-unique fends	365,707	1,866,685
Unique fends	1,233,367	6,343,712
Unique fends male genome count	2,403,065	12,373,352
Median fragment length	2319 bp	451 bp

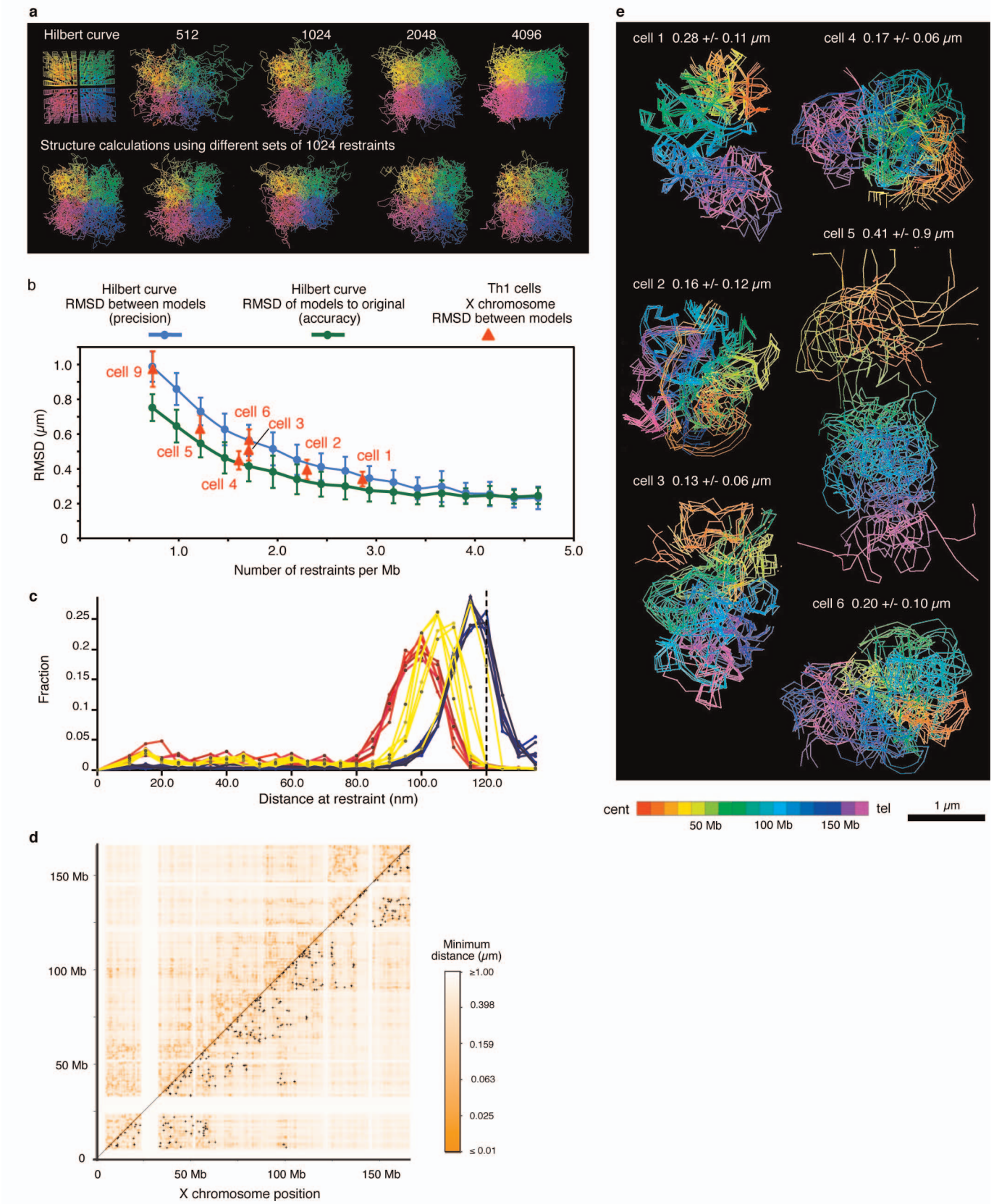
h

Sample	RE1	Unique fend-pairs	% single read fend-pairs	% normal ligation	*cis-trans* log ratio	Autosomal fends with > 2 hits		
						Observed	Predicted	*p*-value
cell 1	Bgl II	30,671	12.7	71.0	7.90	0	24.97	2.91 x e -07
cell 2	Bgl II	27,382	9.6	24.1	5.68	5	17.87	1.17 x e -03
cell 3	Bgl II	24,300	7.9	18.3	5.75	7	12.57	5.80 x e -02
cell 4	Bgl II	19,978	6.3	24.4	5.74	3	7.00	6.52 x e -02
cell 5	Bgl II	19,373	25.9	53.0	6.20	4	6.45	1.67 x e -01
cell 6	Bgl II	16,958	8.5	62.3	7.23	4	4.30	4.43 x e -01
cell 7	Bgl II	13,533	12.0	59.6	6.76	0	2.21	6.86 x e -02
cell 8	Dpn II	12,231	18.6	27.4	5.94	0	0.06	4.01 x e -01
cell 9	Bgl II	11,711	14.1	39.5	6.59	2	1.44	6.81 x e -01
cell 10	Bgl II	11,159	18.5	26.4	6.32	0	1.23	1.34 x e -01

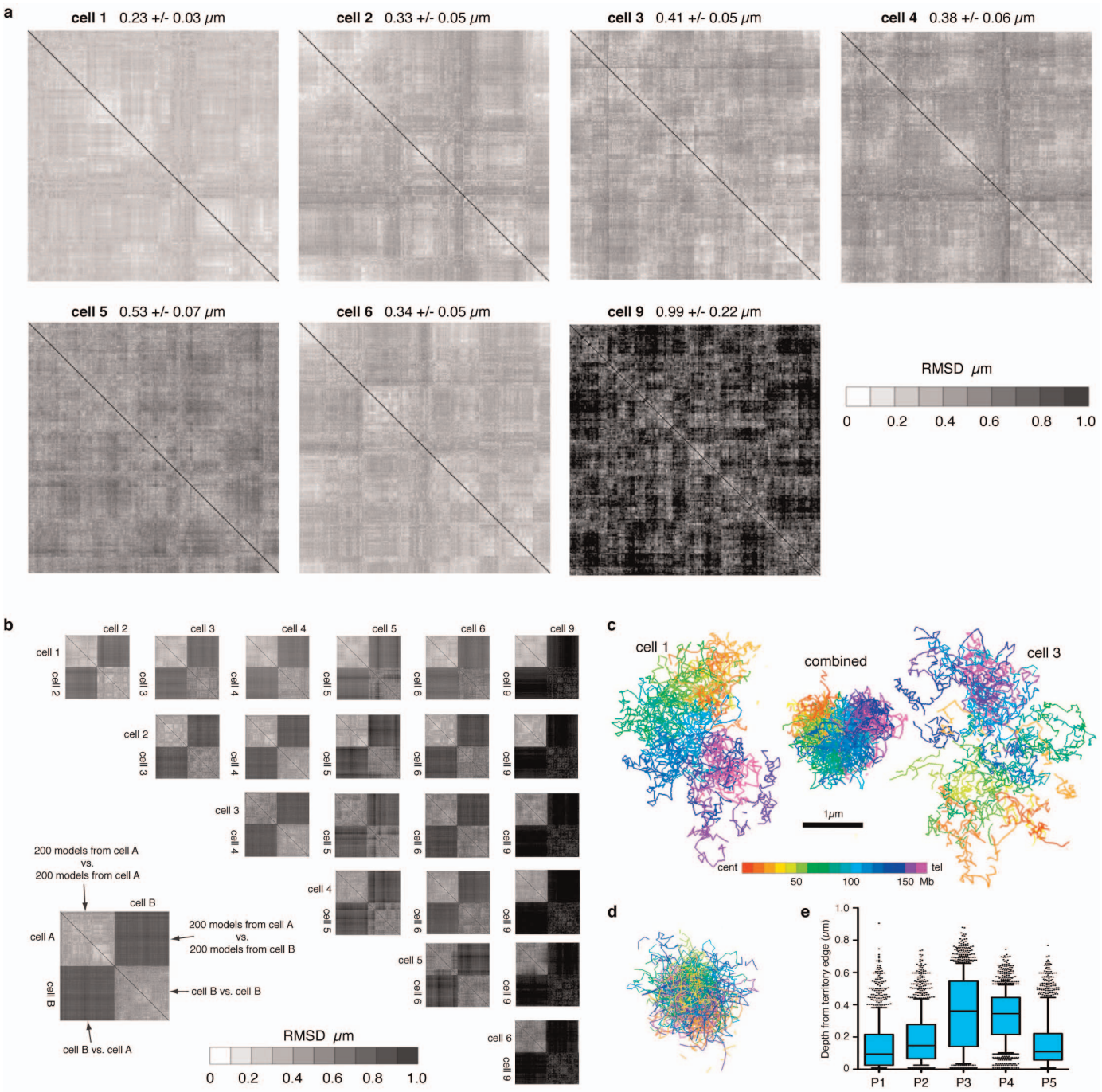
Extended Data Figure 1 | Single-cell Hi-C quality controls. **a**, Efficiency of biotin labelling at Hi-C ligation junctions for two Hi-C ligation products, showing 90–95% efficiency (Supplementary Information). **b**, Read-pair classification. **c**, Discarding the missed RE2 read-pairs removes a uniform 'blanket' of non-specific contacts from the map. **d**, Estimating numbers of multiple covered fends. Shown is the dependency between the number of fend pairs in a sample and the estimated number of autosomal fends covered by more than two fend pairs under different models. The binomial model (grey line) distributes fend pairs to fends randomly without any constraint, as if sampling fend pairs from an infinite number of chromosomes. **e**, Single-cell Hi-C fragments coverage. Number of fends in each 250-kb genomic bin for BglII or DpnII as RE1. Tail of bins with few fends is for bins of low mappability and near the chromosomes edges. **f**, Median fend length (distance from RE1 to the first upstream RE2) in each 250-kb genomic bin for BglII or Dpn II as RE1. Values larger than 300 bp are of poorly mappable bins. **g**, Information on the two restriction enzymes we used for RE1, BglII (6 cutter, which we used predominantly) and DpnII (4 cutter, only used for cell 8). Blind fends do not have a RE2 site in their fragment. Fends in which their first RE2 site starts a non-unique 36-bp sequence are marked as non-unique fends. We discarded both blind and non-unique fends and used only the unique fends. The number of actual fends in a male mouse genome, which have two copies of each autosome and a single X chromosome are shown as well as the median fragment length (chromosome Y and mitochondrial genome were ignored throughout the analysis). **h**, Information on the ten single-cell data sets that successfully passed the quality control filters. *P* value of the number of autosomal fends with more than two covering fend-pairs was calculated from the binomial model (panel **d** and Supplementary Information). **i**, Percentages of read-pair types. **j**, Percentage of fend-pair types. **k**, Distribution of fend-pair coverage (number of read pairs that support each fend pair) in the ten single-cell data sets. **l**, Distribution of mean contacts per fend calculated for each mappable 1 Mb, normalized by the mean value in each cell, and averaged across autosomal or X chromosomes from the ten single cells.



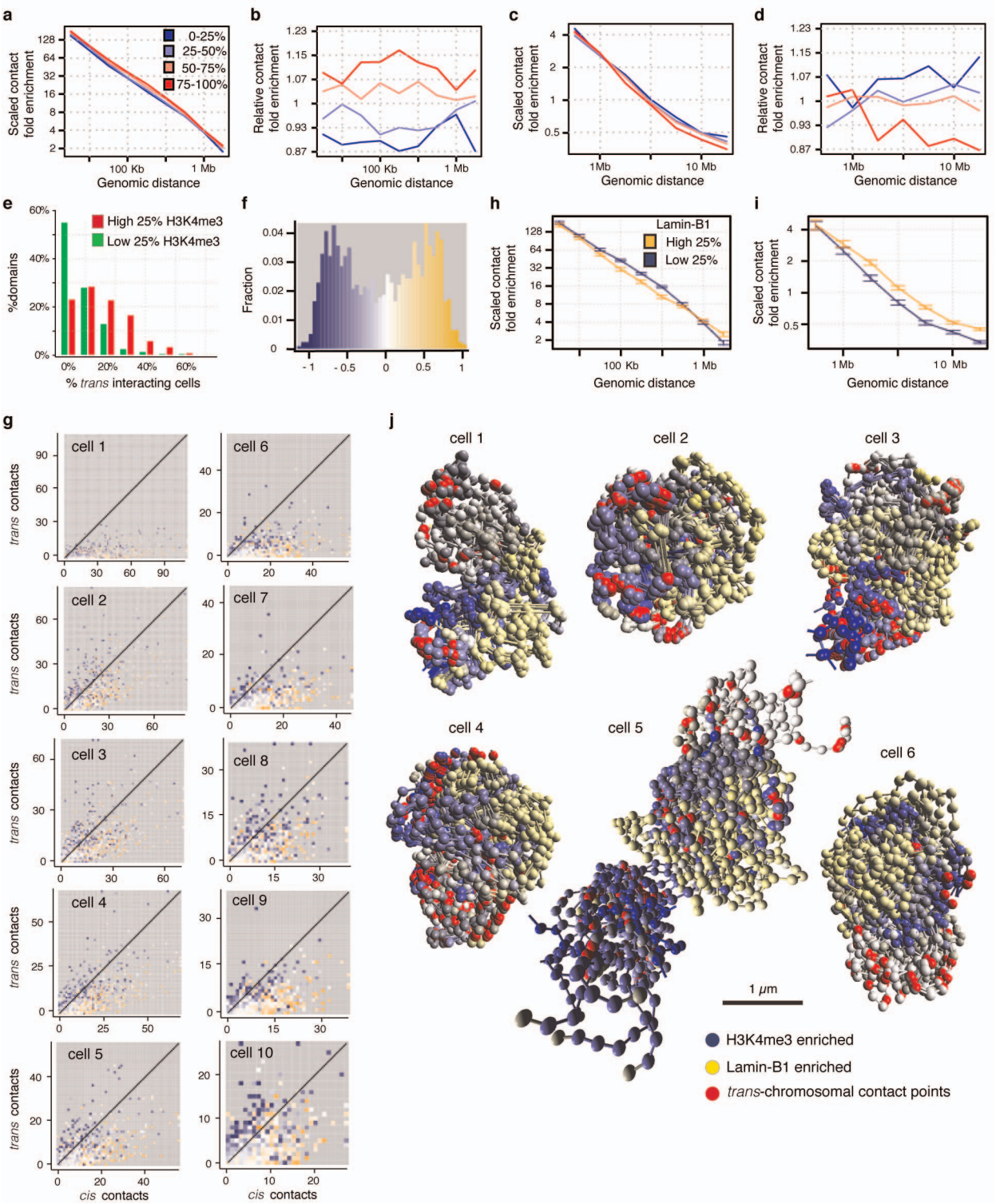
Extended Data Figure 2 | Chromosomal domains. **a**, Ratios between intradomain and interdomain contact enrichments over genomic distance. The mean single-cell trend is shown in black. Chromosomes are grouped into four groups: group 1 (chromosomes 1, 8, 15, 16 and X), group 2 (chromosomes 2, 6, 10, 13 and 18), group 3 (chromosomes 3, 5, 11, 14 and 17) and group 4 (chromosomes 4, 7, 9, 12 and 19). The intra- over interdomain enrichment is persistent in all chromosome groups and does not seem to stem from peculiar chromosomes. **b**, Distribution of correlations between intradomain contact numbers of all domains from pairs of real and reshuffled controls. **c**, Distribution of the insulation score at each fend in nine single-cell Hi-C data sets (where RE1 is BglII; real cells) is shown in red. Fifty sets of reshuffled cells were produced (see Supplementary Methods) and their insulation score distribution is shown in black. Real cells have a heavier tail of highly insulating loci, which is indicative of non-uniform and cell-specific interdomain contact structure.



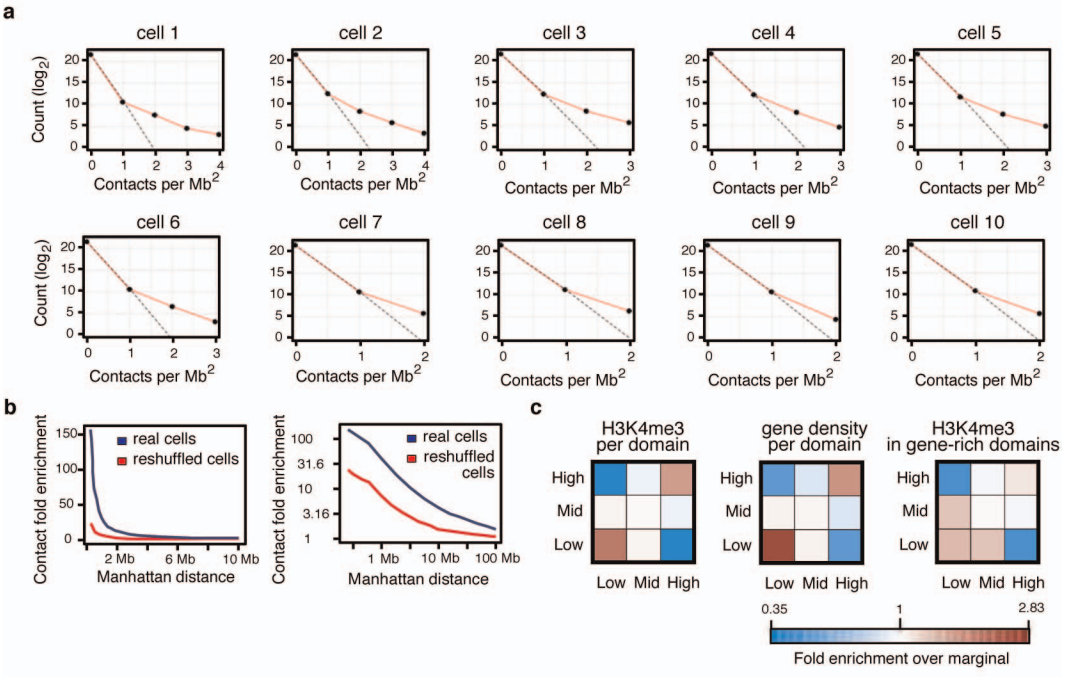
Extended Data Figure 3 | Modelling protocol quality controls. **a**, Results of structure calculations using restraints from a space-filling Hilbert curve test structure with 4,096 particles and four typical results of structure modelling using different numbers of restraints are shown (upper panels). Structure calculations of the Hilbert curve from random positions using different sets of 1,024 restraints (lower panel). **b**, Comparison of r.m.s.d. values from Hilbert curve and single-cell X-chromosome models. Structure calculations for Hilbert curves were repeated 100 times with variable numbers of restraints as shown. The root mean square deviation (r.m.s.d.) values between 100 models (precision) using the indicated number of restrains (mean ± s.d.) are plotted in blue. The r.m.s.d. values between the original Hilbert curve and each of the 100 models (accuracy) for the same numbers of restraints are plotted in green (mean ± s.d.). r.m.s.d. values from 100 repeated calculations of fine-scale (50-kb backbone) X-chromosome structure from the seven single-cell data sets are also plotted (red; mean ± s.d.). **c**, Restraint violation analysis. The distances between directly restrained positions in fine-scale (50-kb backbone) X-chromosome models are shown. Models for the six single-cell data sets (cell 1 to cell 6; red) show no values exceeding the upper bound (dashed line). Calculations with six shuffled interaction maps (created from cell 1 data set; blue) show significant violations. Structure calculations performed on merged pairs of data sets (yellow; all possible combinations of cell 1 to cell 4) have a few violations and are significantly closer to the upper limit. **d**, Comparison of structure-derived distance matrix from 200 fine-scale X-chromosome models from cell 1 (orange) and its single-cell Hi-C contacts (black crosses). The orange colour indicates the minimum distance between backbone particles. **e**, Comparison of X-chromosome structural models for six cells computed using low-resolution (500-kb binned) single-cell Hi-C interaction data. The bundles shown represent minimised structural alignments of five models from repeat calculations for each cell. Colours indicate chromosomal positions as shown. Scale bar, 1 μm.



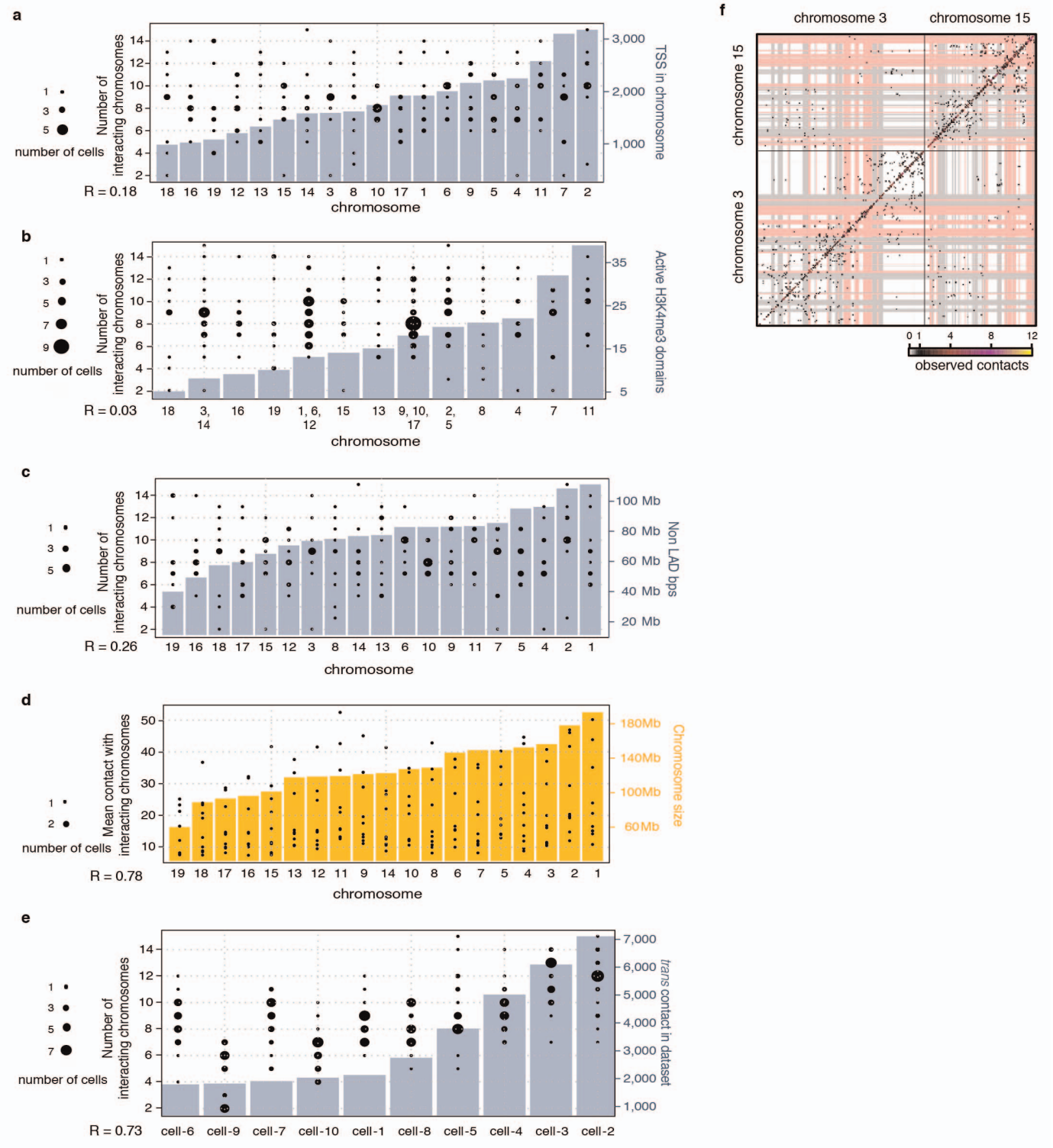
Extended Data Figure 4 | Comparison and investigation of models. **a**, Pair-wise comparison of fine-scale X chromosome structural models by r.m.s.d. analysis. Each pixel represents an r.m.s.d. value for a pair-wise comparison of two models. Lighter pixels indicate structures of higher similarity (low r.m.s.d.). Diagonal elements have been excluded. The order of 200 models in each panel was determined by hierarchical clustering of the r.m.s.d. values. Numbers shown are the mean r.m.s.d. values and the standard deviations for all the comparisons for each cell calculated by comparing the Hi-C contact particles. **b**, Cell-to-cell comparison of 200 fine-scale X chromosome structural models by r.m.s.d. analysis. Each pixel represents an r.m.s.d. value for a pair-wise comparison of two models. **c**, Fine-scale X-chromosome structures calculated from cell 1 and cell 3 data sets, and a structure from the combined data set. Colours indicate chromosomal positions as shown. Scale bar, 1 μm. **d**, Typical structure calculated using a randomized data set, where the interacting points for cell 1 have been shuffled with a pairing probability proportional to one over the square root of the sequence separation. Colours and scale as shown in **c**. **e**, Distribution of measurements of depth from the surface for five loci P1–P5 (Fig. 3e) in 1,200 X-chromosome models (200 fine-scale models for each of the six cells). Whiskers on box plots define 10th and 90th percentiles and the outliers are shown as individual dots.



Extended Data Figure 5 | Epigenomic landscape of chromosomes. **a**, Intradomain contact enrichment for each quartile of *trans*-chromosomal contacting domains. **b**, Same as **a** but subtracting the mean quartile enrichment in each genomic distance emphasizing the differences shown in **a**. **c**, Using the same sets as in **a** but plotting the enrichment of interdomain contacts within the same chromosome. **d**, Same as **c** but subtracting the mean quartile enrichment in each genomic distance. **e**, Percentage of cells in which high and low H3K4me3 enriched domains are *trans*-interacting. For each cell the domains with top 10th percentile *trans* intensity were defined as *trans*-interacting in that cell. We then counted for each domain the fraction of cells in which that domain was *trans*-interacting. Shown are the distributions of these fractions for H3K4me3 enriched and non-enriched domains (the top and bottom 25th percentiles, respectively). **f**, Distribution of the average lamin B1 (ref. 26) enrichment in chromosomal domains, colour coded according to the enrichment value. **g**, Domains plotted according to their number of *trans*- and *cis*-chromosomal (but excluding intradomain) contacts, colour-coded as in **f**. The domain lamin B1 enrichment and H3K4me3 peak density are highly anti-correlated (Spearman's correlation = −0.73). **h**, Intradomain contact enrichment for high versus low quartile of domains stratified by their mean lamin B1 enrichment. Error bars indicate 95% confidence intervals. **i**, Using the same sets as in **h** but plotting the enrichment of interdomain contacts within the same chromosome. Error bars as in **h**. **j**, Lamin B1 domains show a minor decrease in intradomain contact intensities that might suggest less compacted domains, and significantly increased *cis*-interdomain contact, maybe owing to lack of *trans*-chromosomal contacts. Topology of lamin B1, H3K4me3 and *trans*-contacts on five-model bundles of low-resolution X-chromosome models. Regions of low mappability have been excluded. Scale bar, 1 μm.



Extended Data Figure 6 | Interchromosomal contacts. **a**, Comparison of observed coverage of the *trans*-chromosomal 1-Mb square bins of each cell (red lines), versus predicted coverage assuming a binomial model (random uniform distribution of contacts to bins; black dashed line). Observed coverage is consistently higher than the uniform model, indicating the highly non-random distribution of *trans*-chromosomal contacts to genomic bins. **b**, *Trans*-chromosomal contact enrichment around observed *trans*-contacts as a function of the contacts total distance on both chromosomes (Manhattan distance in the contact map). Observed and expected (by random uniform contact distribution) numbers of contacts are counted around each *trans*-contact, and their ratio is shown for the 9 real cells (blue; where RE1 is BglII) and reshuffled cells (red), at two different scales. **c**, Left panel, *trans*-contacts were classified according to H3K4me3 density of the domains they associate: High and low for top and bottom 25th percentiles, respectively, mid for 25th–75th percentiles. Shown is the log ratio of the contingency table counts with the expected counts generated by multiplying the corresponding marginal probabilities for each group (chi-squared test; $P = 5.8 \times 10^{-18}$). To make sure these phenomena are not caused by the *trans* enrichment of active domains and depletion of non-active ones, only the top 15th percentile trans enriched domains from each cell were used. Middle panel, similar to left panel but contacts are classified by their associated domain gene density (chi-squared test; $P = 2.3 \times 10^{-12}$). Right panel, similar to left panel but using domains in the top 40th percentile of gene density, classifying by their H3K4me3 density, to test H3K4me3 enrichment beyond gene density (chi-squared test; $P = 3.6 \times 10^{-06}$). In all cases active or gene-rich domains preferentially interact with each other, although active domains (high H3K4me3 density) show greater interaction than expected from their gene density.



Extended Data Figure 7 | Chromosomal interfaces. **a**, The number of interacting chromosomes per chromosome is depicted in circles sized according to the number of single cells the value was observed in, and the order of chromosomes is shown by the number of transcription start sites (TSSs) in each chromosome (blue bars). Only autosomes are displayed. Spearman correlation between the number of TSSs and the mean value of the number of interacting chromosomes per chromosome is 0.18. Two chromosomes were defined as interacting when they had at least one domain–domain interaction (see main text) supported by two or more contacts. The number of interacting chromosomes per chromosome rises together with the number of TSSs. However, the change is small, and the number of interacting autosomal chromosomes per chromosome (the plotted value divided by two) remains between 4 and 6. **b**, Same as **a** except that chromosomes ordered by the number of active H3K4me3 domains (the top 25th percentile H3K4me3 peak density domains). **c**, Same as **a** except that chromosomes are ordered by the number of non-lamin-associated domains (non-LAD) base pairs in the chromosome. The fraction of a chromosome covered by LADs ranges from 31% to 53% and is correlated with chromosome size (0.52 Spearman). Thus, chromosome lengths span a range of 3.2-fold change, while their non-LAD fraction spans a smaller range of 2.8-fold change. **d**, Examination of the number of contacts between two chromosomes and the chromosomes sizes. The mean number of contacts of each chromosome with others it interacts with is shown for the ten single cells, with chromosomes ordered by their size. Chromosome size is correlated with the number of contacts it has, but the dynamic range of this number is small. **e**, The number of interacting chromosomes per chromosome is depicted in circles sized as the number of single cells the value was observed in, and the ten single-cell data sets are ordered by the number of *trans*-contacts in each data set, shown by blue bars. Only autosomes are displayed. Spearman correlations between the number of *trans*-contacts in each data set and the mean value of the number of interacting chromosomes per chromosome is 0.73. The number of interacting chromosomes per chromosome rises together with the coverage. However, the change is small, and the number of interacting autosomal chromosomes per chromosome (the plotted value divided by two) remains between 4 and 6. **f**, Example of multi-way chromosomal interfaces. Contact map of chromosomes 3 and 15 in cell 5. Shown is the number of contacts in 1-Mb size bins. Top and bottom 30th percentiles of H3K4me3 peak density domains are marked in light pink and light grey, respectively. Note the grid-like *trans*-contacts arrangement, and the correspondence between the two large *trans*-contact clusters and the organization of *cis*-contacts in both chromosomes to large 'mega domains'.

Extended Data Table 1 | Testing sequencing saturation.

a

Sample	Read-pairs		Fend-pairs			%valid non-amplified
	Original	Re-sequenced	Original	Combined	Addition	
cell 5	1,485,494	13,882,742	13,118	19,373	47.68 %	70.20 %
cell 1	1,253,954	6,210,472	26,318	30,671	16.54 %	51.85 %
cell 9	2,413,574	4,079,969	10,040	11,711	16.64 %	34.05 %
cell 8	1,132,934	12,437,493	10,043	12,231	21.79 %	71.71 %
cell 3	1,420,807	4,096,785	21,414	24,300	13.48 %	47.04 %

b

cell 5		Re-sequenced coverage			
		0	1	≥ 2	
Original coverage	0	0	7088	5635	
	1	170	69	563	5.76 %
	≥ 2	28	26	13078	94.24 %
			27.15 %	72.85 %	

cell 8		Re-sequenced coverage			
		0	1	≥ 2	
Original coverage	0	0	2754	1331	
	1	220	87	778	9.74 %
	≥ 2	33	28	9997	90.26 %
			19.16 %	80.84 %	

cell 1		Re-sequenced coverage			
		0	1	≥ 2	
Original coverage	0	0	3908	3240	
	1	821	337	1247	8.36 %
	≥ 2	1092	295	24962	91.64 %
			13.36 %	86.64 %	

cell 3		Re-sequenced coverage			
		0	1	≥ 2	
Original coverage	0	0	1270	1718	
	1	1010	413	1264	11.13 %
	≥ 2	1141	534	19777	88.87 %
			8.88 %	91.12 %	

cell 9		Re-sequenced coverage			
		0	1	≥ 2	
Original coverage	0	0	1378	1204	
	1	724	260	508	12.93 %
	≥ 2	663	261	9126	87.07 %
			14.91 %	85.09 %	

Several single-cell libraries were extensively re-sequenced. **a**, Shown are the numbers of read pairs in the original and re-sequenced runs, the number of fend pairs in the original run, the number of fend pairs when combining the sequences of the two runs, and the addition to fend pairs that the re-sequencing contributed. The percentages of singly covered fend pairs in the original sample that were supported by more read-pairs in the re-sequenced one are shown (%valid non-amplified). These fend pairs were discarded as potential spurious pairs in the original run, but proved by the re-sequencing to be valid pairs. This gives a sense of the fraction of valid pairs we discard when removing the read pairs suspected to be sequencing pairing errors. **b**, Shown are fend-pair coverage contingency tables of the original and the re-sequenced runs for the five single cells.

Extended Data Table 2 | Testing intercellular spurious ligations.

Library	Total read -pairs	%mm9-mm9	mean% unexpected	%hg18-hg18	%hg18-mm9	mean% hg 18-mm9
Group A_1	2232565	99.452	→ 0.103	0.543	0.005	0.007
Group A_2	2181145	99.955		0.039	0.007	
Group A_3	3170076	99.987		0.006	0.007	
Group A_4	4528440	99.988		0.004	0.008	
Group A_5	2214693	99.967		0.025	0.009	
Group A_6	2887996	99.992		0.002	0.006	
Group B_7	3058729	64.053		35.936	0.011	0.009
Group B_8	4341817	99.976		0.017	0.007	
Group B_9	3262211	61.122		38.869	0.009	
Group B_10	3063179	99.986		0.005	0.009	
Group B_12	3662647	99.326		0.668	0.006	
Group C_1	574035	0.132	← 0.057	99.859	0.010	0.008
Group C_2	2186006	0.146		99.847	0.007	
Group C_3	653496	0.010		99.983	0.007	
Group C_4	965577	0.025		99.971	0.004	
Group C_5	1578045	0.031		99.957	0.012	
Group C_6	891922	0.001		99.992	0.007	
Group C_7	1117686	99.975	→ 0.027	0.015	0.010	0.010
Group C_8	970213	99.974		0.019	0.007	
Group C_9	1338728	99.966		0.028	0.006	
Group C_10	1043818	99.982		0.007	0.011	
Group C_11	2351657	99.906		0.083	0.010	
Group C_12	3202531	99.975		0.012	0.013	
Group D_1	1736070	99.962	→ 0.074	0.033	0.006	0.007
Group D_2	1754172	99.796		0.192	0.012	
Group D_4	1093540	99.813		0.180	0.008	
Group D_5	1554973	99.962		0.034	0.005	
Group D_6	1958313	99.885		0.112	0.003	
Group D_7	1730004	99.991		0.000	0.009	
Group D_8	2097397	99.954		0.042	0.004	
Group D_9	2105599	99.963		0.030	0.007	
Group D_12	2059016	99.946		0.047	0.007	

Mouse and human nuclei or single-cell Hi-C samples were mixed in different stages of the experiment (group A, before fixation; group B, before library construction (so all the mouse and human samples in each library have the same identification tag); group C, before library amplification (so mouse and human samples in each library have different identification tags)). We created single-cell (for group A) or human and mouse two-cell (for groups B and C) Hi-C libraries and analysed them. The table shows the percentages of the three possible read pairs: mouse–mouse (mm9–mm9), human–human (hg18–hg18) and human–mouse (hg18–mm9). The expected pair type in each library is marked in blue. Mean percentage of unexpected read pairs per lane are also shown. For group A, we selected mouse cells based on morphology. In Group A, all six libraries contain almost exclusively mouse–mouse read pairs with insignificant human–human or mouse–human pairs. Each group B library has both human–human and mouse–mouse read pairs as expected, and the number of spurious human–mouse read pairs is extremely low. In each group C library, which was created by amplifying the distinctly tagged human (C1–C6) and mouse (C7–C12) single-cell samples in the same tube (for example, C1 and C7, C2 and C8, etc.), the fractions of foreign pairs (human reads with a mouse tag and vice versa) and of spurious pairs (human–mouse) were consistently extremely low.

To estimate the fraction of foreign and spurious pairs that could have originated simply from mapping a truly pure mouse library to a concatenated human-mouse genome, libraries from pure mouse cells (group D) were mapped to such a genome. The mean percentages of both foreign and spurious fend pairs in this lane are the same as those found in the different human–mouse mixed lanes, suggesting there is no intercellular contamination.

Extended Data Table 3 | Sequencing pairing errors.

	lane_A	lane_B	lane_C	lane_D
% phiX loaded / lane capacity	10 %	10 %	25 %	1 %
% phiX reads	12 %	15 %	40 %	2 %
Total number of read pairs	24.4 M	21.5 M	12.8 M	25.3 M
phiX – phiX pairs	3 M	3.3 M	5 M	0.5 M
phiX – mm 9 pairs	674	553	323	40
mm 9 – phiX pairs	3266	1647	3159	200
Estimated spurious mm 9 - mm 9 fend-pairs with the same identification tag	~3000	~1500	~300	~1000

PhiX174 DNA library was added to four lanes of single-cell Hi-C multiplexed libraries. In theory, no mixed mouse–phiX174 read pair is expected, but in fact a small number were detected. Shown are the fraction of phiX174 DNA loaded to each lane capacity, the percentage of phiX174 read ends in the lane, and the observed number of read pairs by type. The pairing probability was crudely estimated from these figures, and from it the number of expected spurious mouse–mouse read-pairs was calculated. Most of these spurious pairs are discarded due to mismatching unique identification tags at the beginning of each read end. Shown is the estimated number of spurious mouse pairs that coincidently have matching identification tag and are therefore not detected and removed.

doi:10.1038/nature12587

Deterministic direct reprogramming of somatic cells to pluripotency

Yoach Rais[1]*, Asaf Zviran[1]*, Shay Geula[1]*, Ohad Gafni[1], Elad Chomsky[1], Sergey Viukov[1], Abed AlFatah Mansour[1], Inbal Caspi[1], Vladislav Krupalnik[1], Mirie Zerbib[1], Itay Maza[1], Nofar Mor[1], Dror Baran[1], Leehee Weinberger[1], Diego A. Jaitin[2], David Lara-Astiaso[2], Ronnie Blecher-Gonen[2], Zohar Shipony[3,4], Zohar Mukamel[3,4], Tzachi Hagai[5], Shlomit Gilad[6], Daniela Amann-Zalcenstein[6], Amos Tanay[3,4], Ido Amit[2], Noa Novershtern[1] & Jacob H. Hanna[1]

Somatic cells can be inefficiently and stochastically reprogrammed into induced pluripotent stem (iPS) cells by exogenous expression of Oct4 (also called Pou5f1), Sox2, Klf4 and Myc (hereafter referred to as OSKM). The nature of the predominant rate-limiting barrier(s) preventing the majority of cells to successfully and synchronously reprogram remains to be defined. Here we show that depleting Mbd3, a core member of the Mbd3/NuRD (nucleosome remodelling and deacetylation) repressor complex, together with OSKM transduction and reprogramming in naive pluripotency promoting conditions, result in deterministic and synchronized iPS cell reprogramming (near 100% efficiency within seven days from mouse and human cells). Our findings uncover a dichotomous molecular function for the reprogramming factors, serving to reactivate endogenous pluripotency networks while simultaneously directly recruiting the Mbd3/NuRD repressor complex that potently restrains the reactivation of OSKM downstream target genes. Subsequently, the latter interactions, which are largely depleted during early pre-implantation development *in vivo*, lead to a stochastic and protracted reprogramming trajectory towards pluripotency *in vitro*. The deterministic reprogramming approach devised here offers a novel platform for the dissection of molecular dynamics leading to establishing pluripotency at unprecedented flexibility and resolution.

Induced pluripotent stem cells can be generated from somatic cells by ectopic expression of different transcription factors, originally Oct4, Sox2, Klf4 and Myc (OSKM)[1]. The reprogramming process requires initial cell proliferation, after which a fraction of the cell progeny successfully converts into an embryonic stem (ES)-like state with different time latencies[2,3]. A variety of chromatin modifiers have been implicated in facilitating epigenetic changes leading to authentic iPS cell reprogramming[4,5]. Despite these advances, the reprogramming efficiency of somatic cells remains extremely low[3]. Furthermore, the outcome of challenging the somatic epigenome with the overexpression of OSKM reprogramming factors is stochastic[2]. Experimental and theoretical modelling approaches for characterizing the nature of stochastic elements acting in iPS cell reprogramming have suggested that the existence of as few as one dominant rate-limiting element may adequately recapitulate the experimentally measured kinetics for clonal iPS cell formation by OSKM factors[2,6]. The identity of such stochastic rate-limiting element(s) remains to be defined. Here we show that the Mbd3/NuRD repressor complex[7] is the predominant molecular block preventing deterministic induction of ground-state pluripotency.

Mbd3 and establishment of naive pluripotency

We tested whether additional genetic manipulations may enable deterministic reprogramming towards ground-state pluripotency by OSKM factors, where all donor somatic cells and their progeny synchronously convert into iPS cells. Recent studies have pointed out the importance of chromatin derepression in converting somatic cells into iPS cells[5,8,9]. Therefore, we aimed to conduct a loss-of-function screen for selected epigenetic repressor factors in an attempt to markedly boost the efficiency of reprogramming to ground-state pluripotency. We initially focused on reverting murine primed pluripotent epiblast stem cells (EpiSCs) that can stochastically convert within 5 days into a naive pluripotent state in 2i/LIF growth conditions (where 2i is ERK1/2 and GSK3β inhibitors, and LIF is leukaemia inhibitory factor)[5]. We used a primed EpiSC line carrying a Nanog–GFP knock-in reporter that can be reactivated in the naive state[5], and applied short interfering RNA (siRNA) screening to identify boosters of EpiSC reversion into Nanog–GFP^{+} naive cells (Fig. 1a and Extended Data Fig. 1a). Notably, only Mbd3 inhibition markedly increased the EpiSC reversion efficiency, where up to 80% of the transfected cells turned on Nanog–GFP in 2i/LIF conditions (Fig. 1a).

Mbd3 is a key component in the NuRD complex, ubiquitously expressed in all somatic cells[10]. Mbd2 and Mbd3 assemble into mutually exclusive distinct NuRD complexes[11], which can mediate gene repression through histone deacetylation and chromatin remodelling activities. To validate the siRNA screening results, we used *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ ES cells and introduced *Rosa26-creER* and *Nanog-GFP* knock-in alleles before converting them into EpiSCs (Fig. 1b and Extended Data Fig. 1b–e). Clonal analysis for epigenetic reversion of EpiSCs demonstrated 95% Nanog–GFP^{+} single-cell reversion efficiency in Mbd3-null cells (Fig. 1b). *Mbd3*$^{fl/-}$ EpiSCs, which retain hypomorphic (~20%) Mbd3 protein expression levels (Fig. 1c), also yielded reverted ES cells with >90% efficiency (Fig. 1b). Both reverted *Mbd3*$^{-/-}$ (after transgenic insertion of Mbd3 to rescue their differentiation deficiency[10,12]) and *Mbd3*$^{fl/-}$ cells can contribute to chimaera formation (Extended Data Fig. 1f). Reconstitution of Mbd3 expression in *Mbd3*$^{-/-}$ and *Mbd3*$^{fl/-}$ EpiSCs inhibited reversion efficiencies (Fig. 1b). These results directly demonstrate that reduction of Mbd3 protein levels renders nearly complete reversion of EpiSCs to naive pluripotency.

We revisited ES-cell derivation experiments from *Mbd3*$^{-/-}$ E3.5 embryos[10], and were able to isolate *Mbd3*$^{-/-}$ ES cells in serum-free 2i/LIF

[1]The Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]The Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel. [3]The Department of Biological Regulation, Weizmann Institute of Science, Rehovot 76100, Israel. [4]The Department of Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel. [5]The Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. [6]The Israel National Center for Personalized Medicine, Weizmann Institute of Science, Rehovot 76100, Israel.
*These authors contributed equally to this work.

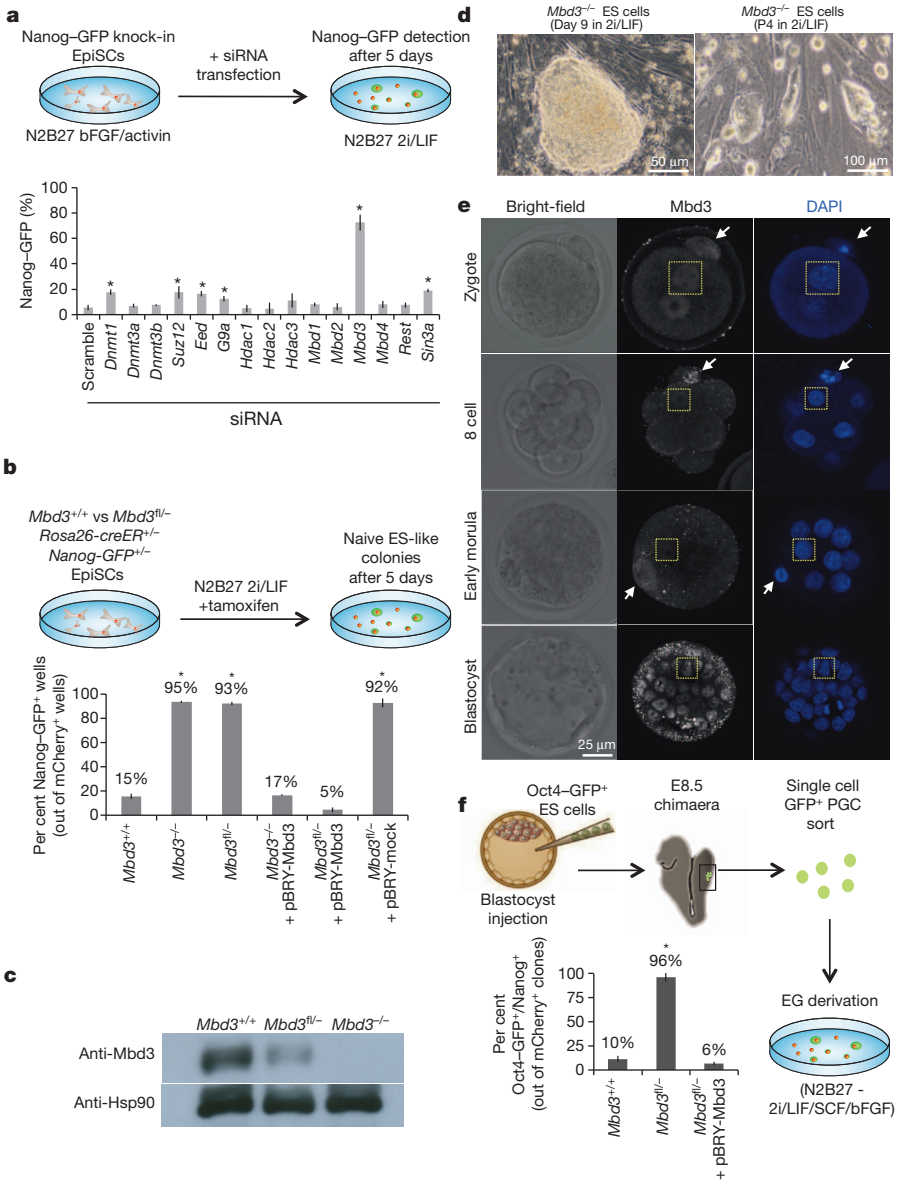


Figure 1 | Boosting primed to naive pluripotency reversion. **a**, An siRNA screen for factors that can boost epigenetic reversion of primed EpiSCs into naive ES cells. Percentage of naive Nanog–GFP$^+$ cells detected by flow cytometry is indicated ($n = 3$). **b**, Single-cell reprogramming efficiency and quantification for EpiSC reprogramming from different mutant lines. The pBRY-Mbd3 rescue construct was stably expressed in the indicated lines ($n = 4$). **c**, Western blot analysis for Mbd3 expression in ES cells. **d**, *Mbd3*$^{-/-}$ ES cell derivation from blastocysts in 2i/LIF. **e**, Representative confocal immunostaining images for temporal Mbd3 expression in developing mouse embryos. Arrows indicate polar body ($n = 15$ embryos stained per stage). Scale bar, 25 µm. **f**, *Mbd3*$^{fl/-}$ and *Mbd3*$^{+/+}$ ES cell lines (with or without pBRY-Mbd3 overexpression allele) were targeted with an Oct4–GFP reporter and a constitutively expressed mCherry reporter, and injected into host chimaeras. Embryonic day (E)8.5 primordial germ cells (PGC) were sorted into defined conditions and evaluated for efficiency to generate EG cells ($n = 6$). Asterisk indicates *t*-test *P* value <0.01 in comparison to *Mbd3*$^{+/+}$. All error bars indicate s.d. from average.

conditions (Fig. 1d and Extended Data Fig. 2a, b). This indicates that Mbd3 is dispensable for establishing the ground state of pluripotency and ES cell derivation. Consistent with an antagonistic role for Mbd3 in establishing pluripotency, Mbd3 is largely depleted after fertilization and throughout pre-implantation development (from 4-cell until early morula stages), and its nuclear expression becomes consolidated at the late morula, blastocyst and post-implantation epiblast (Fig. 1e and Extended Data Fig. 2c, d)[10]. These results indicate that early pre-implantation *in vivo* reprogramming and development are accompanied by depletion of Mbd3 expression, which gets re-expressed as pluripotency is consolidated in the inner cell mass (ICM). The latter dynamic Mbd3 expression pattern is also consistent with a critical role for Mbd3 in restricting aberrant trophoblast lineage specification of the ICM and facilitating adequate differentiation of the assembled pluripotent epiblast[13]. Finally, we aimed to test the influence of reducing Mbd3 expression in reprogramming Oct4$^+$ primordial germ cells (PGCs) into ES-like pluripotent embryonic germ (EG) cells[14]. Single cell isolated *Mbd3*$^{fl/-}$ Oct4–GFP$^+$ E8.5 PGCs from chimaeric mice were proficient in forming EG cell colonies and lines (>95% efficiency), whereas PGCs isolated from chimaeras that were generated by micro-injecting *Mbd3*$^{+/+}$ or *Mbd3*$^{fl/-}$ cells carrying an exogenous *Mbd3* transgene reprogrammed at less than 10% efficiency (Fig. 1f). Collectively, these findings show that neutralizing Mbd3 expression facilitates access to ground-state pluripotency from early embryonic Oct4-expressing cells.

Deterministic reprogramming of somatic cells

We next moved to test whether Mbd3 depletion in somatic cells facilitates their conversion to pluripotency at efficiencies nearing 100%. Recent studies have described a mild positive effect for Mbd3 short-hairpin RNA (shRNA)-mediated knockdown on mouse iPS cell formation[15], and a negative effect on human primed iPS cell induction[16]. We revisited these experiments while using optimized *Mbd3* genetic depletion, OSKM transgene delivery and 2i/LIF containing naive pluripotency conditions (Extended Data Fig. 3a–d). Notably, 95% of *Mbd3*$^{fl/-}$ and *Mbd3*$^{-/-}$ cells were Oct4–GFP$^+$ at day 10, whereas only levels up to 18% were observed in control *Mbd3*$^{+/+}$ fibroblasts (Fig. 2a). To evaluate reprogramming efficiencies accurately, we established 'secondary reprogrammable'[17] *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ transgenic cell lines harbouring a doxycycline-inducible OKSM polycistronic cassette[18], a constitutive nuclear mCherry marker (to track individual cells and control for plating efficiency), and an Oct4–GFP reporter (Extended Data Fig. 3a). Single cell sorting of secondary mCherry$^+$ *Mbd3*$^{fl/-}$ mouse embryonic fibroblasts (MEFs) and subsequent reprogramming in 2i/LIF plus doxycycline conditions reproducibly yielded 100% iPS cell derivation efficiency by day 8. Wild-type

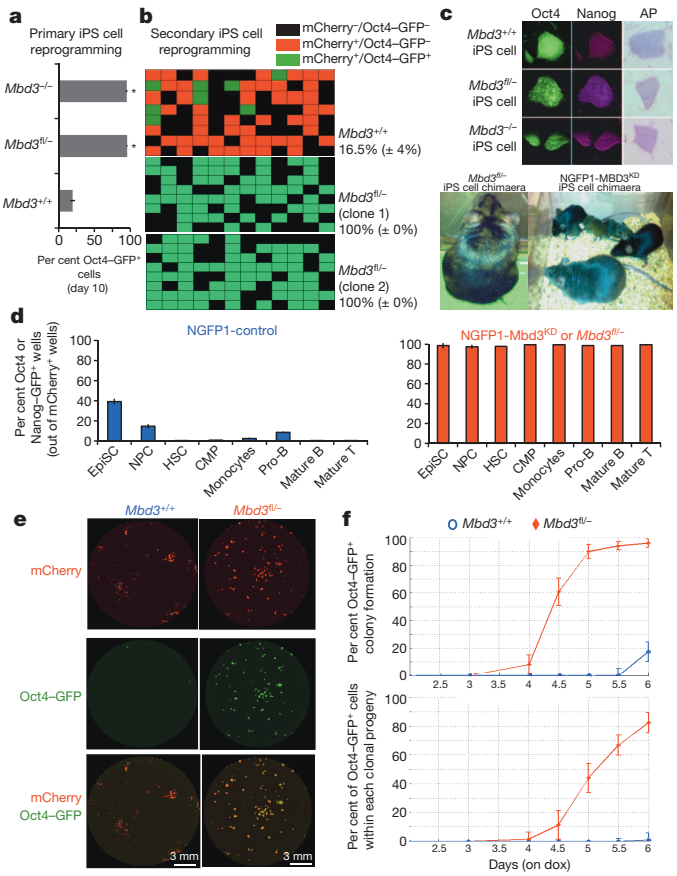


Figure 2 | Deterministic and synchronized iPS cell reprogramming. **a**, Mbd3 wild-type and depleted ($Mbd3^{fl/-}$ or $Mbd3^{-/-}$) MEFs were directly infected with lentiviruses expressing a polycistronic OKSM cassette. Reprogramming efficiency (Oct4–GFP) was measured by FACS at day 10. Asterisk indicates *t*-test *P* value <0.01 in comparison to $Mbd3^{+/+}$ samples. Error bars indicate s.d. from average ($n = 5$). **b**, Secondary reprogrammable fibroblasts carrying an Oct4–GFP reporter and an mCherry constitutively expressed marker were single-cell-seeded and subjected to doxycycline-induced reprogramming. Reprogramming efficiency at day 8 was calculated by dividing the number of Oct4–GFP$^+$ wells by mCherry$^+$ wells ($n = 3$ per clone, ± indicates s.d. from average). **c**, Top: immunostaining of representative iPS cell clones for pluripotency markers (original magnification, ×100). Bottom: agouti-coat-coloured chimaera and germline transmission from Mbd3-depleted iPS cells. **d**, The indicated somatic cell types from NGFP1-control or NGFP1-Mbd3^{KD} adult chimaeras were isolated and subjected to single-cell reprogramming and evaluation of Nanog–GFP expression after 8 days of doxycycline ($n = 4$). CMP, common myeloid progenitor; HSC, haematopoietic stem cell; NPC, neural precursor. **e**, Full-well mosaic images of mCherry, Oct4–GFP and combined channels, shown for $Mbd3^{fl/-}$ and $Mbd3^{+/+}$ at day 6. Scale bar, 3 mm. **f**, Characterization of Oct4–GFP$^+$ dynamic for $Mbd3^{fl/-}$ (red plot) and $Mbd3^{+/+}$ (blue plot) based on live imaging. Top graph indicates cumulative Oct4–GFP$^+$ colonies; bottom graph indicates fraction of Oct4–GFP$^+$ cells within colonies. Dox, doxycycline. Graphs show means and s.d. of all tracked colonies in three biological replicates (one out of four biological data sets is shown).

cells reprogrammed under identical conditions, no more than 20% of clones reactivated Oct4–GFP (Fig. 2b). Teratomas (not shown) and chimaeras were obtained from iPS cell clones (Fig. 2c). High single-cell reprogramming efficiency rates were obtained from a variety of adult progenitor and terminally differentiated cells (Fig. 2d and Extended Data Fig. 3e, f).

We analysed the reprogramming dynamics of 'secondary' $Mbd3^{fl/-}$ and control $Mbd3^{+/+}$ fibroblasts by applying microscopic live imaging and an algorithm that allows segmentation of single mCherry$^+$ colonies and tracking of Oct4–GFP reactivation dynamics during reprogramming (Supplementary Videos 1–4). By day 6 after doxycycline induction, >98% of $Mbd3^{fl/-}$ clonal populations reactivated the Oct4–GFP pluripotency marker, whereas only up to 20% efficiency was detected in control samples reprogrammed in identical growth conditions (Fig. 2e, f). By day 6, approximately 85% of cells within each individual $Mbd3^{fl/-}$ clonal population became Oct4–GFP$^+$ cells, whereas <2% of cells within successfully reprogrammed $Mbd3^{+/+}$ clones turned on the Oct4–GFP marker (bottom panel in Fig. 2f). The latter unbiased quantitative analysis demonstrated a markedly intra- and interclonal synchronized reactivation of Oct4–GFP occurring during a narrow window in $Mbd3^{fl/-}$ clonal populations at days 4.5–5.5 (Fig. 2f and Supplementary Video 2), and highlights a marked increase in reprogramming synchrony and efficiency after Mbd3 depletion in OSKM-transduced somatic cells. Detection of Oct4–GFP by flow cytometry on polyclonal populations demonstrated similar iPS cell reprogramming kinetics (Extended Data Fig. 4a). Re-infection with lentiviruses encoding Mbd3, but not Mbd2, before day 5 of reprogramming had a profound inhibitory effect on iPS cell generation from $Mbd3^{fl/-}$ MEFs, whereas re-infection after day 5 had a diminished effect (Extended Data Fig. 4b). The above kinetic analysis indicates that Mbd3 can inhibit reprogramming when introduced before the final stages of reprogramming. However, once pluripotency is re-established, Mbd3 does not compromise the maintenance of pluripotency.

We next conducted global gene expression analysis on donor MEFs at days 0, 4 and 8 after doxycycline induction without cell passaging, and compared them to iPS cell and ES cell lines. Notably, by day 8, $Mbd3^{fl/-}$ donor cells were transcriptionally indistinguishable from multiple ES cell lines and subcloned established iPS cell lines (Fig. 3a). Genome-wide chromatin mapping for H3K27me3, H3K4me3 and H3K27ac histone marks by chromatin immunoprecipitation followed by sequencing analysis (ChIP-seq) also confirmed that only $Mbd3^{fl/-}$ transduced MEFs had assumed an ES-like chromatin profile by day 8 (Extended Data Fig. 5a). Genome-wide DNA methylation mapping by reduced representation bisulphite sequencing (RRBS) confirmed that an iPS cell/ES cell-like methylation pattern could be seen in the $Mbd3^{fl/-}$ polyclonal population sample after 8 days of doxycycline treatment (Extended Data Fig. 5b, c). Single cell polymerase chain reaction with reverse transcription (RT–PCR) analysis confirmed that nearly 100% of single cells tested expressed key endogenous pluripotency markers only in $Mbd3^{fl/-}$ reprogrammed samples (Extended Data Fig. 5d). Collectively, the above results indicate that Mbd3 depletion after OSKM induction yields authentic molecular re-establishment of pluripotency in the entire population of donor somatic cells and their progeny.

After the depletion of Mbd3 expression, we were not able to isolate stable, partially reprogrammed cells[19,20] that did not reactivate Oct4–GFP or Nanog–GFP and could be stably expanded *in vitro*, as typically can be obtained from OSKM-transduced wild-type somatic cells (Fig. 3b). We next took $Mbd3^{+/+}$ OSKM-transduced partially reprogrammed cells and attempted to complete their reprogramming by Mbd3 inhibition. Notably, by introducing *Mbd3* siRNA, all clones[19] markedly turned on Oct4–GFP or Nanog–GFP pluripotency markers after continued OSKM expression in 2i/LIF (Fig. 3c). To functionally test a conserved inhibitory role for MBD3 in human iPS cell reprogramming, we generated MBD3^{mut} human ES cells by gene editing with TALE nuclease effectors (Fig. 3d), and validated hypomorphic MBD3 protein expression in a selected bi-allelically targeted clone (Fig. 3e). Single-cell reprogramming efficiency was tested after introducing an OCT4–GFP knock-in reporter and a constitutive mCherry in MBD3^{WT} and MBD3^{mut} iPS cells harbouring doxycycline-inducible OSKM factors. Only *in vitro* differentiated secondary fibroblasts from MBD3^{mut} cells reprogrammed at near 100% efficiency in optimized conditions (Fig. 3f and Extended Data Fig. 6a). Markedly increased human iPS cell formation was obtained by applying *MBD3* siRNA starting at 2 days after doxycycline induction in MBD3^{WT} OSKM transgenic secondary fibroblasts (Extended Data Fig. 6b). *MBD3* siRNA treatment allowed reproducible generation of human iPS cells from adult human patient-specific fibroblasts only after two rounds of OSKM with

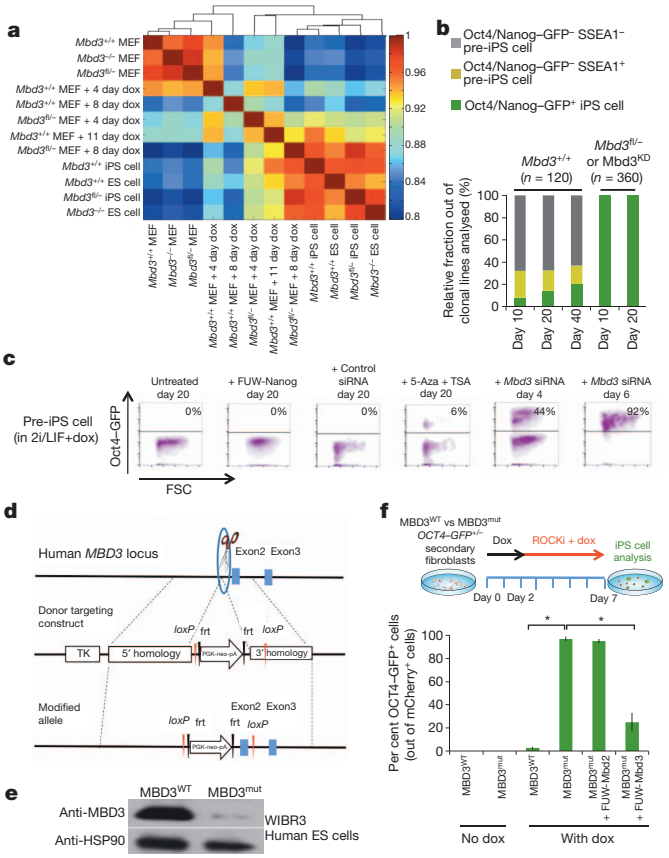


Figure 3 | Alleviating Mbd3 expression facilitates transition to pluripotency. **a**, Spearman correlation matrix between the indicated mouse samples, measured over gene expression levels of all 16,620 expressed genes. The matrix is clustered with hierarchical clustering. **b**, Monoclonal lines established from *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ secondary cells and reprogrammed in 2i/LIF plus doxycycline. Fraction of pre-iPS clones that did not reactivate Oct4/Nanog GFP markers (either SSEA1 positive or negative) is shown. **c**, Representative partially reprogrammed cell line containing OSKM transgenes, and that did not reactivate GFP reporter, was subjected to the indicated manipulations and analysed for completion of reprogramming as assayed by FACS. FUW indicates lentiviral backbone vector used. **d**, Targeting strategy for human *MBD3* locus. **e**, Biallelically targeted MBD3^{mut} clone displayed ~90% reduction in MBD3 protein expression levels. **f**, MBD3^{WT} and MBD3^{mut} iPS cells carrying doxycycline-inducible OKSM transgenes were labelled with constitutively expressed mCherry and targeted with an *OCT4-GFP* knock-in allele. ROCKi, Rho-associated protein kinase inhibitor. *In vitro* differentiated fibroblasts from the latter lines were reprogrammed as indicated in the scheme. Error bars indicate s.d. of average ($n = 6$). Asterisks indicate *t*-test *P* value <0.01 in comparison to *MBD3*^{WT}.

LIN28 mRNA transfection[21] (Extended Data Fig. 6c). Taken together, these results demonstrate that inhibiting MBD3 alleviates predominant obstacles for human iPS cell reprogramming.

Numerical modelling of reprogramming

We next sought to characterize the reprogramming latency distribution for both wild-type and Mbd3-depleted samples quantitatively, as this may allow a comparison of reprogramming dynamics to known deterministic and stochastic dynamic models. We applied a previously described approach for monoclonal pre-B-cell weekly follow-up for reactivation of Nanog–GFP (Fig. 4a)[2]. A secondary OSKM transgenic NGFP1-iPS cell line[2], carrying a Nanog–GFP reporter, was rendered transgenic for a doxycycline-inducible *Mbd3* shRNA construct (NGFP1-Mbd3^{KD}). Indeed, NGFP1-Mbd3^{KD}-derived monoclonal B-cell populations converted into Nanog–GFP^{+} iPS cells at day 7 at near 100% efficiency (Fig. 4b and Extended Data Fig. 7a). Subsequently, during the first 10 days of reprogramming we conducted daily Nanog–GFP

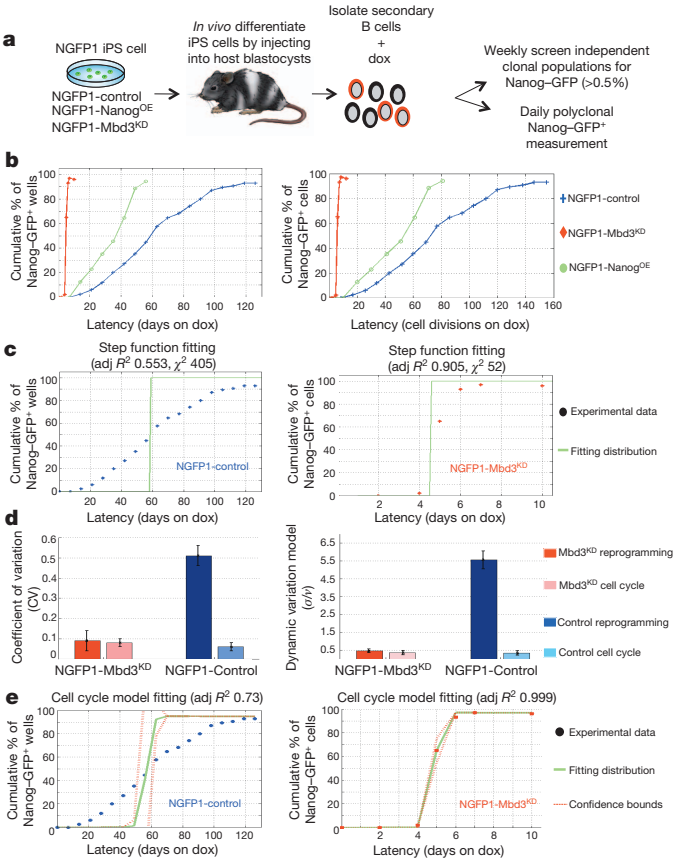


Figure 4 | Numerical description of reprogramming after Mbd3 depletion. **a**, Scheme demonstrating the monoclonal and polyclonal follow-ups for Nanog–GFP reactivation. KD, knockdown; OE, overexpression. **b**, Cumulative percentage of Nanog–GFP^{+} wells versus time on doxycycline, measured for various clonal B-cell-derived populations. **c**, Goodness of fit plots for the fitting of deterministic step function model to the observed reprogramming latency. **d**, Comparison of the calculated variation of *Mbd3*^{KD} and control (dark red and blue colours, respectively), and the cell cycle variation of each sample (light red and blue colours, respectively). Sample variation was calculated using coefficient of variation (left panel) and dynamic variation (right panel) (Extended Data Fig. 7b). Graphs show maximum likelihood estimations. Error bars indicate 95% confidence intervals of maximal likelihood value (Mbd3^{KD} reprogramming $n = 7$; control reprogramming $n = 13$; cell cycle for control and Mbd3^{KD} $n = 20$). **e**, Goodness of fit plots for the fitting of cell-cycle time distribution to the observed reprogramming latency.

detection on polyclonal NGFP1-control and NGFP1-Mbd3^{KD} B-cell populations. We next tested whether Mbd3^{KD} cell reprogramming behaves like a deterministic function. Such deterministic behaviour is well approximated by a step function with 0% reprogramming before a fixed deterministic time after doxycycline induction, and 100% iPS cell formation afterwards. Fitting the clonal cell reprogramming dynamics to such deterministic step-function revealed a tight fit ($R^2 > 0.9$, chi-squared = 52) of *Mbd3*^{KD} cells, but not of control *Mbd3*$^{+/+}$ cells ($R^2 = 0.55$, chi-squared = 405) (Fig. 4c). Despite the observed similarity to a deterministic behaviour, variability was still evident in our Mbd3^{KD} sample under the optimized conditions devised herein. Thus, we sought to quantify and compare the variability detected in the reprogramming latency measurements in both *Mbd3*$^{+/+}$ and *Mbd3*^{KD} samples, and further compare it to the inherent measured cell-cycle variability[2]. For this purpose we used two modelling schemes. By calculating the dynamic variability (s.d./drift) by a Brownian motion (BM) model (Extended Data Fig. 7b), and the coefficient of variation (s.d./mean) by Gaussian model fitting (Extended Data Fig. 7c), we show a consistent reduction in Mbd3^{KD} reprogramming variability and tight proximity to cell-cycle variability (Fig. 4d). To further support reduction in rate-limiting barrier(s)

on Mbd3 depletion, we modelled the dynamics of the OSKM reprogramming process using a multi-step Markov chain model (phase-type (PH) modelling[22]) (Extended Data Fig. 7d). Although this model does not directly argue for stochasticity, still there was a clear connection between the reduction in barriers and the reduction in process variability (Extended Data Fig. 7d). Finally, we proposed that the reduction in rate-limiting barriers in Mbd3^{KD} samples may reduce reprogramming variability to variability explained by cell cycle alone. For this purpose we fitted the observed reprogramming latency to a cell-cycle model that captures the required reprogramming duration[2,23]. We obtained a profound fit ($R^2 = 0.999$) between the Mbd3^{KD} dynamic and cell-cycle model, but not in the control *Mbd3*$^{+/+}$ dynamics ($R^2 = 0.73$) (Fig. 4e). Altogether, these results consistently show a reduction in OSKM reprogramming variability and increased proximity to deterministic dynamic behaviour upon Mbd3 depletion.

Deterministic reprogramming mechanism

We aimed to define the mechanisms for Mbd3 inhibition of iPS cell reprogramming. Inhibiting Mbd3 expression was not sufficient to induce iPS cell formation in the absence of exogenous OSKM overexpression in somatic cells (Extended Data Fig. 3e, f). Contrary to previous reports[15], Mbd3 depletion without OSKM expression does not independently lead to endogenous reactivation of bona fide pluripotency genes (Extended Data Fig. 8a). We established that Flag-tagged OCT4, KLF4, SOX2 and MYC specifically co-immunoprecipitated with Mbd3 after exogenous overexpression in HEK293 cells (Fig. 5a and Extended Data Fig. 9a). OSKM specifically co-immunoprecipitated with Mbd3/NuRD components in MEFs undergoing reprogramming (Extended Data Fig. 9b). These interactions were mediated via the MBD domain of Mbd3, as defined deletions introduced into the MBD domain abrogated co-immunoprecipitation of Mbd3 with OSKM (Extended Data Fig. 9c). Consistent with the direct protein interactions of the Mbd3/NuRD complex with OSKM reported above, genome-wide ChIP-seq analysis of Mbd3 binding in doxycycline-induced wild-type MEFs identified a global increase in Mbd3 recruitment and binding after OSKM induction (1,177 binding regions in MEFs compared to 8,657 after OSKM induction) (Supplementary Data set 1). Importantly, in somatic MEFs before OSKM induction, Mbd3 is not localized to pluripotency factor target genes (Fig. 5b). Only after doxycycline induction are Mbd3-bound genes enriched for targets of Klf4, Oct4, Sox2 and Esrrb ($P < 10^{-22}$) (Fig. 5b). The NuRD component Chd4 was similarly recruited to downstream targets only after doxycycline induction, indicating NuRD recruitment with Mbd3 (Fig. 5b). Chd4 knockdown in *Mbd3*$^{+/+}$ MEFs undergoing reprogramming enhanced iPS cell formation (Extended Data Fig. 8b). Transcription levels of Mbd3 target genes after 4 days of doxycycline were significantly upregulated in Mbd3-depleted samples (Fig. 5c), consistent with the predominant function of the Mbd3/NuRD complex as a repressor of pluripotency. Chromatin of Mbd3 and/or OSKM direct targets was significantly more active and open in Mbd3-depleted samples during reprogramming, including

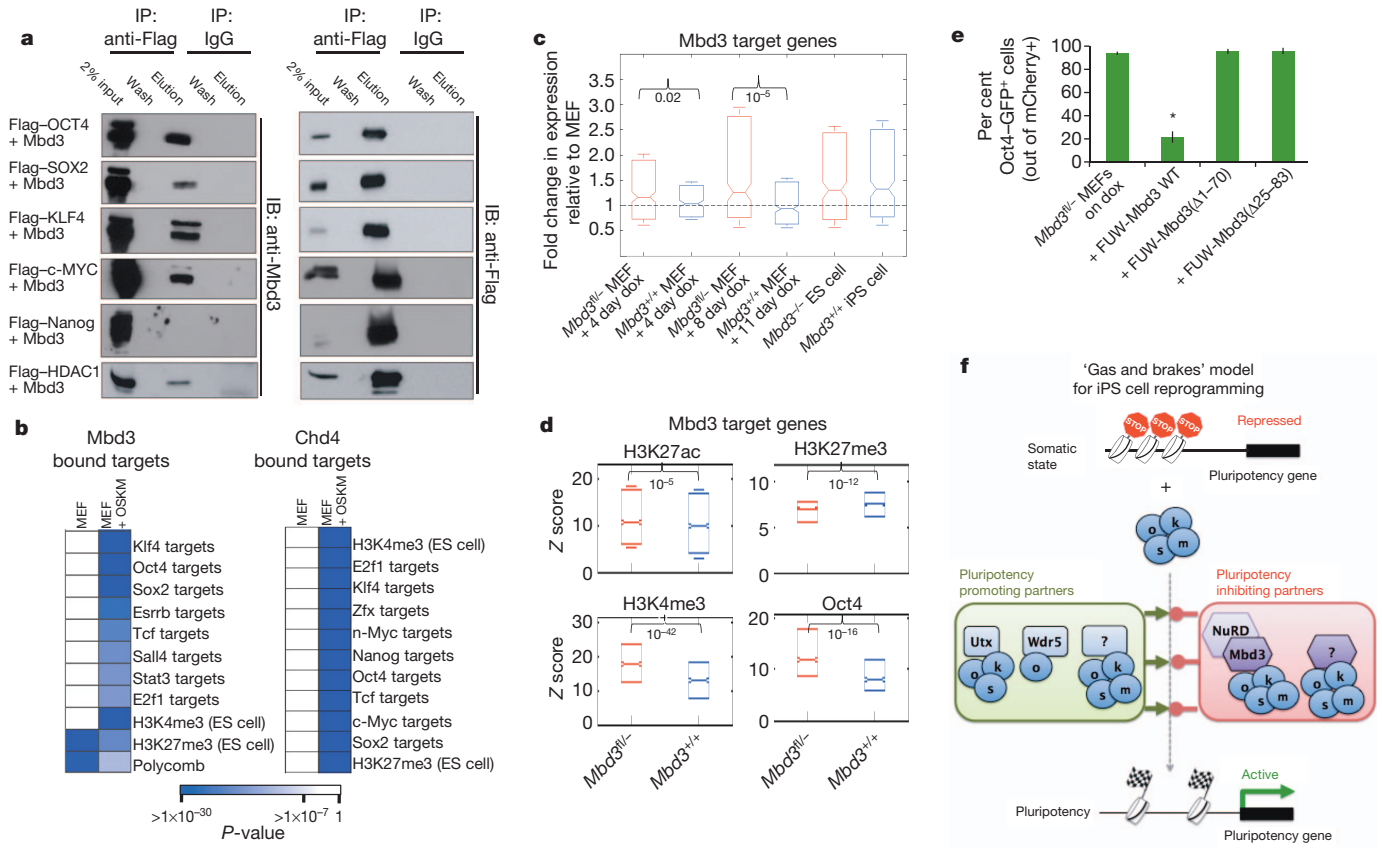


Figure 5 | Mechanisms for Mbd3 inhibitory effect on induced pluripotency. **a**, Constructs encoding Flag-tagged OCT4, SOX2, KLF4, MYC, Nanog or HDAC1 were transfected into HEK293T cells in combination with Mbd3. The cell lysates were immunoprecipitated (IP) with an anti-Flag antibody (or anti-IgG as control), followed by an immunoblot analysis (IB) ($n = 3$ biological replicates). **b**, Functional enrichment of Mbd3 and Mi2β (Chd4) direct targets, measured in MEFs before and after OSKM induction. Colour levels indicate enrichment *P* values (by Fisher's exact test) that pass the false discovery rate (FDR) threshold of 0.0001%. **c**, Distribution of gene expression fold change relative to MEFs of *Mbd3*$^{+/+}$ (blue) and *Mbd3*$^{fl/-}$ (red) samples throughout reprogramming. Graphs show box-plot medians and 25th/75th percentiles, and *P* values by paired sample *t*-test. **d**, Distribution of histone marks and Oct4 binding levels in *z*-score values at day 4 after OSKM (doxycycline) induction. Graphs show box-plot medians and 25th/75th percentiles, and *P*-values by paired sample *t*-test. **e**, Reprogramming efficiency of *Mbd3*$^{fl/-}$ MEFs after infection with lentiviruses encoding wild type and different mutant Mbd3 inserts. Error bars indicate s.d. from average ($n = 6$). Asterisk in **e** indicates *t*-test *P* value <0.01. **f**, Mechanistic model scheme.

statistically significant higher levels of Oct4 binding, H3K4me3 and H3K27ac, and reduced H3K27me3 repressive chromatin mark (Fig. 5d and Extended Data Fig. 8c–e). Mbd3 mutants with a compromised ability to interact with OSKM reprogramming factors directly (Extended Data Fig. 9d) were deficient in reducing reprogramming efficiency of *Mbd3*$^{fl/-}$ somatic cells, supporting the notion that direct OSKM–Mbd3 interactions are important for inhibiting iPS cell formation (Fig. 5e).

We noted that a minimum 5-day exogenous transgene (doxycycline) induction was similarly required to obtain iPS cells from *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ cells (Extended Data Fig. 10a), and that the expression of Utx and Wdr5 (refs 4, 5)—OSK interacting partners that positively propel reprogramming to pluripotency—was also essential for iPS cell formation in Mbd3-depleted cells (Extended Data Fig. 10b, c). Collectively, these results establish a 'gas and brakes' paradigm. Whereas exogenous OSKM factors interact with multiple epigenetic complexes that de-repress pluripotency-promoting gene networks (such as Wdr5- or Utx-containing complexes), they also directly assemble with the Mbd3/NuRD repressor complex (Fig. 5f). As a result, Mbd3/NuRD is directly recruited to downstream OSKM target genes that are essential for propelling the reprogramming process, and potently counteracts their robust reactivation. In the absence of an Mbd3 inhibitory effect, OSKM interactions with pluripotency-promoting epigenetic regulators predominate functionally, and drive uninterrupted progression of direct reprogramming to pluripotency.

Concluding remarks

Here we show that the stochastic and asynchronized trajectory of direct reprograming by OSKM factors[3] can be coaxed to become nearly synchronized and deterministic with modified reprogramming approaches. We highlight a NuRD repressor complex component, Mbd3, which is normally depleted during early pre-implantation development, as a predominant barrier preventing epigenetic reversion of EpiSCs, primordial germ cells and somatic cells to ground-state pluripotency by defined signalling and transcriptional input. Several critical reprogramming factors directly interact and recruit Mbd3/NuRD complex, and thus form a highly potent negative regulatory complex that restrains pluripotency gene reactivation throughout the process. It will be of great interest to explore whether direct reprogramming in the absence of Mbd3 repression also improves the quality of reprogrammed cells and reduces the frequency of obtaining aberrantly reprogrammed iPS cells[24]. Finally, the deterministic reprogramming strategy reported here may allow the dissection of authentic molecular events accompanying a synchronized and non-saltatory progression pattern towards iPS cells, which is critical for deciphering the black box of reprogramming.

METHODS SUMMARY

Details of cell lines, plasmids, siRNAs and antibodies used, as well as descriptions of methods for reprogramming, immunofluorescence, immunoprecipitation, embryo micromanipulation, time-lapse microscopic imaging, bioinformatics, statistical analyses and mathematical modelling, are provided in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 April; accepted 23 August 2013.
Published online 18 September 2013.

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
2. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).
3. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143,** 508–525 (2010).
4. Ang, Y.-S. *et al.* Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145,** 183–197 (2011).
5. Mansour, A. A. *et al.* The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. *Nature* **488,** 409–413 (2012).
6. Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nature Biotechnol.* **28,** 521–526 (2010).
7. Hu, G. & Wade, P. A. NuRD and pluripotency: a complex balancing act. *Cell Stem Cell* **10,** 497–503 (2012).
8. Pawlak, M. & Jaenisch, R. *De novo* DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state. *Genes Dev.* **25,** 1035–1040 (2011).
9. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151,** 994–1004 (2012).
10. Kaji, K., Nichols, J. & Hendrich, B. Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development* **134,** 1123–1132 (2007).
11. Le Guezennec, X. *et al.* MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol. Cell. Biol.* **26,** 843–851 (2006).
12. Kaji, K. *et al.* The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nature Cell Biol.* **8,** 285–292 (2006).
13. Reynolds, N. *et al.* NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. *Cell Stem Cell* **10,** 583–594 (2012).
14. Leitch, H. G. *et al.* Embryonic germ cells from mice and rats exhibit properties consistent with a generic pluripotent ground state. *Development* **137,** 2279–2287 (2010).
15. Luo, M. *et al.* NuRD blocks reprogramming of mouse somatic cells into pluripotent stem cells. *Stem Cells* **31,** 1278–1286 (2013).
16. Onder, T. T. *et al.* Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483,** 598–602 (2012).
17. Hanna, J. *et al.* Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* **133,** 250–264 (2008).
18. Sommer, C. A. *et al.* Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem Cells* **27,** 543–549 (2009).
19. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).
20. Sridharan, R. *et al.* Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136,** 364–377 (2009).
21. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7,** 618–630 (2010).
22. Bolch, G., Greiner, S., de Meer, H. & Trivedi, K. S. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications* (John Wiley, 2006).
23. Duffy, K. R. *et al.* Activation-induced B cell fates are selected by intracellular stochastic competition. *Science* **335,** 338–341 (2012).
24. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465,** 175–181 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements J.H.H. is supported by a generous gift from I. and P. Mantoux; and grants from the Leona M. and Harry B. Helmsley Charitable Trust, ERC (StG-281906) grant, BIRAX initiative, Israel Science Foundation (BIKURA, ICORE and Regular programs), ICRF, Fritz Thyssen Stiftung, The Benoziyo Endowment fund, Alon Scholar Program, and the Clore research prize. I.A is supported by the HFSP Career Development Award, an ISF-Bikura and the ERC (StG-309788). A.A.M. is supported by a Weizmann Dean fellowship. We thank N. Barkai and her group, K. Saha, B. Hendrich, J. Nichols and A. Surani, for reagents and advice. We thank the Weizmann Institute management for providing critical financial and infrastructural support.

Author Contributions Y.R., A.Z., S.Ge., N.N. and J.H.H. conceived the idea for this project, designed and conducted experiments and wrote the manuscript. S.Ge. conducted protein biochemical analysis. A.Z. conducted numerical modelling analysis. O.G., L.W. and N.M. assisted in chromatin immunoprecipitation experiments. N.N. and A.Z. conducted bioinformatics analysis. Y.R. and A.Z. conducted live imaging experiments and analysis. S.V. engineered human stem cell lines. I.A., D.A.J., D.L.-A., S.Gi., D.A.-Z. and R.B.-G. assisted with ChIP-seq experiments. E.C., Z.S., Z.M. and A.T. conducted RRBS analysis. Y.R. and M.Z. conducted microinjections. Y.R. and A.A.M. conduced embryo staining. Y.R., S.Ge. and J.H.H. conducted reprogramming experiments with help from I.C., I.M., V.K., T.H. and D.B.

Author Information Chromatin immunoprecipitation data are available at the National Center for Biotechnology Information Gene Expression Omnibus database under the series accession number GSE49766. Microarray data are available at the National Center for Biotechnology Information Gene Expression Omnibus database under the series accession number GSE45352. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.H.H. (jacob.hanna@weizmann.ac.il) or N.N. (noa.novershtern@weizmann.ac.il).

METHODS

Mouse stem cell lines and cell culture. Reprogramming and maintenance of murine naive pluripotent cells were conducted in serum-free chemically defined N2B27-based media: 500 ml KO-DMEM (Invitrogen), 5 ml N2 supplement (Invitrogen; 17502048), 5 ml B27 supplement (Invitrogen; 17504044), 15–20% knockout serum replacement (Invitrogen; 10828), 1 mM glutamine (Invitrogen), 1% non-essential amino acids (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma), 1% penicillin–streptomycin (Invitrogen), 5 mg ml^{-1} BSA (Sigma). Naive conditions for murine iPS cells, ES cells and EGs included 5 μg recombinant human LIF (Peprotech). Throughout the study 2i was applied in reprogramming 48 h after OSKM induction: small-molecule inhibitors CHIR99021 (CH, 3 μM; Axon Medchem) and PD0325901 (PD, 1μM; Axon Medchem). Primed N2B27 media for murine cells (EpiSCs) contained 8 ng ml^{-1} recombinant human bFGF (Peprotech) and 20 ng ml^{-1} recombinant human activin (Peprotech). Stem-cell lines and mice deficient for Mbd3 and their derived ES lines were obtained as previously described[10,12]. For additional gene targeting of mouse pluripotent stem cell lines (Nanog-GFP reporter, pBRY-Mbd3 rescue constructs, Rosa26-CreER), 50 μg DNA of the targeting construct was linearized and electroporated into the indicated pluripotent cell lines, which were then subjected to selection with puromycin (1 μg ml^{-1}). After 10 days of antibiotic selection, drug-resistant or GFP^+ clones were analysed for correct targeting by PCR or Southern blot analysis. *Mbd3*$^{+/-}$ male and female mice were mated and E3.5 blastocysts were collected and explanted for ES cell derivation in defined mouse 2i/LIF conditions on gelatin/MEF-coated plates. NGFP1-$Mbd3^{KD}$ was established by infection and sub-cloning of secondary NGFP1 iPS cell line with a shRNA pLKO-Tet-On vector (Addgene) as previously described[2]. Mbd3 target sequences selected for the latter strategy were CTAAGTGGATTGAGTGCCTTT and GCGCTATGATTCTTCCAACCA. Mycoplasma detection tests are conducted weekly to ensure exclusion of any contaminated cells.

Epigenetic reversion of mouse primed epiblast cells. Male naive V6.5 (*Mbd3*$^{+/+}$) and Nanog–GFP ES cells[15] maintained in 2i/LIF conditions were injected into BDF2 blastocysts. Chimaeric embryos were dissected at day E6.5 and explanted on gelatin/vitronectin-coated plates in N2B27 bFGF/activin conditions supplemented with 1 μg ml^{-1} puromycin, allowing the isolation of Nanog–GFP EpiSCs. For epigenetic reversion of murine EpiSCs to naive pluripotency, cells were passaged into N2B27 2i/LIF conditions on vitronectin (1 μg ml^{-1}) and gelatin (0.2%)-coated plates (without overexpression of exogenous reprogramming factors). When epigenetic reversion assay involved single-cell plating, EpiSC growth medium was supplemented with ROCK inhibitor (Y-27632) for 24 h before trypsinization. siRNAs (ON-TARGETplus SMARTpool) and the control siRNA (ON-TARGETplus non-targeting pool D-001810-10-05) were obtained from Dharmacon. 10 nM siRNA or control was used for each transfection with RNAiMAX (Invitrogen). For EG derivation experiments, Oct4–GFP^+ primordial germ cells were sorted from E8.5 dissected chimaeric embryos and plated in N2B27 15% KSR, LIF (20 ng ml^{-1})/SCF (10 ng ml^{-1})/bFGF (8 ng ml^{-1}) medium and 2i (supplemented 48 h later). Nuclear mCherry labelling of EpiSCs and their derived primordial germ cells was used to allow calculating plating efficiency and reprogramming efficiency (reprogramming efficiency % = Oct4 or Nanog–GFP^+ clones/$mCherry^+$ clones).

Reprogramming of mouse somatic cells and cell infection. Virus-containing supernatants of the different reprogramming viruses: STEMCCA-OKSM polycistronic vector (doxycycline-inducible and constitutive expression)[5], STEMCCA-OKS polycistronic vector (doxycycline-inducible and constitutive expression), FUW-tetO-lox-KLF4, FUW-tetO-lox-OCT4 and FUW-tetO-lox-SOX2, FUW-tetO-Klf4, FUW-tetO-Oct4, FUW-tetO-Sox2, FUW-tetO-c-Myc, FUW-Oct4-2A-Sox2, FUW-Oct4-2A-Klf4, FUW-tetO-lox-SOX2, pMXs-OCT4, pMXs-SOX2, pMXs-KLF4, pMXs-MYC was supplemented with the FUW-M2rtTA virus (when necessary) and an equal volume of fresh culture medium for infection. Mouse fibroblast and other somatic cells types were isolated and single-cell sorted from secondary transgenic reprogrammable chimaeras[5,25]. iPS cells were reprogrammed using mouse naive ES cell medium 2i/LIF plus doxycycline (1 μg ml^{-1}) (without 2i in the first 48 h) (under physiological 5% pO_2 conditions for fibroblast cells). *Mbd3*$^{-/-}$ MEFs (but not pluripotent cells) experience accelerated senescence and proliferation capacity loss[26], and thus *Mbd3*$^{-/-}$ somatic cells were reprogrammed by applying tamoxifen on *Mbd3*$^{fl/-}$ cells only after 48 h of OSKM induction. Similarly, for acute knockdown of Mbd3 in somatic cells with *Mbd3* siRNAs in an attempt to boost reprogramming, transfections were conducted at least after 48 h of OSKM induction. Alternatively, somatic cells with hypomorphic expression (rather than complete ablation) of Mbd3 do not demonstrate proliferation defects or accelerated senescence due to the residual Mbd3 expression levels, yet they retain sufficiently reduced Mbd3 levels that allow deterministic synchronized reprogramming by OSKM (Fig. 2b). Thus, the following systems were preferably and predominantly used throughout this study: *Mbd3*$^{fl/-}$ or WIBR3-$MBD3^{mut}$ genetic backgrounds. Notably, in our reprogramming conditions, single-cell plating of MEFs yielded approximately 70% survival efficiency (with or without doxycycline). Thus, for live imaging, upon plating 150 MEFs per well we observed formation of 100–120 colonies that were tracked in *Mbd3*$^{fl/-}$ samples. Notably, constitutive mCherry allowed one to control for survival after plating to obtain accurate and unbiased reprogramming efficiencies (reprogramming efficiency % = Oct4 or Nanog–GFP^+ clones (cells)/mCherry + clones (cells)). Equivalent reprogramming efficiencies were obtained on mouse irradiated feeder cells or gelatin-, Matrigel- or gelatin/vitronectin-coated plates (data not shown). Reprogramming on irradiated DR4 MEFs was preferably used for live imaging and single-cell reprogramming experiments to enhance cell survival and adherence. Mbd3 Stealth siRNA mix that includes MSS-237238, MSS-275658 and MSS-275659 components (Invitrogen), and Chd4 Stealth siRNA mix that includes MSS-200894, MSS-200895 and MSS-200896 (Invitrogen), were used for efficient knockdown in mouse cells. Transfections were conducted with RNAiMAX (Invitrogen) according to the manufacturer's instructions.

DNA plasmids and TALEN gene editing. The following lentiviral and mammalian constitutive overexpression vectors were used in somatic and pluripotent cells: pBRY-Mbd3-IRES-zeocin or pBRY-Mbd3-IRES-puromycin. Constitutively expressed lentiviruses FUW-Mbd2 and FUW-Mbd3 were generated by insert cloning into EcoRI sites of FUW lentiviral vector to generate constitutive expression following viral transduction and stable integration in somatic or iPS cell/ES cell lines. Flag–Mbd3 mutations and deletions were done by PCR with Q5 DNA polymerase (NEB). TALEN-expressing plasmids were designed and cloned using GoldenGate TALEN kit 2.0 purchased from Addgene[27] according to the published protocol. For targeting G we have used NN-type repeat. Donor construct was made with DNA fragments amplified from WIBR3 human ES cell genomic DNA. 10^7 ES cells were electroporated with 30 μg of donor plasmid and 10 μg each of the TALEN-expressing plasmids and grown in the presence of G418 (75 μg ml^{-1}) and ganciclovir (1 μM). Resistant clones were isolated and genomic DNA was extracted for Southern blot and PCR analysis. The WIBR3-$MBD3^{mut}$ subcloned cell line was correctly targeted on both endogenous alleles as confirmed by 5′ and 3′ Southern blot strategies (not shown), and was subsequently used for generating a doxycycline-inducible secondary reprogramming system[28] (without excision of PGK-neo-pA cassette, to maintain interference with *MBD3* expression). For generating OCT4–GFP reporter subcloned cell lines, 10^7 $MBD3^{WT}$ and $MBD3^{mut}$ cells were electroporated with 30 μg of previously described OCT4-GFP-2A-PURO knock-in donor plasmid[29] (provided by R. Jaenisch through Addgene) and 10 μg each of the TALEN-expressing plasmids and grown in the presence of puromycin (0.4 μg ml^{-1}). Resistant clones were isolated and genomic DNA was extracted for Southern blot and/or PCR analysis.

Reprogramming of human somatic cells and cell infection. Reprogramming was conducted at 5% pO_2 in doxycycline (1–2 μg ml^{-1}) supplemented conditions. In the first 48 h, cells were incubated in conventional human ES medium (hESM; see below). Afterwards, cells were transferred until day 7–8 into modified hESM with ROCKi (Y-27632; 5 μM final concentration) and , more optimally, with 2i/LIF containing supplement. After 8 days doxycycline was withdrawn and cells were expanded as stable iPS cell lines. MBD3 Stealth siRNAs that include HSS147580 and HSS147581 components (catalogue number 1299003) were used for efficient MBD3 knockdown in human cells. Transfections were conducted with RNAiMAX (Invitrogen) according to the manufacturer's instructions. Conventional human ES conditions (hESM) include: 475 ml knockout DMEM (Invitrogen 10829), 15–20% KSR (Invitrogen), 8 ng ml^{-1} recombinant bFGF (Peprotech) and 1 ng ml^{-1} recombinant TGFβ1 (Peprotech), 1 mM glutamine (Invitrogen), 1% nonessential amino acids (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma), 1% penicillin–streptomycin (Invitrogen). $Mbd3^{mut}$ human ES cells were generated by genetic engineering with TALE nuclease effectors and differentiated *in vitro* into fibroblasts. Next, $MBD3^{WT}$ and $MBD3^{mut}$ iPS cells carrying doxycycline-inducible OKSM transgene were generated and labelled with constitutively expressed mCherry and targeted with an *OCT4-GFP* knock-in allele. *In vitro* differentiated fibroblasts from human ES cells/iPS cells were generated as previously described[28]. The Weizmann Institute ESCRO committee approved human cell line studies.

Immunofluorescence staining of pre- and post-implantation embryos. Immunostaining was performed as described previously with modifications[30]. Embryos were collected from the oviducts and uteri of hormone-primed B6D2F1 6-week-old females mated with C57BL/6 males. At least 15 embryos at each stage were analysed in total, and representative images were shown. Identical simultaneous staining and imaging analysis was conducted for all embryonic stages analysed. Briefly, embryos were transferred to a watch-glass dish (Genenet), fixed for 15 min in 4% PFA in phosphate buffer (PB), rinsed three times in PBS containing 3 mg ml^{-1} PVP, permeabilized in PBS/PVP with 0.5% Triton X-100 for 30 min, and blocked in blocking solution (2% normal donkey serum, 0.01% Tween in PBS) for 1 h. Embryos were then incubated overnight at 4 °C in primary antibodies diluted in blocking solution, washed three times in blocking solution for 15 min each, incubated with secondary antibodies for 1 h at room temperature, counterstained with DAPI for 15 min, washed twice in PBS, and placed in 96-well glass-bottom

plates for confocal imaging. Post-implantation embryos in the maternal decidua were fixed in 4% PFA/PB overnight at 4 °C, washed three times in PBS for 30 min each, dehydrated and embedded in paraffin using standard procedure. Embryonic paraffin sections (5–7 μm) were rehydrated, treated with antigen retrieval, rinsed in PBS, permeabilized in 0.1% Triton/PBS for 10 min, rinsed in PBT (0.02% Tween/PBS), and blocked in blocking solution (5% normal donkey serum, 0.05% BSA, in PBT) for 1 h. Slides were then incubated in the appropriate primary and secondary antibodies diluted in blocking solution as described above, and processed as described previously[5]. The following antibodies were used: mouse anti-Oct4 (1:100, C-10; Santa Cruz SC-5279), goat anti-Mbd3 (1:50, C-18; Santa Cruz SC-9402).

Immunoprecipitation and immunoblotting analyses. HEK293T cells were transfected with each cDNA clone in an expression vector using jetPEI (Polyplus transfection) and were lysed 48 h later in lysis buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% Triton, 0.1% NP40 and 1.5 mM EDTA). The following plasmids were used for transfections in different combinations: pCaggs-Mbd3, FUW-OCT4, FUW-KLF4, FUW-SOX2, FUW-c-MYC, FUW-Nanog, pCaggs-Flag-Mbd3, pMSCV-Flag-OCT4, pMSCV-Flag-SOX2, pMSCV-Flag-KLF4, pCaggs-Flag-c-MYC, pCaggs-Flag-Nanog, pcDNA3.1-Flag-HDAC1 (obtained through addgene). 30 μl of anti-FlagM2 Magnetic beads (Sigma) were incubated for 6 h in cell lysate fractions, for IgG control 6 μg of IgG and 50 μl of protein-G Dynabeads (Invitrogen) were added to the cell lysate for 6 h. Both fractions (the anti-Flag and anti-IgG) were loaded on Invitrogen magnetic separator and the beads were washed six times with lysis buffer. The binding proteins were eluted with 0.5 μg ml^{-1} of ×3Flag peptide (Sigma) buffer for the anti-FlagM2 beads or by boiling with sample buffer and analysed by SDS–polyacrylamide gel electrophoresis and immunoblotting. The immunoblot analyses were performed using the following primary antibodies: anti-Flag (clone M2, F3165, Sigma), anti-Mbd3 (A300-258A, Bethyl), anti-Nanog (A300-397A, Bethyl), anti-OCT4 (sc-9081, H134, Santa Cruz), anti-KLF4 (sc20691, H180, Santa Cruz), anti-SOX2 (2748s, Cell Signaling) and anti-c-Myc (9402s, Cell Signaling).

Mouse embryo micromanipulation and teratoma formation. Pluripotent stem cells (ES cells or iPS cells) were injected into BDF2 diploid blastocysts. Microinjection into blastocysts placed in M16 medium under mineral oil was done by a flat-tip microinjection pipette. A controlled number of 10–12 cells were injected into the blastocyst cavity. After injection, blastocysts were returned to KSOM media (Invitrogen) and placed at 37 °C until transferred to recipient females. Ten to fifteen injected blastocysts were transferred to each uterine horn of 2.5 days post coitum pseudo-pregnant females. Embryos were recovered for analysis at different time points during development or allowed to develop into full term. Determining germline transmission was performed by mating chimaeric animals with C57BL/6 females, and continuous checking for agouti-coloured pups. For teratoma formation and analysis, ES cells and iPS cells were collected by trypsinization before injection. Cells were injected subcutaneously into female 4–8-week-old NOD-SCID mice (Jackson laboratories). Tumours generally developed within 4–6 weeks and animals were killed before tumour size exceeded 1.5 cm in diameter. All animal studies were conducted according to the guideline and following approval by the Weizmann Institute IACUC (approval 00960212-3). We did not exclude animals from our analysis, and did not apply randomization by blinding.

RT–PCR analysis. Total RNA was isolated using the RNeasy kit (Qiagen). 3 μg of total RNA was treated with DNase I to remove potential contamination of genomic DNA using a DNA Free RNA kit (Zymo Research). 1 μg of DNase-I-treated RNA was reverse transcribed using a First Strand Synthesis kit (Invitrogen) and ultimately re-suspended in 100 μl of water. Quantitative PCR analysis was performed in triplicate using 1/50 of the reverse transcription reaction on Viia7 platform (Applied Biosystems). Error bars indicate standard deviation of technical triplicate for each measurement. For single-cell RT–PCR analysis, single cells from different samples were single cell plated, and Ambion Single Cell-to-CT kit was used for sample processing according to manufacturer instructions. TaqMan probe based chemistry and TaqMan Real-Time PCR master mix were used on Viia7 platform for gene expression detection. The following TaqMan (Invitrogen) probes were used: Sall4 Mm00453037_s1, Utf1 Mm00447703_g1; Sox2 (endogenous mouse allele specific) Mm03053810_s1; Nanog Mm02384862_g1; Gapdh Mm99999915_g1. Undetected expression indicates lack of amplification even after 50 amplification cycles (red boxes).

Immunocytochemistry and FACS analysis. Cells were fixed in 4% paraformaldehyde in PBS and immunostained according to standard protocols using the following primary antibodies: mouse anti-TRA-1-60 (1:500, Abcam; ab16288), mouse anti-TRA-1-81 (1:500, Abcam; ab16289), mouse anti-SSEA1 (1:100, Abcam; MC480 ab16285), mouse anti-SSEA4 (1:50, Abcam; MC813 ab16287), rat anti-SSEA3 (1:50, Abcam; MC631 ab16286), rabbit anti-Nanog (1:400, Bethyl; A300-397A), rabbit anti-Oct3/4 (1:400, Santa Cruz; H134 SC9081), mouse anti-Oct4 (1:200, Santa Cruz; C-10 SC5279), rabbit anti-Sox2 (1:500, Millipore; AB5603). Appropriate Alexa Fluor dye-conjugated secondary antibodies (1:200, Jackson ImmunoResearch) were used. FACS data were collected on BD FACS ARIA III and analysed with Flowjo software.

Microscopy image acquisition and analysis. Secondary OKSM inducible *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ MEFs carrying the Oct4–GFP pluripotency reporter and constitutively expressed nuclear mCherry marker were plated in 24-well plates at low densities (∼150 cells per well) and imaged using AxioObserver Z1 (Zeiss) in 5% O_2, 5% CO_2, 37 °C controlled conditions. Plates were taken out at day 3–4 for media replacement (but without passaging/splitting) and put back for the automated live imaging stage. Full-well mosaic images were taken every 12 h for 6 days at ×50 magnification, including phase contrast and two fluorescent wavelength images. In-house automated segmentation protocol was developed and implemented in Matlab to analyse time-lapse measurements of full-well mosaics with fluorescent mCherry and Oct4–GFP markers.

The challenge in this protocol was to implement fast segmentation of unknown number of colonies in 10^8 pixels mosaic image. The protocol includes the following main steps. Adaptive detection: erasing the plate margins with circular filter. Defining detection threshold using median with offset (10% of the dynamic range), and creating a binary image of detected pixels. These steps were carried out separately for each time point and each fluorescent wavelength. Complexity reduction: for this task we applied a morphological filter to isolate mCherry colonies using median sliding filter (60 μm×60 μm)[31]. This filter retains only dense colonies, erasing noise and single isolated cells (a single nucleus is approximately 6 μm×6 μm), this step is crucial for reducing the dimension of the clustering task. Colony segmentation: the segmentation was done using moving average filter (low-pass filter) (60 μm×60 μm)[31] to merge adjacent colony fragments into large connected colonies and then apply connected components clustering, labelling connected objects using 8-connected neighbourhood. Colony feature extraction: extracting the features of each mCherry colony including area, bounding box and centroid. By overlaying mCherry colony segmentation on the GFP binary image (detected pixels) we extract for each colony the GFP^+ indicator (0/1) and the fraction of GFP^+ and $mCherry^+$ pixels out of all $mCherry^+$ pixels.

This segmentation protocol was run over time-lapse mosaics collecting information on colony formation dynamics, colony GFP^+ dynamics and ratios of offspring Oct4–GFP^+ cells. Colony and reprogramming dynamics features were then statistically analysed using Matlab program, including estimation of the cumulative distribution and density function. In addition, videos characterizing the process dynamics were produced using customized Matlab program. The above algorithm was validated by artificial input matrix and by ES mosaic image collection. In addition, robustness of detection threshold and filter size were measured with varying parameters (data not shown).

Chromatin immunoprecipitation and sequencing library preparation. Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) was measured for the proteins H3K4me3, H3K27me3, H3K27ac and Mbd3 in four different time points throughout reprogramming: 0 (MEF), 4 days, 8 days, subcloned iPS or ES lines. The binding of each protein was measured in both *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ conditions, as well as in *Mbd3*$^{-/-}$ ES cells. Oct4 was measured in all the above conditions, excluding 8 days. Approximately 40×10^6 cells were crosslinked in formaldehyde (1% final concentration, 10 min at room temperature), and then quenched with glycine (5 min at room temperature). Fixed cells were lysed in 50 mM HEPES KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 alternative, 0.25% Triton supplemented with protease inhibitor at 4 °C (Roche, 04693159001), centrifuged at 950*g* for 10 min and re-suspended in 0.2% SDS, 10 mM EDTA, 140 mM NaCl and 10 mM Tris-HCl. Cells were then fragmented with a Branson Sonifier (model S-450D) at −4 °C to size ranges between 200 and 800 bp, and precipitated by centrifugation. 10 μg of each antibody was pre-bound by incubating with Protein-G Dynabeads (Invitrogen 100-07D) in blocking buffer (PBS supplemented with 0.5% TWEEN and 0.5% BSA) for 2 h at room temperature. Washed beads were added to the chromatin lysate, and then incubated overnight. Samples were washed five times with RIPA buffer, twice with RIPA buffer supplemented with 500 mM NaCl, twice with LiCl buffer (10 mM TE, 250mM LiCl, 0.5% NP-40, 0.5% DOC), once with TE (10Mm Tris-HCl pH 8.0, 1mM EDTA), and then eluted in 0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris HCl pH 8.0 at 65 °C. Eluate was incubated in 65 °C for 8 h, and then treated sequentially with RNaseA (Roche, 11119915001) for 30 min and proteinase K (NEB, P8102S) for 2 h. DNA was purified with The Agencourt AMPure XP system (Beckman Coulter Genomics, A63881). Libraries of cross-reversed ChIP DNA samples were prepared according to a modified version of the Illumina Genomic DNA protocol, as described previously[32]. Briefly, ChIP DNA was ligated to Illumina adaptors and subjected to 14 cycles of PCR amplification. Amplified products between 200 and 800 bp were purified on a 2% agarose gel. Roughly 5 pmol of DNA library was then applied to each lane of the flow cell and sequenced on Illumina Hiseq2000 sequencer according to standard Illumina protocols. The

following antibodies were used for chromatin immunoprecipitation experiments: control IgG (ChIP grade, ab46540, Abcam), anti-histone H3 trimethyl K4 (ChIP grade, ab8580, Abcam), anti-histone H3 acetyl K27 (ChIP grade, ab4729, Abcam), anti-histone H3 trimethyl K27 (ChIP grade, 07-449, Millipore), anti-Oct4 (sc5729 (C-10), Santa Cruz), anti-Chd4 (ChIP Grade, ab70469, Abcam). For Mbd3 chip 1:1 antibody mix was used: anti-Mbd3 (Bethyl laboratories A302-528\9A) and anti-Mbd3 (ab16057, Abcam). Chromatin immunoprecipitation data are available at the National Center for Biotechnology Information Gene Expression Omnibus database under the series accession number GSE49766.

Alignment and peak detection. We used bowtie software version 0.12.5 to align reads to mouse mm9 reference genome (UCSC, July 2007). We only considered reads that were uniquely aligned to the genome with up to a single mismatch, taking the single best match of each read. We identified enriched intervals of H3K4me3, H3K27me3, H3K27ac, Mbd3 and Oct4 using MACS version 1.4.1. We used sequencing of whole-cell extract as control to define a background model. Duplicate reads aligned to the exact same location are excluded by MACS default configuration. Enriched intervals were mapped to genes if they overlapped a single kilobase symmetric interval around their transcription start sites (TSS; taken from RefSeq known gene table in UCSC genome browser). ChIP-seq data on wild-type samples were highly compatible with those provided in previous publications[19,20].

Histone mark profiles. Histone mark profiles were calculated using in-house script. Briefly, this script generates a matrix of read densities in given genomic intervals. In this case, the profiles of all 29,952 Entrez genes (mm9, taken from UCSC known gene tables) were calculated between 1 kb upstream to TSS and TES (transcription end site). These read densities were then converted to z-score by normalizing each position by the mean and standard deviation of the sample noise ($\hat{X}_j = \frac{X_j - \mu_{\text{Noise}}}{\sigma_{\text{Noise}}}$), where X_j corresponds to read density, and μ_{noise} and σ_{noise} are the estimated noise mean and standard deviation, respectively . Noise parameters were estimated for each sample from 6×10^7 random base pairs across the genome. Finally, to present aligned profiles, the z-score profile of each gene was binned to 20 bins upstream to TSS and another 100 quantiles between TSS to TES. The value of each bin or quantile was selected to be the maximum value within that interval. In the histone mark distribution analysis and in the correlation and clustering of histone marks (Extended Data Fig. 5a), each gene and each histone mark is represented with the maximal z-score measured in the profile of that gene, where the profiles were calculated as described above. Clustering of histone marks was carried out on concatenated vectors that include all marks for every gene in tandem.

Annotation enrichment analysis. Mbd3 target genes were tested for enrichment of functional gene sets taken from Gene Ontology (GO, http://www.geneontology.org). Protein–DNA binding annotations were taken from various publications[33–36]. Enrichment P values were calculated using Fisher exact test[37] and corrected for multiple hypotheses using false discovery rate (FDR) threshold of 0.0001%.

Gene expression data acquisition. Total RNA was isolated from indicated cell lines. The concentration of RNA was quantified and subjected to quality control on Agilent Bioanalyzer. 250 ng of RNA was simultaneously processed from each sample. cDNA was fragmented, labelled, and hybridized to Affymetrix Mouse Gene 1.0 ST GeneChip (Affymetrix), which contain 35,557 probes. Transcript levels were processed from image files using RMA method[38], which corrects for non-biological sample variation using quantile normalization, implemented by the Affymetrix 'Expression Console' software. Microarray data are available at the National Center for Biotechnology Information Gene Expression Omnibus database under the series accession number GSE45352.

Gene expression analysis. Probes were mapped to Entrez Gene IDs and further filtered to include IDs that have at least one call higher than 32 ($=2^5$), resulting in 16,620 gene IDs. For gene expression analysis, we used Matlab version R2011b. Gene signatures differentially expressed between MEF samples (*Mbd3*$^{+/+}$, *Mbd3*$^{fl/-}$, *Mbd3*$^{-/-}$ MEF samples) and ES samples (ES V6.5, *Mbd3*$^{-/-}$ ES, Mbd3$^{fl/-}$ iPS and *Mbd3*$^{+/+}$ iPS) were characterized using a two-sample t-test and corrected for multiple hypotheses using false discovery rate (FDR)[39]. Differentially expressed gene signatures include genes that are under FDR threshold of 5%, as well as above fourfold change, resulting in 1,323 genes. Sample clustering with all 16,620 genes (Fig. 3a) was done with hierarchical clustering using Spearman correlation as a distance metric and average linkage. Single gene progression in reprogramming (Extended Data Fig. 8a) were quantified using the following transformation

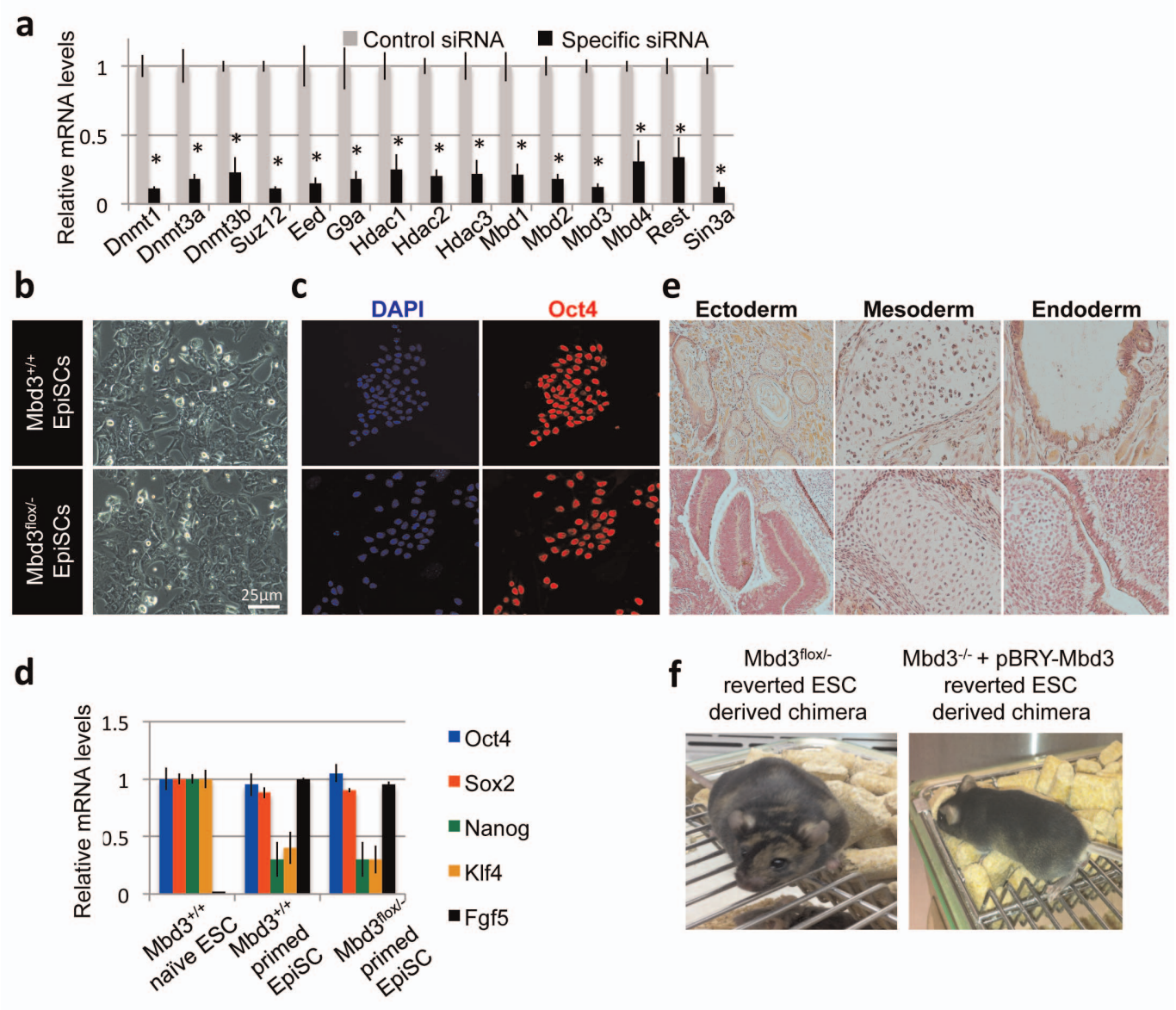
$$\hat{X}_j(t) = \max\left(\frac{X_j(t) - X_j(\text{MEF } Mbd3^{+/+})}{\bar{X}_j(\text{iPS}) - \bar{X}_j(\text{MEF})}, 0\right)$$

where, $X_j(t)$ denotes gene j expression value at time t (for example, X_j(4 days) or X_j(MEF)) and $\bar{X}_j$(iPS), $\bar{X}_j$(MEF) denote the averaged expression value for iPS and MEF samples, respectively. The above transformation represents a distance from MEF expression values (set to 0) towards iPS values (set to 1), where genes whose expression changes towards (up/down-regulating) their iPS value show $\hat{X}_j(t) > 0$. Distribution of gene expression fold-change, relative to MEF, is presented by box plots (Fig. 5c and Extended Data Fig. 8c). Distribution difference significance was calculated with a paired samples t-test.

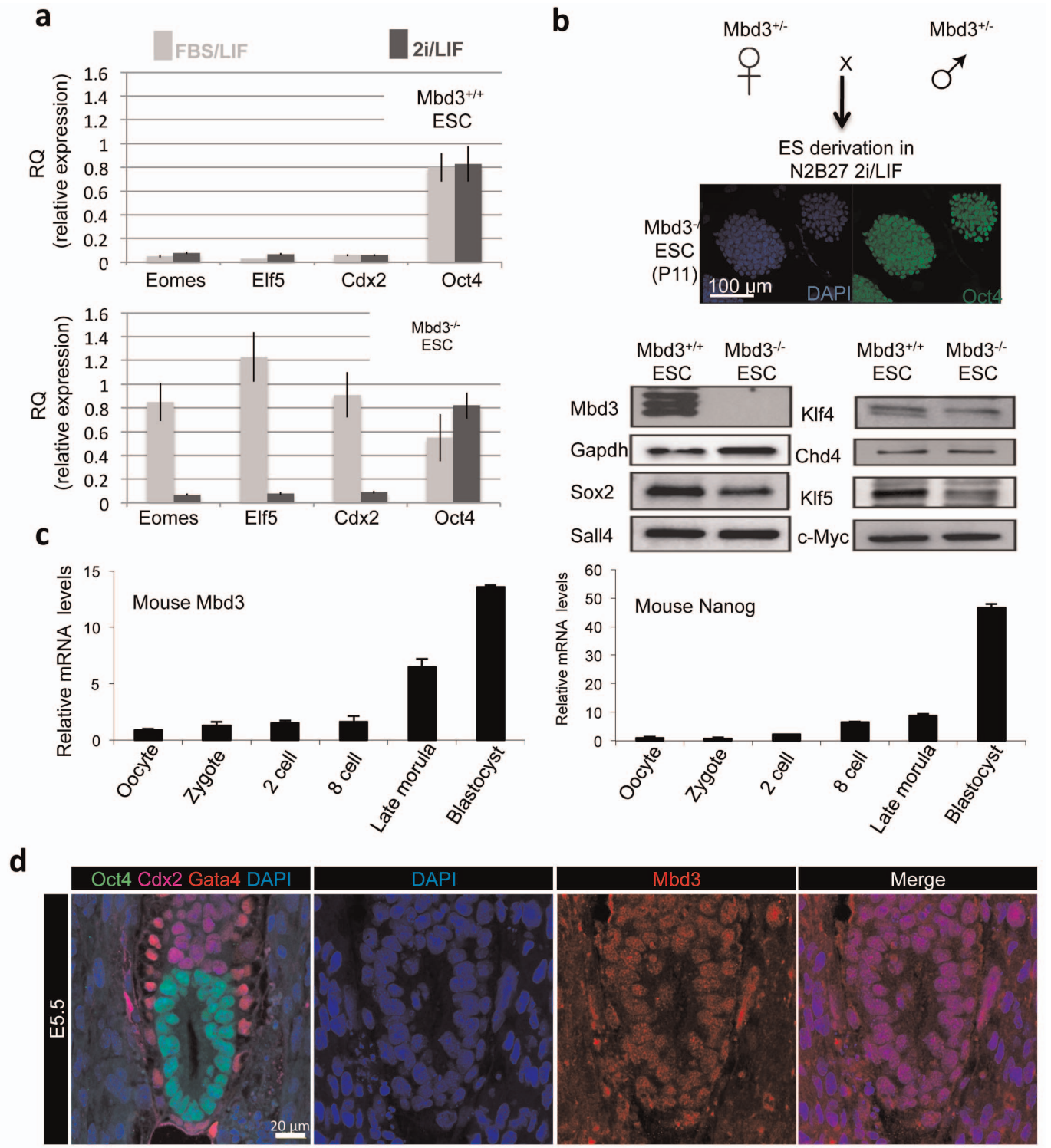
Preparation and analysis of reduced representation bisulphite sequencing libraries. RRBS libraries were generated as described previously with slight modifications[40]. Briefly, DNA was isolated from snap-frozen cell pellets using the Quick-gDNA mini prep kit (Zymo). Isolated DNA was then subjected to MspI digestion (NEB), followed by end repair using T4 PNK/T4 DNA polymerase mix (NEB), A-tailing using Klenow fragment (3′→5′ exo-) (NEB), size selection for fragments shorter than 500 bp using SPRI beads (Beckman Coulter) and ligation into a plasmid using quick T4 DNA ligase (NEB). Plasmids were treated with sodium bisulphite using the EZ DNA Methylation-Gold kit (Zymo) and the product was PCR amplified using GoTaq Hot Start DNA polymerase (Promega). The PCR products were A-tailed using Klenow fragment, ligated to indexed Illumina adapters using quick T4 DNA ligase and PCR amplified using GoTaq DNA polymerase. The libraries were then size-selected to 200–500 bp by extended gel electrophoresis using NuSieve 3:1 agarose (Lonza) and gel extraction (Qiagen). Libraries were pooled and sequenced on an Illumina HiSeq 2500 system. The sequencing reads were aligned to the Mouse Genome Build 37 (mm9) using Bismark. Methylation levels were calculated and averaged only for CpGs that were covered by 5 or more distinct sequencing reads across all libraries.

Numerical modelling analysis. Clonal reprogramming measurements and some data sets used were previously published[2]. NGFP1-*Mbd3*^{KD} and NGFP1-*Mbd3*$^{+/+}$ (control) reprogramming distributions were fitted to multiple modelling schemes, comparing reprogramming variability to known deterministic and stochastic dynamic models. This includes: (1) fitting to a deterministic step function, where the deterministic transition time was estimated by the fitting procedure. (2) Fitting to Gaussian distribution and calculating the mean, variance and coefficient of variation (CV = s.d./mean) for each sample. (3) Fitting to inverse Gaussian distribution, according to the 'first passage time of Brownian motion' model, and calculating the dynamic variability as the ratio of the Brownian motion standard deviation divided by the Brownian motion drift parameter. (4) Fitting to multi-step Markov chain (phase-type) model that infers possible structure for the reprogramming process. For this purpose, we constructed a nested fitting procedure for the fitting of *Mbd3*^{KD} and control *Mbd3*$^{+/+}$ dynamics to multiple models with 1 to 5 exponential transitions. We also compared reprogramming dynamics to a cell-cycle model, where we estimated the number of generations according to the reprogramming duration, and fit the cell cycle time distribution to the observed reprogramming latency. All model fittings were implemented by Matlab program performing nonlinear regression fitting with adjusted R^2 statistic and/or by using maximum likelihood estimator. For more detailed information see Supplementary Information numerical modelling analysis section.

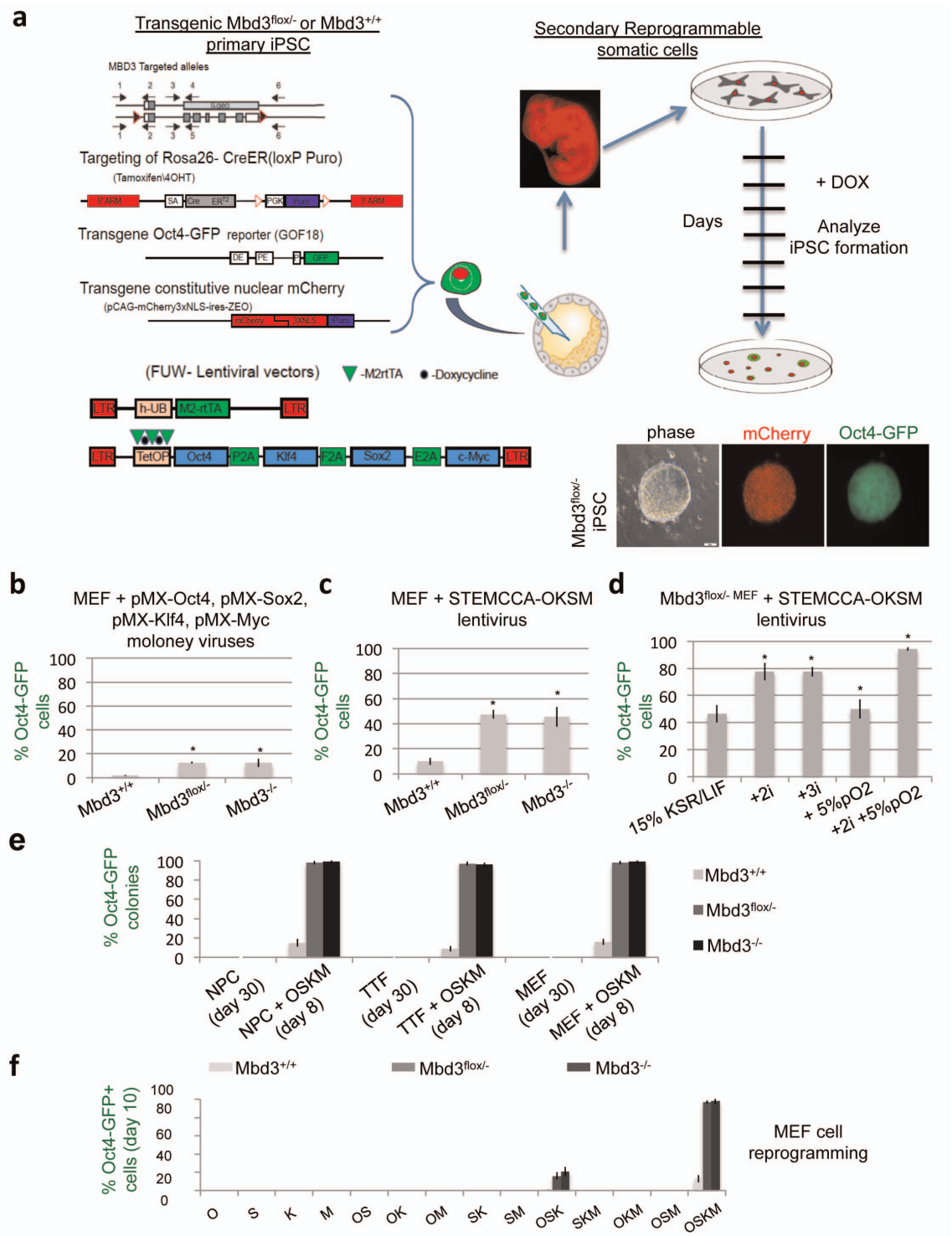
25. Hanna, J. *et al.* Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell* **4,** 513–524 (2009).
26. Reynolds, N. *et al.* NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression. *EMBO J.* **31,** 593–605 (2011).
27. Bedell, V. M., Wang, Y., Campbell, J. M. & Poshusta, T. L. *In vivo* genome editing using a high-efficiency TALEN system. *Nature* **491,** 114–118 (2012).
28. Hockemeyer, D. *et al.* A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell* **3,** 346–353 (2008).
29. Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nature Biotechnol.* **29,** 731–734 (2011).
30. Silva, J. *et al.* Nanog is the gateway to the pluripotent ground state. *Cell* **138,** 722–737 (2009).
31. Arce, G. R. *Nonlinear Signal Processing: A Statistical Approach* (Google Books, 2005).
32. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of *in vivo* protein-DNA interactions and epigenomic states. *Nature Protocols* **8,** 539 (2013).
33. Boyer, L. A., Gifford, D. K., Melton, D. A., Jaenisch, R. & Young, R. A. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
34. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
35. Kim, J. *et al.* An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–1061 (2008).
36. Loh, Y.-H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* **38,** 431–440 (2006).
37. Fisher, S., Genetiker, S., Fisher, R. A. & Genetician, S. *Statistical Methods for Research Workers* (Oliver and Boyd, 1970).
38. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
40. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484,** 339–344 (2012).



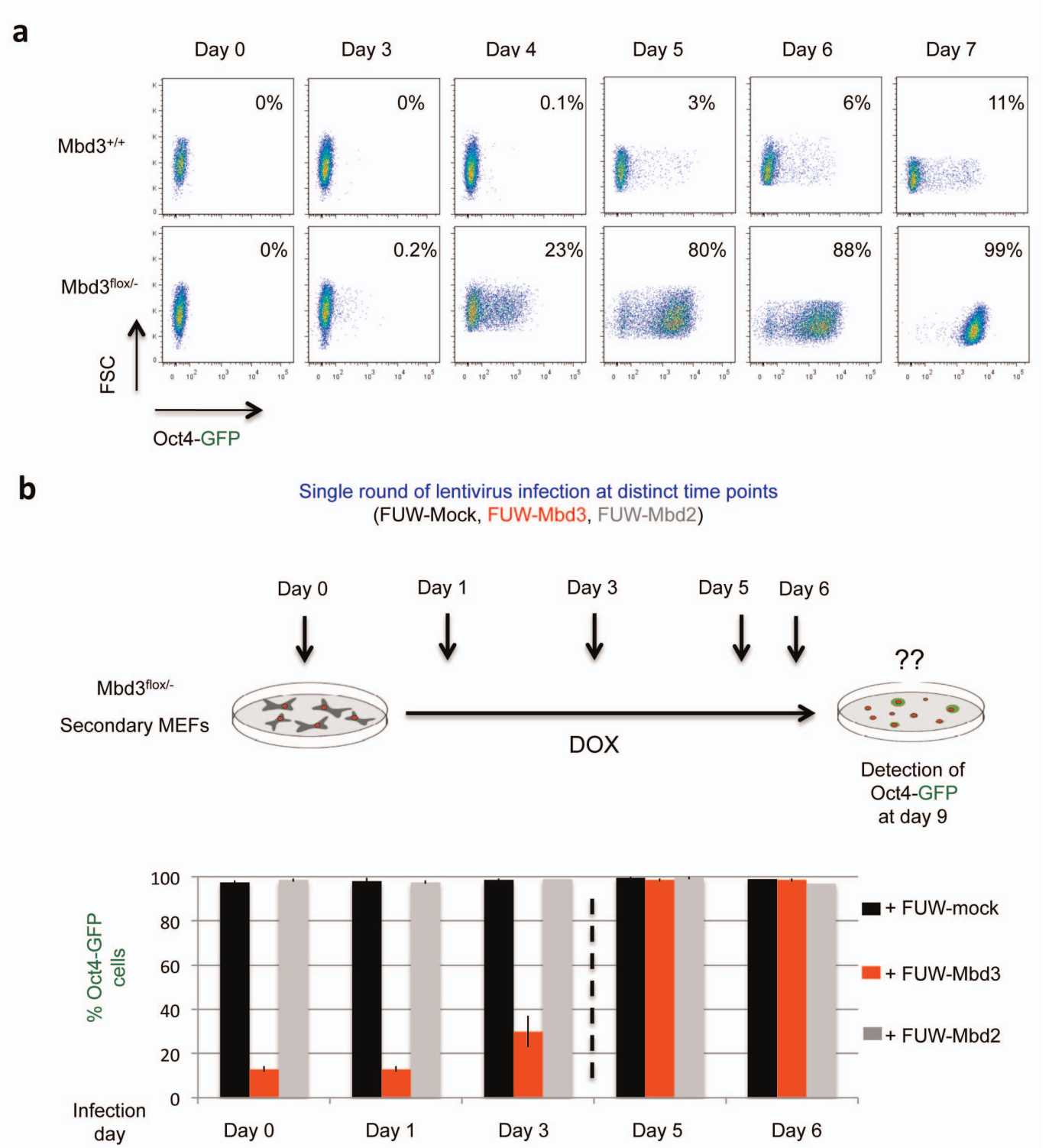
Extended Data Figure 1 | Knockdown screen for epigenetic repressors in EpiSCs. **a**, Knockdown efficiency of the indicated siRNA pools in EpiSCs measured by qRT–PCR. Expression values for each gene were normalized to those measured in control siRNA. Error bars indicated s.d. from average. Asterisks indicate *t*-test *P* value <0.05. **b**, Phase images of $Mbd3^{+/+}$ and $Mbd3^{fl/-}$ EpiSC lines in this study. **c**, Oct4 immunostaining on EpiSC lines. **d**, RT–PCR expression level validation for pluripotency genes in naive V6.5 ES cells and primed $Mbd3^{+/+}$ and $Mbd3^{fl/-}$ EpiSCs. In comparison to naive ES cells, primed EpiSCs downregulate naive pluripotency markers Nanog and Klf4, and upregulate FGF5 transcription ($n = 3$). **e**, EpiSC lines were pluripotent as evident by their ability to form mature differentiated teratomas. **f**, Representative agouti-coloured chimaeras obtained from reverted EpiSCs after Mbd3 depletion.



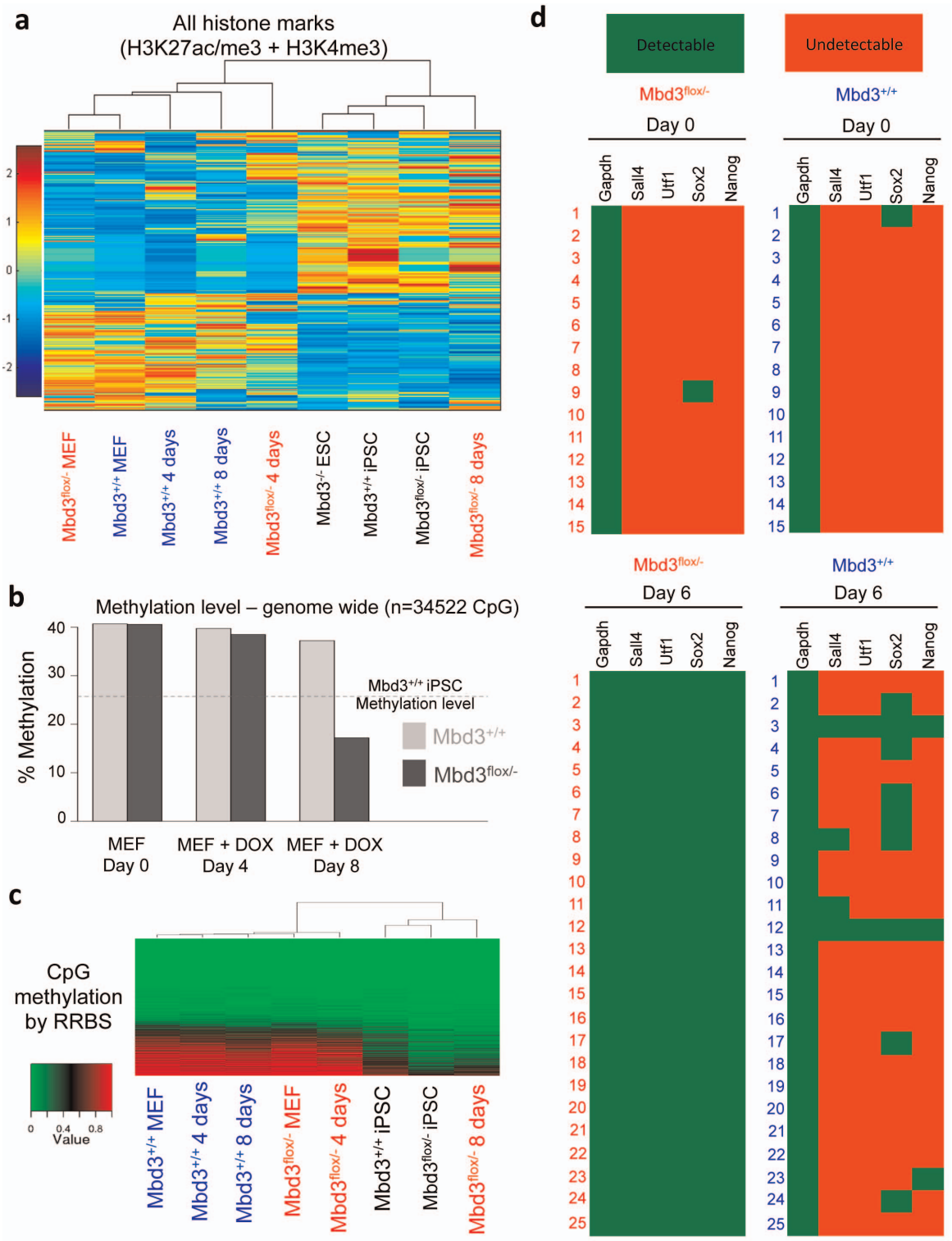
Extended Data Figure 2 | Derivation of ES cells from *Mbd3*$^{-/-}$ blastocysts. **a**, RT–PCR analysis for Oct4 and trophoblast marker expression of *Mbd3*$^{+/+}$ and *Mbd3*$^{-/-}$ ES cells expanded either in FBS/LIF or 2i/LIF conditions. Only *Mbd3*$^{-/-}$ ES cells, and only in serum conditions, upregulate trophoblast differentiation markers. Error bars indicate s.d. from average ($n = 3$). **b**, *Mbd3*$^{+/-}$ heterozygous mice were mated, and *Mdb3*$^{-/-}$ ES cells were derived from blastocysts in naive defined 2i/LIF conditions. Western blot for pluripotency marker expression also indicated that the derived *Mbd3*$^{-/-}$ ES cell lines adequately expressed all pluripotency factors tested. **c**, Transcriptional expression of Mbd3 and Nanog during pre-implantation development. RT–PCR analysis demonstrating the expression of Mbd3 during early mouse development, presented as a relative quantification column scheme. Error bars indicate s.d. from average ($n = 3$). *Mbd3* transcript is detected at low levels in oocytes whereas Mbd3 protein is weakly detected by immunostaining in oocytes and zygotes (Fig. 1e), consistent with maternal inheritance. *Mbd3* transcription becomes increased towards the end of pre-implantation development at the morula and blastocyst stages, consistent with strong re-expression of Mbd3 protein at the blastocyst stage (Fig. 1e). **d**, Immunostaining for Mbd3 and lineage markers in E5.5 post-implantation epiblast, indicating prominent expression ($n = 3$ embryos stained).



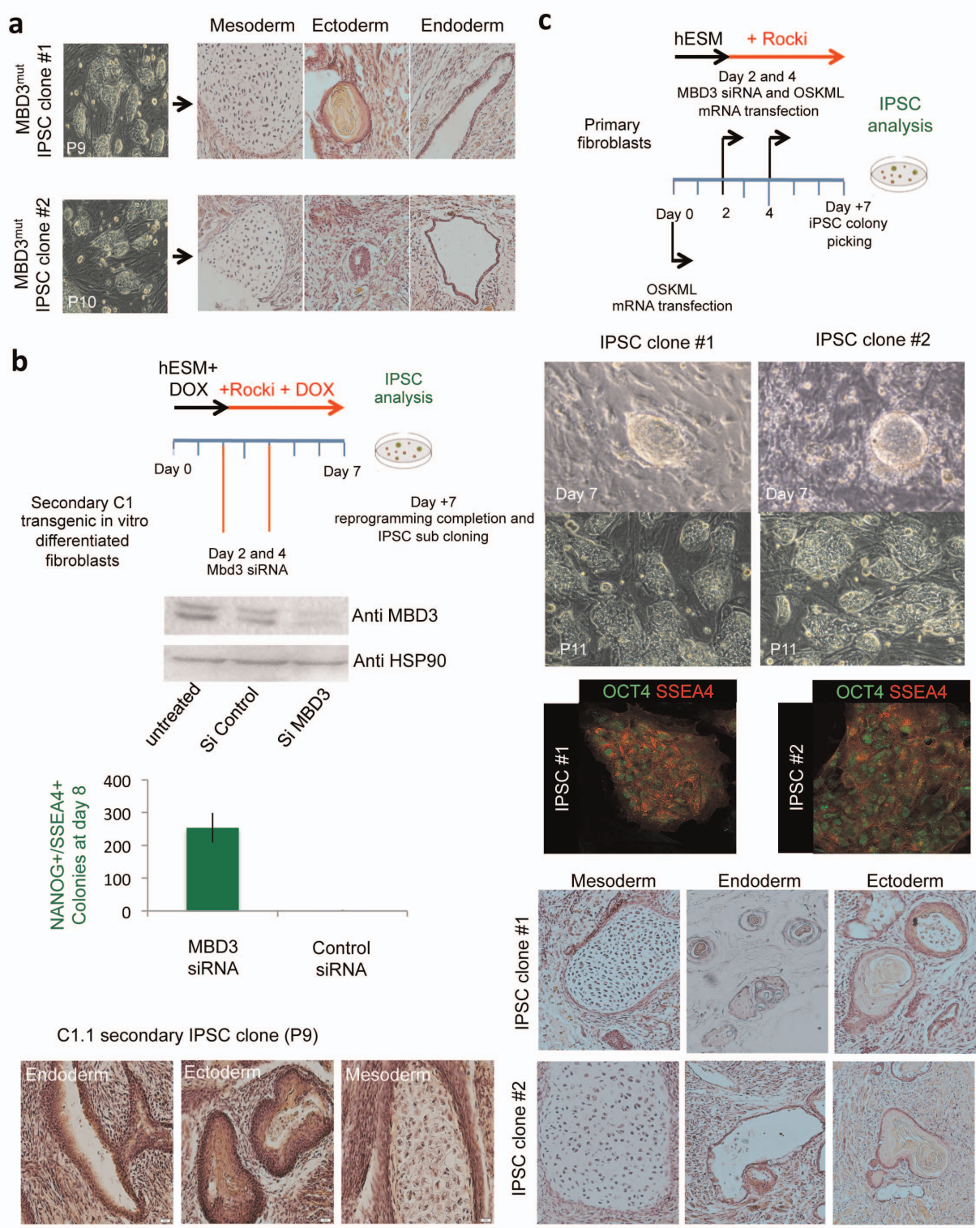
Extended Data Figure 3 | Genetically engineered systems for deterministic reprogramming in mouse cells. **a**, We established a reprogrammable mouse *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ iPS cell lines carrying (1) an Oct4–GFP reporter, (2) nuclear mCherry constitutively expressed marker, (3) m2RtTa transgene and (4) a TetO inducible STEMCCA-OKSM polycistronic cassette. These lines were injected into host blastocysts, and their differentiated derivatives were re-isolated *in vitro*. Subsequently, reprogramming efficiency and progression were analysed after doxycycline induction. **b**, Reprogramming efficiency after infection with indicated MEF lines with moloney retroviruses encoding individual factors. **c**, Reprogramming efficiency after infection with indicated MEF lines with polycistronic OKSM encoding lentivirus. **d**, *Mbd3*$^{fl/-}$ MEFs were infected with polycistronic OKSM vector in LIF-containing ES medium with or without the indicated exogenous supplements. Reprogramming efficiency was evaluated by Oct4–GFP levels on day 9 after transduction without cell splitting during the process. **e**, *Mbd3*$^{+/+}$, *Mbd3*$^{-/-}$ and *Mbd3*$^{fl/-}$ MEFs, adult tail-tip-derived fibroblast (TTF) and neural precursor cells (NPC) were tested for iPS cell formation in 2i/LIF with or without OKSM lentiviral transduction. Our analysis indicates that OKSM is essential for iPS formation, and that Mbd3 depletion alone is not sufficient to reprogram any of these cells types to pluripotency (even after 30 days of follow up). **f**, Reprogramming efficiency of MEFs after transduction with the indicated combinations of reprogramming factors at day 10. Polycistronic lentiviral vectors were used for OSK and OSKM combinations. Asterisk indicates *t*-test *P* value <0.01 relative to *Mbd3*$^{+/+}$ control. Error bars indicate s.d. from average ($n = 4$).



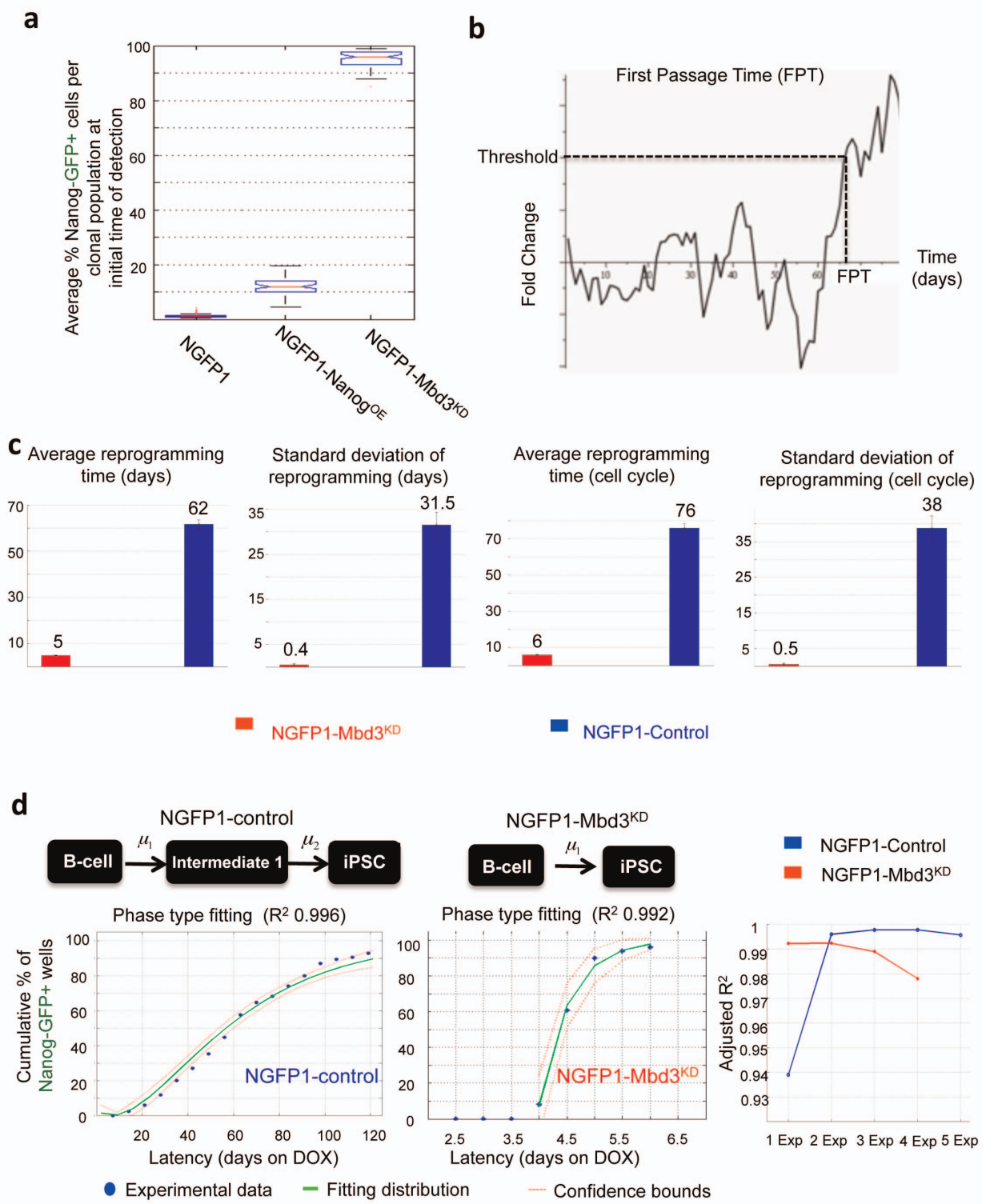
Extended Data Figure 4 | Reprogramming kinetics on perturbation of Mbd3 expression. a, Flow cytometry measurements of Oct4–GFP reactivation dynamics in 2i/LIF after doxycycline (OSKM) induction. Notably, wells at the indicated time points were collected for analysis without prior passaging and splitting during the reprogramming course. 1 out of 3 independent experiments is shown. FSC, forward scatter. **b**, Characterizing the effect for Mbd3 expression reconstitution during deterministic reprogramming of somatic cells to pluripotency. Scheme demonstrates experimental strategy for defining the temporal ability of Mbd3 during reprogramming to inhibit iPS formation. Secondary OSKM reprogrammable *Mbd3*$^{fl/-}$ MEFs were tested for their amenability to reprogramming after overexpression of Mbd3, Mbd2 or empty FUW lentiviruses at different time points during reprogramming. Mbd2 or mock-vector transfection did not result in a decrease in iPS cell reprogramming efficiency. Error bars indicate s.d. from average ($n = 3$). One out of two representative data sets is shown.



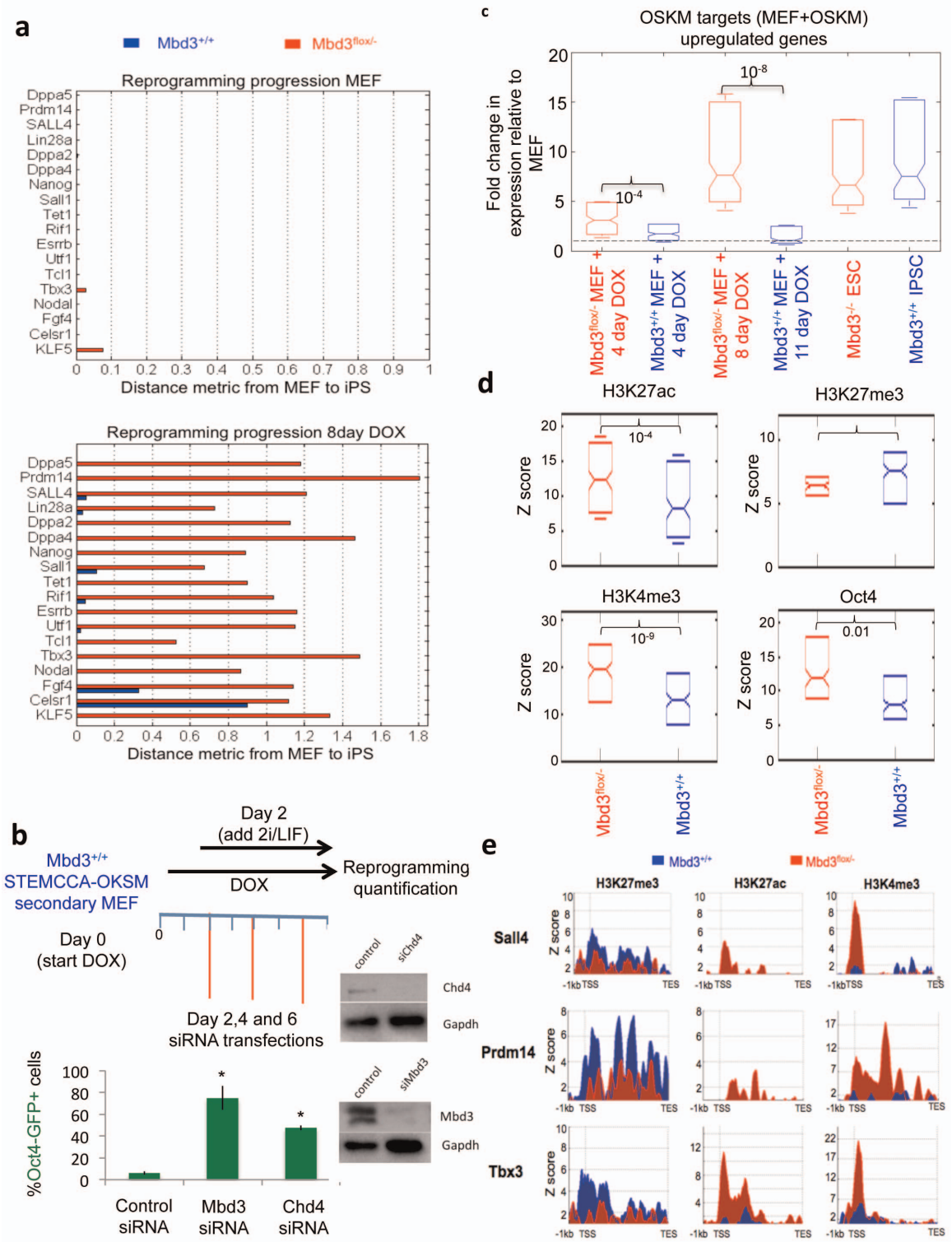
Extended Data Figure 5 | Genetic and epigenetic changes during iPS cell reprogramming after Mbd3 depletion. **a**, Hierarchical clustering was carried out on chromatin IP-seq measurements in fibroblasts before and after doxycycline induction. Clustering was calculated over concatenate vectors including *z*-scores of all histone marks (H3K4me3, H3K27me3 and H3K27ac) for each gene ($n = 1{,}323$ genes with differential gene expression between MEFs and ES cells). Spearman correlation was used as a distance metric and average linkage. **b**, Graph shows genome-wide methylation levels as measured by reduced representation bisulphite sequencing (RRBS). Results are averaged over all CpGs that were covered by five or more distinct sequencing reads (34,522 CpG sites in total). The average methylation level of low-passage $Mbd3^{+/+}$ iPS cells is provided as a dashed line for reference. **c**, Hierarchical clustering for CpG methylation was made using Ward's method and the Pearson correlation score as the similarity matrix. **d**, Single cell RT–PCR analysis for detection of pluripotency gene markers. Analysis was conducted on $Mbd3^{+/+}$ and $Mbd3^{fl/-}$ MEFs before and 6 days after doxycycline induction. Undetected expression (marked by red boxes) indicates lack of amplification even after 50 amplification cycles are marked in red. Expressed genes are marked by green boxes. One biological replicate is shown of two performed.



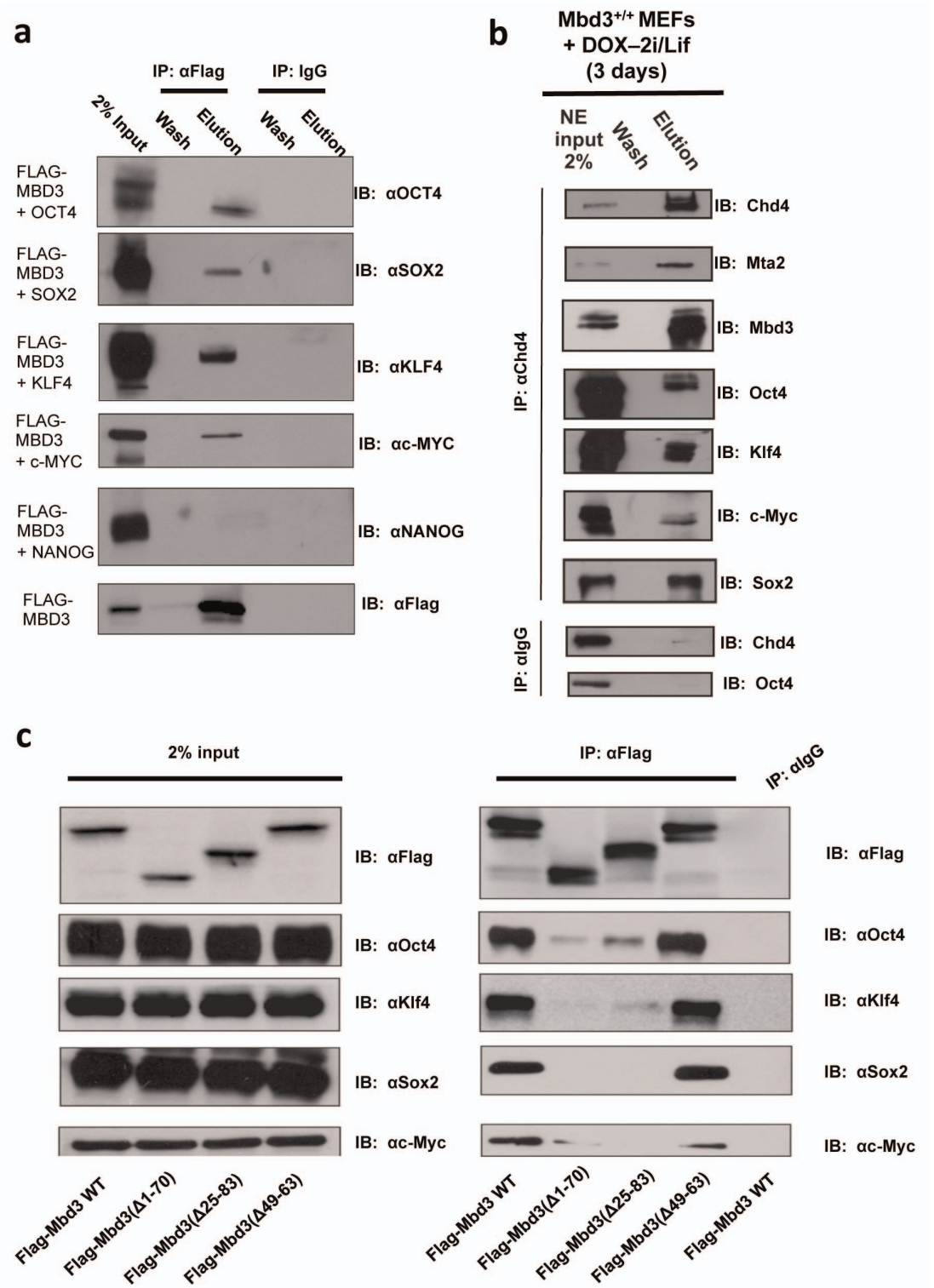
Extended Data Figure 6 | Depleting Mbd3 expression facilitates human iPS cell formation. **a**, *In vitro* differentiated fibroblasts from MBD3^{WT} and MBD3^{mut} iPS cells carrying the doxycycline-inducible OKSM transgenes, were reprogrammed as indicated in Fig. 3f. Pluripotency of randomly selected iPS cell clones is shown as evident by teratoma. **b**, Secondary human reprogrammable C1 fibroblasts carrying doxycycline-inducible OSKM transgenes were subjected to the depicted reprogramming protocol. Knockdown of Mbd3 at days 2 and 4, but not with scrambled control siRNA, markedly increased the reprogramming efficiency as evaluated by formation of NANOG/SSEA4^{+} colonies. Pluripotency of a randomly selected iPS cell clone expanded and validated by *in vivo* teratoma formation. Western blot confirmed specific and significant decrease in MBD3 protein expression after *MBD3* siRNA transfection. Error bars indicate s.d. from average ($n = 3$). One out of three representative experiments is shown. **c**, *MBD3* siRNA treatment of human primary fibroblasts allows generation of iPS cells by only two rounds of reprogramming with mRNA transfection with OSKM and LIN28 (OSKML) factors. Representative human iPS cell clones are shown at different time points and passages (P indicates passage number). Pluripotency of randomly selected clones is shown by specific staining for OCT4 and SSEA4 pluripotency markers and teratoma formation. These results indicate that inhibition of MBD3 expression and/or function promotes iPS cell formation by transient mRNA or other transient transfection protocols for iPS cell reprogramming.



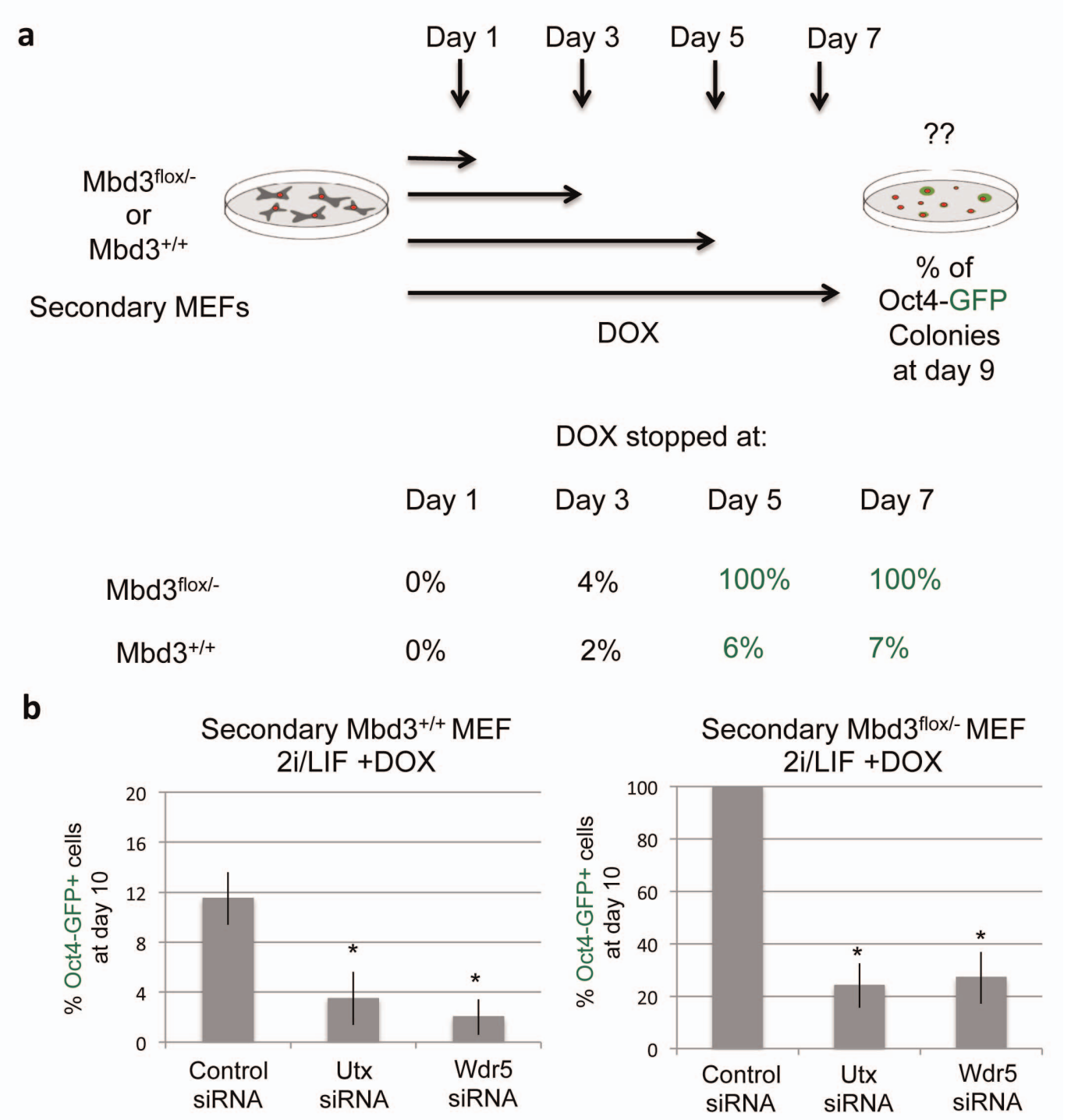
Extended Data Figure 7 | Statistical analysis of iPS cell reprogramming after Mbd3 depletion. **a**, Distribution of Nanog–GFP$^+$ cells at initial time of detection, by quantifying the amount of Nanog–GFP$^+$ cells detected above the 0.5% threshold. Graphs show box-plot medians and 25th/75th percentiles. **b**, Illustration of the first passage time model. In this model, we assume that reprogramming time depends on the first time in which some master regulator (that is, Nanog or Oct4) makes a transition from a low state to a high state of expression. **c**, *Mbd3*^{KD} and *Mbd3*$^{+/+}$ reprogramming dynamics were fit to Gaussian distribution. Figures show maximum likelihood estimates of mean and standard deviation, with 95% confidence intervals. **d**, *Mbd3*^{KD} and *Mbd3*$^{+/+}$ reprogramming dynamics were fit to multiple tandem rate-limiting step models, where convergence of adjusted R^2 indicates the best fit (right panel). Results show that *Mbd3*$^{+/+}$ (blue) fit best to a multi-phase process with one or two intermediate states, whereas *Mbd3*^{KD} (red) fit best to a single exponential transition with no intermediate states.



Extended Data Figure 8 | Effect of Mbd3 depletion on OSKM target genes. **a**, Normalized single gene expression for selected group of genes in MEF and 8 days after doxycycline induction. Expression values represent distance from MEF expression values (set to 0) towards iPS values (set to 1), indicating absence of transcription in MEF and fast activation after doxycycline induction in Mbd3 depleted samples. **b**, Reprograming efficiency of *Mbd3*$^{+/+}$ secondary MEFs after knockdown of Mbd3 or Chd4. Error bars indicate s.d. from average ($n = 3$). Asterisks indicate Student's *t*-test *P* value <0.01. Western blot indicating protein depletion efficiency on siRNA transfection of either Mbd3 or Chd4 targeting siRNA pools. **c**, Distribution of gene expression fold-change relative to MEF, calculated over 2,928 genes bound by at least one of the OSKM factors[20] and upregulated during reprogramming. Graphs show box-plot medians and 25th/75th percentiles, and *P* values by paired sample *t*-test. **d**, Distribution of histone marks and Oct4 binding levels in *z*-score values at day 4 after OSKM (doxycycline) induction, calculated over the same set of 2,928 genes described above. **e**, Histone mark *z*-score profiles for three representative OSKM target genes, calculated between 1 kb upstream to TSS and TES.



Extended Data Figure 9 | Direct interaction of Mbd3 with OSKM pluripotency factors during reprogramming. **a**, Overexpression of Flag-tagged Mbd3 simultaneously with OCT4, SOX2, KLF4, MYC or Nanog in HEK293 cells was followed by co-immunoprecipitation (co-IP) assay. Immunoblot analysis (IB) using antibodies against Oct4, Sox2, Klf4, Myc and Nanog showed specific binding between Mbd3 and the pluripotent factors except Nanog ($n = 2$). **b**, Co-immunoprecipitation assay of Chd4 (Mi2b), the core subunit of the NuRD complex, in secondary *Mbd3*$^{+/+}$ fibroblasts 3 days after doxycycline induction. Co-immunoprecipitation for NuRD component, Chd4, followed by immunoblot analysis indicated specific pull-down of other Mbd3/NuRD components (Mbd3 and Mta2) and OSKM reprogramming factors ($n = 3$). **c**, Deletion mutations in the MBD site of Mbd3 was planned to find the binding region of Mbd3. Flag-tagged mutation constructs were co-transfected with Oct4, Sox2, Klf4 and Myc in HEK293T cells for 48 h followed by co-immunoprecipitation with anti-Flag beads and immunoblotted against OSKM. This analysis shows loss of binding and interaction between OSKM and selected Mbd3 mutants ($n = 3$).



Extended Data Figure 10 | Pluripotency-promoting epigenetic activators are essential for both deterministic and stochastic iPS cell formation. **a**, Requirement for doxycycline-mediated transgene induction during iPS cell reprogramming form *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ secondary MEFs. Percentage of Oct4–GFP colonies was quantified at final set time point on day 9. Similar time frame for minimal doxycycline induction was required for iPS cell formation in both cell samples (irrespective of the total iPS formation efficiency obtained). Representative data from one out of three biological replicates conducted. **b**, **c**, Specific knockdown of Utx and Wdr5 epigenetic regulators that are required for iPS cell formation significantly inhibited iPS cell formation in both *Mbd3*$^{+/+}$ and *Mbd3*$^{fl/-}$ cells. Asterisks indicate *t*-test *P* value <0.01 in comparison to control siRNA sample. Error bars indicate s.d. from average ($n = 3$).

doi:10.1038/nature12512

Attractive photons in a quantum nonlinear medium

Ofer Firstenberg[1]*, Thibault Peyronel[2]*, Qi-Yu Liang[2], Alexey V. Gorshkov[3]†, Mikhail D. Lukin[1] & Vladan Vuletić[2]

The fundamental properties of light derive from its constituent particles—massless quanta (photons) that do not interact with one another[1]. However, it has long been known that the realization of coherent interactions between individual photons, akin to those associated with conventional massive particles, could enable a wide variety of novel scientific and engineering applications[2,3]. Here we demonstrate a quantum nonlinear medium inside which individual photons travel as massive particles with strong mutual attraction, such that the propagation of photon pairs is dominated by a two-photon bound state[4–7]. We achieve this through dispersive coupling of light to strongly interacting atoms in highly excited Rydberg states. We measure the dynamical evolution of the two-photon wavefunction using time-resolved quantum state tomography, and demonstrate a conditional phase shift[8] exceeding one radian, resulting in polarization-entangled photon pairs. Particular applications of this technique include all-optical switching, deterministic photonic quantum logic and the generation of strongly correlated states of light[9].

Interactions between individual photons are being explored in cavity quantum electrodynamics, where a single, confined electromagnetic mode is coupled to an atomic system[10–12]. Our approach is to couple a light field propagating in a dispersive medium to highly excited atomic states with strong mutual interactions[13,14] (Rydberg states). Similar to previous studies of quantum nonlinearities involving Rydberg states that were based on dissipation[15–19] rather than dispersion[20], we make use of electromagnetically induced transparency (EIT) to slow down the propagation of light[21] in a cold atomic gas. By operating in a dispersive regime away from the intermediate atomic resonance (Fig. 1b), where atomic absorption is low and only weakly nonlinear[22], we realize a situation in which Rydberg-atom-mediated coherent interactions between individual photons dominate the propagation dynamics of weak light pulses. Previous theoretical studies have proposed various scenarios for inducing strong interactions between individual photons[2,3,23] and for creating bound states of a few quanta[4,5,7,24], a feature generic to strongly interacting quantum field theories. The main result reported here is the experimental realization of a photonic system with strong attractive interactions, including evidence for a predicted two-photon bound state.

Our experiment (outlined in Fig. 1a) makes use of an ultracold rubidium gas loaded into a dipole trap, as described previously[19]. The probe light of interest is σ^+-polarized, coupling the ground state, $|g\rangle$, to the Rydberg state, $|r\rangle$, via an intermediate state, $|e\rangle$, of linewidth $\Gamma/2\pi = 6.1$ MHz by means of a control field that is detuned by Δ below the resonance frequency of the upper transition, $|e\rangle \rightarrow |r\rangle$ (Fig. 1b). Under these conditions, for a very weak probe field with mean incident photon rate $R_i = 0.5\ \mu s^{-1}$, EIT is established when the probe detuning matches that of the control field (see Fig. 1c, which shows the probe transmission and phase shift). However, the Rydberg medium is extremely nonlinear: a probe photon rate of $R_i = 5\ \mu s^{-1}$ saturates the medium as a result of the Rydberg blockade[25], yielding a probe spectrum close to the bare two-level response. Given the measured system bandwidth of about $5\ \mu s^{-1}$, this implies a substantial nonlinear response with average pulse energies corresponding to less than one photon per inverse bandwidth. We perform our experiments on the two-photon resonance $|g\rangle \rightarrow |r\rangle$, where, for $|\Delta| > \Gamma$, the transmission is approximately independent of the probe photon rate for our experimental parameters, yielding a purely dispersive nonlinearity. The linear dispersion at this resonance corresponds to a reduced probe group velocity of typically $v_g = 400\ \mathrm{m\,s^{-1}}$, and the group velocity dispersion endows the photons with an effective mass[26] of $m \approx 1{,}000\hbar\omega/c^2$, where ω is the optical frequency, $\hbar$ is Planck's constant divided by 2π and c is the speed of light in vacuum.

To explore the quantum dynamics in the propagation of photon pairs, we measure time-dependent two-photon correlation functions of the transmitted light (Fig. 1a). To determine both the amplitude and the phase of the σ^+-polarized probe field, we prepare input coherent light in a linearly polarized state, $|V\rangle = (|\sigma^+\rangle + |\sigma^-\rangle)/\sqrt{2}$, where the σ^--polarized component (which is approximately non-interacting owing to the 15-fold-smaller transition strength) serves as a phase reference. To analyse the properties of photon pairs, we measure two-photon correlation functions, $g^{(2)}_{\alpha\beta}$, in different polarization bases, α and β (Supplementary Fig. 3). The component $g^{(2)}_{++}$ directly gives the probability density of the σ^+-polarized interacting photon pairs. Figure 1d shows $g^{(2)}_{++}$, for a control detuning of $\Delta/2\pi = 14$ MHz, as a function of the time separation, $\tau = t_1 - t_2$, between the photons detected at times t_1 and t_2, converted into a relative distance, r, in the medium using the group velocity, v_g. A prominent feature is the cusp at $r = v_g\tau = 0$, which is characteristic of a predicted two-photon bound state[5,7], as discussed below.

The measured $g^{(2)}_{\alpha\beta}$ allow us to reconstruct the two-photon density matrix, ρ, using quantum state tomography and maximum-likelihood estimation[27,28]. From ρ, we define an interaction matrix, $\tilde{\rho}$, by factoring out the linear response, such that $\tilde{\rho}$ directly quantifies the nonlinearity (Methods). The density matrix approach is necessary to account for decoherence and technical imperfections. The probability density of two interacting σ^+ photons, $g^{(2)}_{++} = \tilde{\rho}_{++,++}$, and the nonlinear phase, $\phi = \arg[\tilde{\rho}_{++,--}]$, acquired by the $\sigma^+\sigma^+$ pair relative to a non-interacting $\sigma^-\sigma^-$ pair, are shown in Fig. 2a, b for $\Delta/2\pi = 14$ MHz. The time dependence allows us to extract the nonlinear phase as a function of the photon–photon separation. Clearly visible is the bunching of photons, that is, an increased probability for photons to exit the medium simultaneously ($t_1 \approx t_2$), and a substantial nonlinear two-photon phase shift of -0.5 rad in that region. Figure 2c shows the intensity correlation in the dissipation-dominated antibunching regime[19] at $\Delta = 0$ and in the dispersive regime at $|\Delta| > \Gamma$, where there is bunching. Figure 2d displays the nonlinear phase for two different detunings. The transition from the dissipative regime to the dispersive is summarized in Fig. 3a, b. In the dispersive regime, the nonlinear phase shift, $\phi(\tau = 0)$, can reach $(-0.32 \pm 0.02)\pi$, at a detuning $\Delta/2\pi = 9$ MHz and a linear transmission of order 50%. The observed signals, particularly ϕ, are asymmetrical under a sign change of Δ.

The origin of the quantum nonlinearity underlying these observations is explained by the following simple model. The repulsive van der Waals interaction between two Rydberg atoms, $V(r) = \hbar C_6/r^6$, tunes

[1]Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Department of Physics and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA. †Present address: Joint Quantum Institute, NIST/University of Maryland, College Park, Maryland 20742, USA.
*These authors contributed equally to this work.

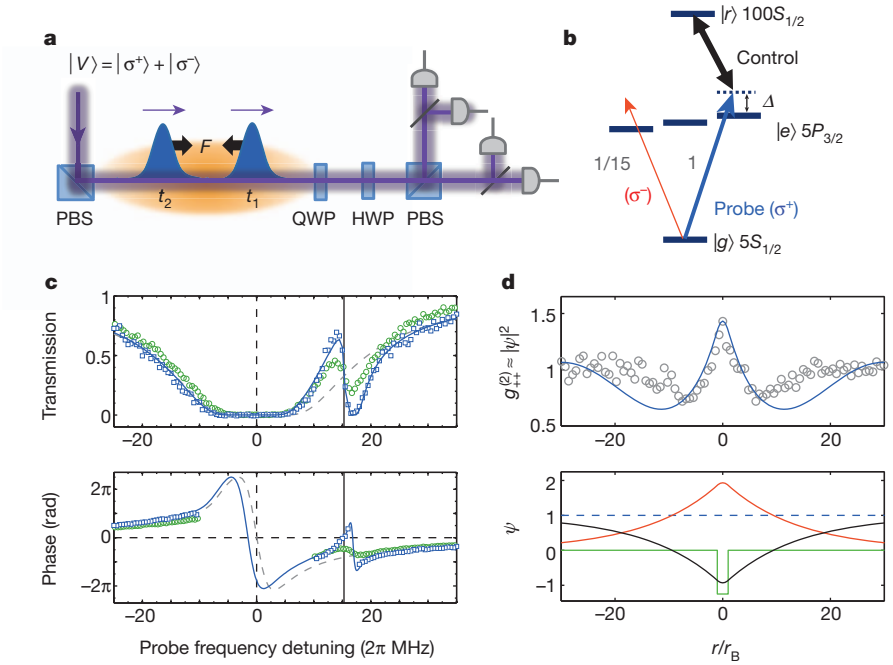


Figure 1 | Photons with strong mutual attraction in a quantum nonlinear medium. **a**, **b**, A linearly polarized weak laser beam near the transition $|g\rangle \rightarrow |e\rangle$ at 780 nm is sent into a cold rubidium gas driven by a control laser near the transition $|e\rangle \rightarrow |r\rangle$ at 479 nm. Strong nonlinear interactions between σ^+-polarized photons are detected using photon–photon correlation functions of the transmitted light for a set of different polarization bases, as determined by a quarter-wave plate (QWP), a half-wave plate (HWP) and a polarizing beam splitter (PBS). Here σ^- photons serve as a phase reference. **c**, Transmission spectra (top) and phase shift (bottom) for σ^+ photons with an incoming rate of $R_i = 0.5\ \mu s^{-1}$ (blue squares) or $R_i = 5\ \mu s^{-1}$ (green circles), for a control field red-detuned by $\Delta/2\pi = 15$ MHz. The blue line shows the theoretical spectrum. The spectrum at high probe rate approaches that of the undriven two-level system (dashed grey; see also Supplementary Fig. 2). The solid vertical line corresponds to the EIT resonance. **d**, Photon bunching and two-photon bound state. Theoretically predicted photon–photon correlation function in the Schrödinger equation approximation (top, blue line) for $\Delta/2\pi = 14$ MHz, with a potential well of width $2r_B$ (bottom, green line). The bound state (bottom, red) and the superposition of scattering states (bottom, black) form the initial wavefunction, $\psi = 1$ (bottom, dashed blue). The two-photon bound state results in the observed bunching in the correlation function, $g^{(2)}_{++} \approx |\psi|^2$ (top, grey circles), where time has been converted into distance using the group velocity, v_g. The boundary effects resulting from the finite extent of the atom cloud become important for $|r| \geq 5r_B$.

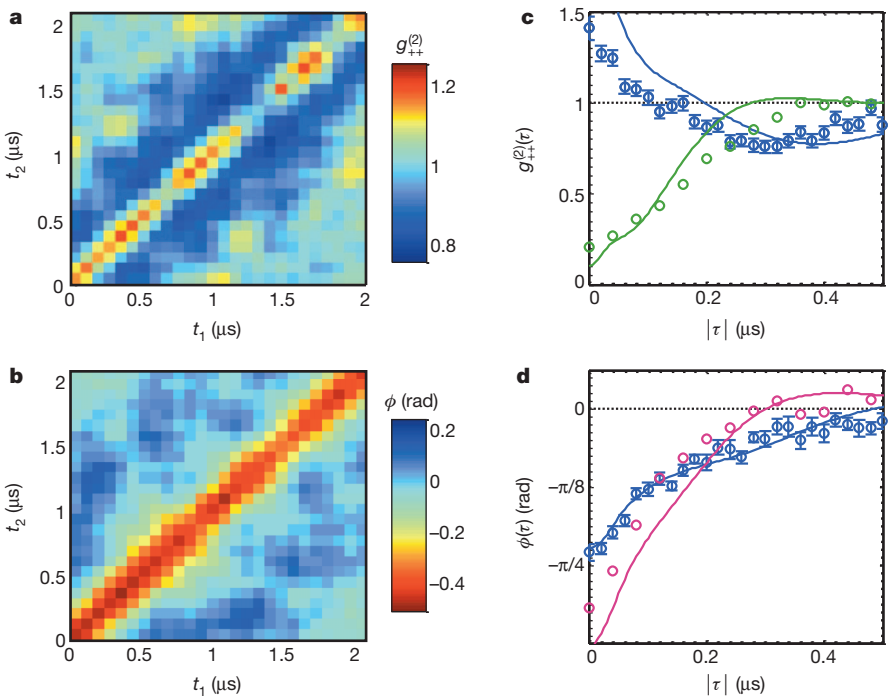


Figure 2 | Propagation of interacting photon pairs. **a**, **b**, Measured second-order correlation function (**a**) and nonlinear phase shift (**b**) of interacting photon pairs at $\Delta = 2.3\Gamma$. The photons are detected at times t_1 and t_2. **c**, Second-order correlation function displayed as a function of the time difference, $|\tau| = |t_1 - t_2|$, between the photons, showing the transition from antibunching on resonance ($\Delta = 0$, green) to bunching at large detuning ($\Delta = 2.3\Gamma$, blue). Points are experimental data; lines are full numerical simulations. All $g^{(2)}_{++}$ measurements are rescaled by their value at $\tau > 1.5\ \mu s$ (Supplementary Information). **d**, Nonlinear phase shift versus $|\tau|$ for two different detunings ($\Delta = 1.5\Gamma$, purple, and $\Delta = 2.3\Gamma$, blue). The 1 s.d. error is ± 30 mrad, dominated by photon shot-noise.

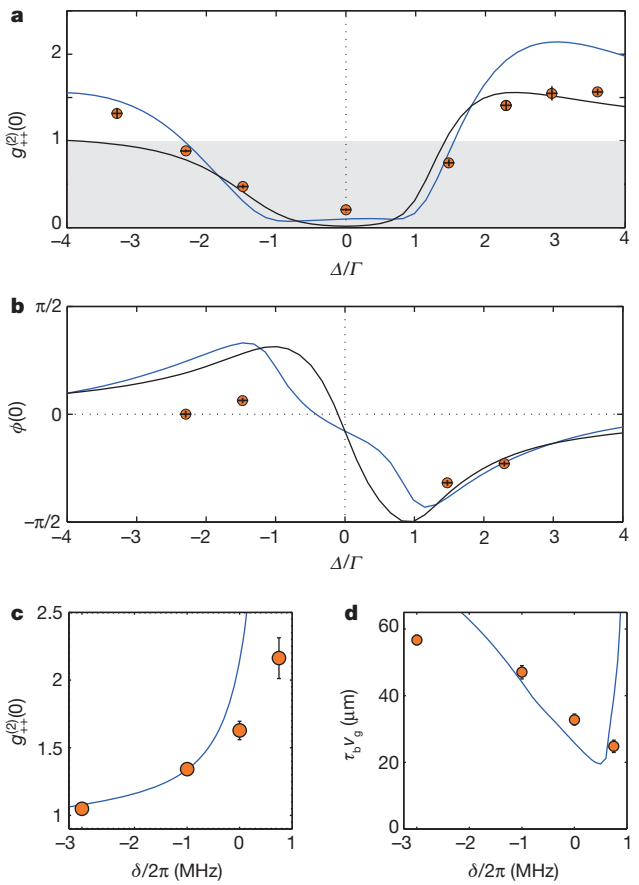


Figure 3 | Dependence of the photon–photon interaction on detuning. **a**, **b**, Equal-time two-photon correlation, $g^{(2)}_{++}(0)$ (**a**), and nonlinear phase, $\phi(0)$ (**b**), versus detuning, Δ, from the intermediate state, $|e\rangle$. Blue lines are full theoretical simulations and black lines are the result of the Schrödinger equation approximation, assuming a simplified δ-function potential. Vertical error bars represent 1 s.d. and horizontal error bars are $\pm 0.5 \times 2\pi$ MHz. **c**, **d**, Equal-time correlation function (**c**) and spatial extent of the bunching feature (**d**) versus Raman detuning, δ, from the EIT resonance $|g\rangle \rightarrow |r\rangle$ for $\Delta = 3\Gamma$, showing increased photon–photon attraction due to a deeper potential near Raman resonance. The characteristic bunching timescale, τ_b, is the half-width of the cusp feature of $g^{(2)}_{++}$, defined at half-height between the peak value at $\tau = 0$ and the local minimum closest to $\tau = 0$. Error bars, 1 s.d. The theoretical model (solid line) breaks down close to the Raman resonance at $\delta = 1.3 \times 2\pi\,\text{MHz} \approx \Omega_c^2/4\Delta$, where the single-photon component of the probe field is strongly absorbed.

the doubly excited Rydberg state far off EIT resonance for distances $|r| < r_B$, where $r_B = \sqrt[6]{C_6/\gamma}$ is the Rydberg blockade radius[14,29,30], C_6 is the van der Waals coefficient, $\gamma = \Omega_c^2/|4\Delta|$ is the EIT linewidth at detuning $|\Delta| \gg \Gamma$ and Ω_c is the Rabi frequency of the control field. Although the phase shift that would originate from the bare $|g\rangle \rightarrow |e\rangle$ probe transition is suppressed by EIT for photons with large separation in the medium ($|r| > r_B$), the light acquires this phase shift for small photon separations ($|r| \leq r_B$) (Fig. 1c). This explicit dependence of the refractive index on photon–photon separation can be modelled in one dimension as a potential well with a characteristic width of $2r_B$. Qualitatively, a substantial two-photon phase shift arises for $(r_B/l_a)(\Gamma/|\Delta|) \gtrsim 1$, where l_a is the resonant attenuation length in the medium, that is, for sufficiently high atomic density. Furthermore, the probe field must also be compressed in the transverse direction to a waist size $w < r_B$ to ensure that interactions occur. Using the Rydberg state $100S_{1/2}$, and for $\Omega_c/2\pi = 10$ MHz, we obtain $r_B \approx 18$ μm at detunings of a few Γ, $l_a = 4$ μm at the peak density, and $w = 4.5$ μm, fulfilling the conditions for strong interactions for $|\Delta| \lesssim 5\Gamma$.

The propagation of σ^+-polarized photon pairs in such a medium can be understood by first considering an idealized situation with no decoherence between the Rydberg state and the ground state. Then the steady state in a one-dimensional homogenous medium can be described by a two-photon wavefunction, $\psi(z_1, z_2)$, whose evolution is approximately governed by a simple equation[19] in terms of the centre-of-mass coordinate, $R = (z_1 + z_2)/2$, and the relative coordinate, $r = z_1 - z_2$:

$$i\frac{\partial\psi}{\partial R} = 4l_a\left[i + \frac{2\Delta}{\Gamma} - \mathcal{V}(r)\frac{\Omega_c^2}{\Gamma^2}\right]\frac{\partial^2\psi}{\partial r^2} + \frac{\mathcal{V}(r)}{l_a}\psi \qquad (1)$$

Here the effective potential, $\mathcal{V}(r) = \left[i + 2(\Delta/\Gamma)(1 + 2r^6/r_B^6)\right]^{-1}$, approaches $(i + 2\Delta/\Gamma)^{-1}$ inside the blockaded volume ($|r| < r_B$) and approaches zero outside. The solution relates approximately to our measurements in the time domain for small $|\tau|$ via $\psi(R = L, r = v_g\tau) \approx \sqrt{g^{(2)}_{++}(\tau)}e^{i\phi(\tau)}$ (see Supplementary Information for the exact relation). Far off resonance ($|\Delta| \gg \Gamma$, Ω_c), equation (1) corresponds to a Schrödinger equation with R playing the part of effective time. The photons' effective mass, $m \propto -\Gamma/16l_a\Delta$, can be positive or negative depending on the sign of the detuning, Δ. Because the sign of the potential also changes with Δ (potential well for $\Delta < 0$; barrier for $\Delta > 0$), the effective force (F in Fig. 1a) in both cases is attractive and the resulting dynamics similar (Supplementary Information). However, the potential for $\Delta < 0$ also has additional features near the edges of the well, corresponding to a Raman resonance $|g\rangle \rightarrow |r\rangle$ for the interaction-shifted Rydberg state at some interatomic distance near $|r| = r_B$. These features are probably responsible for the deviation from symmetry (or antisymmetry) under the change of the sign of Δ displayed in Fig. 3a, b.

In the experimentally relevant regime, the effective potential supports only one bound-state, $\psi_B(r)$ (Fig. 1d). The initial wavefunction, $\psi(R = 0, r) = 1$, is a superposition of $\psi_B(r)$ and the continuum of scattering states. The accumulation of probability near $r = 0$ can then be understood as arising from the interference between the bound and scattering states that evolve at different frequencies, and the observed bunching feature in $g^{(2)}_{++}$ reveals the wavefunction of the two-photon bound state (Supplementary Information). As shown in Fig. 3a, b, the solution of the Schrödinger-like equation (1) with a simplified δ-function potential captures the essential features of the nonlinear two-photon propagation. Additional experimental evidence for the bound-state dynamics is obtained by tuning the probe field relative to the EIT resonance, thereby varying the strength of the two-photon interaction potential. As the probe detuning approaches the Raman resonance, the difference in refractive indices inside and outside the blockade radius increases and the potential deepens (Supplementary Information and Fig. 1c). Consequently, the bound state becomes more localized and the bunching, quantified by $g^{(2)}_{++}(0)$, is enhanced (Fig. 3c, d). We note that the size of the two-photon bound state and, correspondingly, the width of the bunching feature, $2\tau_b v_g \approx 70$ μm, exceed the width of the potential well, $2r_B \approx 35$ μm, as expected for a potential with one weakly bound state.

Figures 2 and 3 also show the results of our full theoretical model, in which we numerically solve the set of propagation equations for the light field and atomic coherences. The model incorporates the longitudinal atomic density distribution and the decoherence of the Rydberg state (Supplementary Information). These simulations are in good agreement with our experimental results and the predictions of the simplified model (equation (1)), confirming that the evolution of the two-photon wave packet is dominated by the attractive force between the photons.

Finally, we study the quantum coherence and polarization properties of the transmitted photon pairs. In Fig. 4a, we compare the purity of the two-photon density matrix, $\rho(\tau)$, which includes photon interactions, with the purity of the product of one-photon matrices, $\rho^{(1)} \otimes \rho^{(1)}$, for non-interacting photons. At large photon separation, τ, the purity, $P(\tau)$, of the two-photon density matrix is dominated by

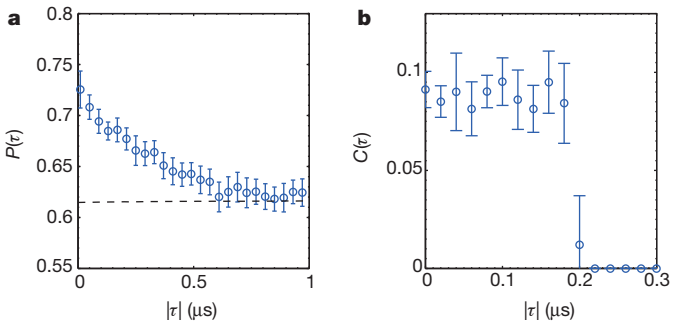


Figure 4 | Quantum coherence and entanglement. **a**, Purity, $P(\tau) = \mathrm{Tr}[\rho(\tau)^2]$, of the measured two-photon density matrix, ρ, for $\Delta = 2.3\Gamma$ (blue symbols), which at large photon separation approaches the purity expected from the measured one-photon density matrix, $\mathrm{Tr}[(\rho^{(1)} \otimes \rho^{(1)})^2]$ (dotted black line). Interacting $\sigma^+\sigma^+$ photon pairs near $\tau = 0$ exhibit lower decoherence. Error bars (1 s.d.) are derived from the uncertainty in the density matrix due to detection shot-noise. **b**, Concurrence, $C(\tau)$, calculated from ρ, indicating polarization entanglement of proximal photons on transmission through the quantum nonlinear medium.

the one-photon decoherence due to partial depolarization of the transmitted light (Supplementary Information). This depolarization is attributed to the difference in group delay, τ_d, between the σ^+ photons and the faster σ^- photons ($\tau_d^{\sigma^+} - \tau_d^{\sigma^-} = 280$ ns), which is not negligible compared with the coherence time of the probe laser (650 ns). At the same time, σ^+ photons bound to each other travel faster and are more robust against this decoherence mechanism, as evidenced by the greater purity at small τ. Even in the presence of this depolarization, the coherent nonlinear interaction in the dispersive medium produces entanglement in the outgoing polarization state of two photons. We quantify the degree of polarization entanglement in terms of a time-dependent concurrence, $C(\tau)$ (Fig. 4b and Supplementary Information). The obtained value $C(0) = 0.09 \pm 0.03$ indicates deterministic entanglement of previously independent photons on passage through the quantum nonlinear medium. The measured value is in reasonable agreement with the theoretical prediction, $C_{\mathrm{th}}(0) = 0.13$, calculated for a conditional phase $\phi(0) = \pi/4$, purity $P(0) = 0.73$ and 50% σ^+ linear transmission.

In our experiment, the transmission and achievable nonlinear phase are limited by the laser linewidth, the atomic motion and the available control-field intensity. These technical limitations can be circumvented by using stronger control lasers with improved frequency stability and colder atomic clouds trapped in both the ground state and the Rydberg state. Although in our present system the nonlinear phase would not be uniformly acquired by a bandwidth-limited two-photon pulse, a high-fidelity two-photon phase gate may be achievable using, for example, a counter-propagating geometry and greater optical depths[14].

The realization of coherent, dispersive photon–photon interactions opens several new research directions. These include the exploration of a novel quantum matter composed from strongly interacting, massive photons[9]. Measurements of higher-order correlation functions may give direct experimental access to quantum solitons composed of a few interacting bosons[24], or to the detection of crystalline states of a photonic gas[9]. By colliding two counter-propagating photons, it may be possible to imprint a spatially homogeneous phase shift of π on the photon pair, corresponding to a deterministic quantum gate[14] for scalable optical quantum computation[13]. Finally, by accessing other Rydberg states via, for example, microwave transitions, it may become possible to control the state of multiphoton pulses with just one quantum of light, thereby realizing a single-photon transistor[6,31] for applications in quantum networks, and the creation of multiphoton entangled states.

METHODS SUMMARY

The experimental setup is detailed in ref. 19. The average resonant optical depth along the atomic cloud is 22, and the peak atomic density is $10^{12}\,\mathrm{cm}^{-3}$. Probe pulses at an average photon rate of $1.6\,\mu\mathrm{s}^{-1}$ are sent into the cell during the 5.5-μs dark-time periods of a modulated optical trap. For the quantum state tomography, we measure the photon coincidence rates in six polarization bases, $\{q, h\} = \{\pi/4, \pi/4\}, \{0, 0\}, \{\pi/8, \pi/8\}, \{0, \pi/16\}, \{\pi/8, \pi/16\}$ and $\{\pi/8, 0\}$, where q and h are the angles of the quarter- and half-wave plates. The duration of the coincidence time bins, varying between 20 and 80 ns, is chosen to capture the temporal dynamics of the correlation functions with reasonable signal-to-noise ratio. For each (t_1, t_2) time bin, we numerically optimize a Hermitian, positive-semidefinite two-photon density matrix, $\rho(t_1, t_2)$, and one-photon density matrix, $\rho^{(1)}(t)$. To extract the nonlinear phase from $\rho(t_1, t_2)$, we rescale for the linear dispersion and loss effects by defining the interaction matrix $\tilde{\rho}_{i,j}(t_1,t_2) = \rho_{i,j}(t_1,t_2) / \left[\rho^{(1)}(t_1) \otimes \rho^{(1)}(t_2)\right]_{i,j}$ in the circular-polarization basis. The interaction matrix generalizes the standard $g^{(2)}_{\alpha,\beta}$ definition to account for nonlinear phases and decoherence, and all its elements are equal to 1 in the absence of nonlinearity.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 May; accepted 29 July 2013.
Published online 25 September 2013.

1. Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, 1997).
2. Milburn, G. J. Quantum optical Fredkin gate. *Phys. Rev. Lett.* **62,** 2124–2127 (1989).
3. Imamoğlu, A. Schmidt, H. Woods, G. & Deutsch, M. Strongly interacting photons in a nonlinear cavity. *Phys. Rev. Lett.* **79,** 1467–1470 (1997).
4. Deutsch, I. H., Chiao, R. Y. & Garrison, J. C. Diphotons in a nonlinear Fabry-Pérot resonator: bound states of interacting photons in an optical "quantum wire". *Phys. Rev. Lett.* **69,** 3627–3630 (1992).
5. Shen, J.-T. & Fan, S. Strongly correlated two-photon transport in a one-dimensional waveguide coupled to a two-level system. *Phys. Rev. Lett.* **98,** 153003 (2007).
6. Chang, D. E., Srensen, A. S., Demler, E. A. & Lukin, M. D. A single-photon transistor using nanoscale surface plasmons. *Nature Phys.* **3,** 807–812 (2007).
7. Cheng, Z. & Kurizki, G. Optical "multiexcitons": quantum gap solitons in nonlinear Bragg reflectors. *Phys. Rev. Lett.* **75,** 3430–3433 (1995).
8. Turchette, Q. A., Hood, C., Lange, W., Mabuchi, H. & Kimble, H. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* **75,** 4710–4713 (1995).
9. Chang, D. E. *et al.* Crystallization of strongly interacting photons in a nonlinear optical fibre. *Nature Phys.* **4,** 884–889 (2008).
10. Fushman, I. *et al.* Controlled phase shifts with a single quantum dot. *Science* **320,** 769–772 (2008).
11. Rauschenbeutel, A. *et al.* Coherent operation of a tunable quantum phase gate in cavity QED. *Phys. Rev. Lett.* **83,** 5166–5169 (1999).
12. Kirchmair, G. *et al.* Observation of quantum state collapse and revival due to the single-photon Kerr effect. *Nature* **495,** 205–209 (2013).
13. Saffman, M., Walker, T. G. & Mølmer, K. Quantum information with Rydberg atoms. *Rev. Mod. Phys.* **82,** 2313–2363 (2010).
14. Gorshkov, A. V., Otterbach, J., Fleischhauer, M., Pohl, T. & Lukin, M. D. Photon-photon interactions via Rydberg blockade. *Phys. Rev. Lett.* **107,** 133602 (2011).
15. Pritchard, J. D. *et al.* Cooperative atom-light interaction in a blockaded Rydberg ensemble. *Phys. Rev. Lett.* **105,** 193603 (2010).
16. Maxwell, D. *et al.* Storage and control of optical photons using Rydberg polaritons. *Phys. Rev. Lett.* **110,** 103001 (2013).
17. Dudin, Y. O. & Kuzmich, A. Strongly interacting Rydberg excitations of a cold atomic gas. *Science* **336,** 887–889 (2012).
18. Petrosyan, D., Otterbach, J. & Fleischhauer, M. Electromagnetically induced transparency with Rydberg atoms. *Phys. Rev. Lett.* **107,** 213601 (2011).
19. Peyronel, T. *et al.* Quantum nonlinear optics with single photons enabled by strongly interacting atoms. *Nature* **488,** 57–60 (2012).
20. Parigi, V. *et al.* Observation and measurement of interaction-induced dispersive optical nonlinearities in an ensemble of cold Rydberg atoms. *Phys. Rev. Lett.* **109,** 233602 (2012).
21. Kasapi, A., Jain, M., Yin, G. Y. & Harris, S. E. Electromagnetically induced transparency: propagation dynamics. *Phys. Rev. Lett.* **74,** 2447–2450 (1995).
22. Venkataraman, V., Saha, K. & Gaeta, A. L. Phase modulation at the few-photon level for weak-nonlinearity-based quantum computing. *Nature Photon.* **7,** 138–141 (2012).
23. Rajapakse, R. M., Bragdon, T., Rey, A. M., Calarco, T. & Yelin, S. F. Single-photon nonlinearities using arrays of cold polar molecules. *Phys. Rev. A* **80,** 013810 (2009).
24. Drummond, P. D. & He, H. Optical mesons. *Phys. Rev. A* **56,** R1107–R1109 (1997).
25. Lukin, M. D. *et al.* Dipole blockade and quantum information processing in mesoscopic atomic ensembles. *Phys. Rev. Lett.* **87,** 037901 (2001).
26. Fleischhauer, M., Imamoglu, A. & Marangos, J. P. Electromagnetically induced transparency: optics in coherent media. *Rev. Mod. Phys.* **77,** 633–673 (2005).

27. James, D. F. V., Kwiat, P. G., Munro, W. J. & White, A. G. Measurement of qubits. *Phys. Rev. A* **64,** 052312 (2001).
28. Adamson, R. B. A., Shalm, L. K., Mitchell, M. W. & Steinberg, A. M. Multiparticle state tomography: hidden differences. *Phys. Rev. Lett.* **98,** 043601 (2007).
29. Sevinçli, S., Henkel, N., Ates, C. & Pohl, T. Nonlocal nonlinear optics in cold Rydberg gases. *Phys. Rev. Lett.* **107,** 153001 (2011).
30. Heidemann, R. *et al.* Evidence for coherent collective Rydberg excitation in the strong blockade regime. *Phys. Rev. Lett.* **99,** 163601 (2007).
31. Chen, W, *et al.* All-optical switch and transistor gated by one stored photon. *Science* **341,** 768–770 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. P. Büchler, T. Pohl, J. Otterbach, P. Strack, M. Gullans and S. Choi for discussions. This work was supported by the NSF, the CUA, DARPA and the AFOSR Quantum Memories MURI and the Packard Foundations. O.F. acknowledges support from the HQOC. A.V.G. and M.D.L. thank KITP for hospitality. A.V.G. acknowledges funding from the Lee A. DuBridge Foundation and the IQIM, an NSF Physics Frontiers Center with support of the Gordon and Betty Moore Foundation.

Author Contributions The experiment and analysis were carried out by O.F., T.P. and Q.-Y.L. The theoretical modelling was done by A.V.G. All experimental and theoretical work was supervised by M.D.L. and V.V. All authors discussed the results and contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.D.L. (lukin@fas.harvard.edu) or V.V. (vuletic@mit.edu).

METHODS

The experimental setup is detailed in ref. 19, with the following modifications. The dipole trap is periodically switched off with a 5.5-μs half-period, and the measurements are performed during the dark time to avoid inhomogeneous broadening. Photons detected in the first 1.5 μs after the dipole trap is turned off are not included in the analysis, to guarantee steady-state EIT. For each experimental cycle, data are accumulated over 400 periods of the dipole-trap modulation. The trapped atomic cloud has a longitudinal root mean squared length of $\sigma_{ax} = 36\,\mu m$ and a peak atomic density of $\rho_0 = 10^{12}\,cm^{-3}$. The average resonant optical depth is 22, with less than 20% variation over the measurement time. The probe and control beams are counter-propagating to reduce the residual Doppler broadening to $50 \times 2\pi$ kHz. Linearly polarized probe laser light enters the medium at an average photon rate of $1.6\,\mu s^{-1}$. Quarter- and half-wave plates at angles q and h, respectively, followed by a polarizing beam splitter, project the outgoing probe light onto a chosen polarization basis (Fig. 1a). Four single-photon counting modules measure the pair correlation events at times t_1 and t_2. Normalized second-order correlation functions, $g^{(2)}_{\alpha\beta}$, are calculated using the photon coincidence counts between the different detectors and the average count rates. The time bins (80 ns for $g^{(2)}_{\alpha\beta}(t_1,t_2)$ and 20 or 40 ns for $g^{(2)}_{\alpha\beta}(\tau)$) were chosen to capture the temporal dynamics of the correlation functions with reasonable signal-to-noise ratio.

In the quantum state tomography, we numerically optimize a Hermitian, positive-semidefinite two-photon density matrix

$$\rho = \begin{pmatrix} \rho_{++,++} & \rho_{++,S} & \rho_{++,--} & 0 \\ \rho_{S,++} & \rho_{S,S} & \rho_{S,--} & 0 \\ \rho_{--,++} & \rho_{--,S} & \rho_{--,--} & 0 \\ 0 & 0 & 0 & \rho_{A,A} \end{pmatrix}$$

in the two-qubit basis $\{|\sigma_1^+\sigma_2^+\rangle, |S\rangle, |\sigma_1^-\sigma_2^-\rangle, |A\rangle\}$, where $|S\rangle = (|\sigma_1^+\sigma_2^-\rangle + |\sigma_1^-\sigma_2^+\rangle)/\sqrt{2}$ and $|A\rangle = (|\sigma_1^+\sigma_2^-\rangle - |\sigma_1^-\sigma_2^+\rangle)/\sqrt{2}$. Because the two photons have the same frequency and spatial mode, there is no coherence between the 3×3 symmetric and 1×1 antisymmetric subspaces[28]. We measure in six required polarization bases, chosen as $\{q, h\} = \{\pi/4, \pi/4\}, \{0, 0\}, \{\pi/8, \pi/8\}, \{0, \pi/16\}, \{\pi/8, \pi/16\}$ and $\{\pi/8, 0\}$, to set the ten degrees of freedom in $\rho(t_1, t_2)$. The optimization follows the maximum-likelihood estimate[27], where all coincidence measurements are considered. The one-photon density matrix, $\rho^{(1)}(t)$, is reconstructed using the same technique. To extract the nonlinear phase from $\rho(t_1, t_2)$, we rescale for the linear dispersion and loss effects by defining the interaction matrix $\tilde{\rho}_{i,j}(t_1,t_2) = \rho_{i,j}(t_1,t_2) \big/ \left[\rho^{(1)}(t_1) \otimes \rho^{(1)}(t_2)\right]_{i,j}$ in the basis $\{|\sigma_1^+\sigma_2^+\rangle, |\sigma_1^+\sigma_2^-\rangle, |\sigma_1^-\sigma_2^+\rangle, |\sigma_1^-\sigma_2^-\rangle\}$. The interaction matrix generalizes the standard definition of $g^{(2)}$ to account for nonlinear phases and decoherence, and all its elements are equal to 1 in the absence of nonlinearity. In Supplementary Fig. 3, we compare the measured photon–photon correlation functions with those calculated from $\tilde{\rho}$ (see also Supplementary Fig. 4). The colour maps in Fig. 2 presenting values derived from $\rho(t_1, t_2)$ have been smoothed using an unweighted, nearest-neighbour, rectangular sliding average.

doi:10.1038/nature12541

Microscopic observation of magnon bound states and their dynamics

Takeshi Fukuhara[1], Peter Schauß[1], Manuel Endres[1], Sebastian Hild[1], Marc Cheneau[1,2], Immanuel Bloch[1,3] & Christian Gross[1]

The existence of bound states of elementary spin waves (magnons) in one-dimensional quantum magnets was predicted almost 80 years ago[1]. Identifying signatures of magnon bound states has so far remained the subject of intense theoretical research[2–5], and their detection has proved challenging for experiments. Ultracold atoms offer an ideal setting in which to find such bound states by tracking the spin dynamics with single-spin and single-site resolution[6,7] following a local excitation[8]. Here we use *in situ* correlation measurements to observe two-magnon bound states directly in a one-dimensional Heisenberg spin chain comprising ultracold bosonic atoms in an optical lattice. We observe the quantum dynamics of free and bound magnon states through time-resolved measurements of two spin impurities. The increased effective mass of the compound magnon state results in slower spin dynamics as compared to single-magnon excitations. We also determine the decay time of bound magnons, which is probably limited by scattering on thermal fluctuations in the system. Our results provide a new way of studying fundamental properties of quantum magnets and, more generally, properties of interacting impurities in quantum many-body systems.

The study of non-equilibrium processes in quantum spin models can provide fundamental insight into elementary aspects of magnetism. Magnons are the basic quasiparticle excitations around the ground state of ferromagnets and govern their low-temperature physics[9,10]. Owing to the ferromagnetic interaction, two spin excitations can remain bound together, forming a so-called two-magnon bound state[1,9,11]. In one and two dimensions, bound states exist for all centre-of-mass momenta, which prohibits the description of low-energy properties in terms of free magnon states[9]. In the classical limit, magnon bound states can be regarded as the basic building blocks of magnetic solitons[12,13]. The study of non-equilibrium dynamics in quantum spin chains is also important for a variety of applications. The evolution of two localized spin excitations realizes an interacting quantum walk[14,15] in the spin domain, which can be a versatile tool for the study of complex many-body systems[16]. It is also of importance in the context of quantum information[17], where transport properties in a one-dimensional chain of qubits can be strongly influenced by magnon bound states[18].

The spin-1/2 Heisenberg model is one of the foundational models of interacting quantum spins. The corresponding Hamiltonian was solved analytically in one dimension in the early 1930s by Bethe, who used a systematic Ansatz for the form of the eigenvectors[1]. Later, the Bethe Ansatz proved to be far more general and allowed the solving of many more one-dimensional problems, such as the Lieb–Liniger model or the fermionic Hubbard model[19]. Recent powerful extensions of this approach include the investigation of the dynamics of one-dimensional quantum many-body systems. One of the first results of Bethe's analysis was the prediction of magnon bound states. Experimentally, spectroscopic studies of solid-state materials provided the first evidence for the existence of such states[20–22]. For ultracold atoms in optical lattices, high-energy bound states have been observed in the form of repulsively bound atom pairs[23,24]. Optical lattice systems can also be used to realize the Heisenberg model with (in principle) tunable anisotropy[25,26], where bound states occur as low-energy excitations of the many-body system. Recent technological advances even allow for the *in situ* control and detection of atomic spins in these experiments[6,7,27].

In our system, we make use of a one-dimensional chain of bosonic atoms in an optical lattice. Starting from an initial Mott insulating state and a fully magnetized chain, we flip two neighbouring spins in the centre of the chain (see Fig. 1 and ref. 28). Making use of our detection technique, which has single-site and single-spin resolution[7], we are able to observe individual magnons and their bound states directly and to identify them by correlation measurements after letting the system evolve.

The system is described by the one-dimensional two-species single-band Bose–Hubbard Hamiltonian at unity filling. In the strong coupling limit, where the on-site interaction energy is much larger than the tunnelling matrix element, this Hamiltonian can be mapped onto the ferromagnetic spin-1/2 Heisenberg chain (also known as the XXZ spin-1/2 chain)[25,26]:

$$\hat{H} = -J_{\mathrm{ex}} \sum_i \left[\frac{1}{2} \left(\hat{S}_i^+ \hat{S}_{i+1}^- + \hat{S}_i^- \hat{S}_{i+1}^+ \right) + \Delta \hat{S}_i^z \hat{S}_{i+1}^z \right] \qquad (1)$$

where J_{ex} is the superexchange coupling and Δ is the anisotropy between the transverse and longitudinal spin couplings. The pseudo-spin operators are defined in terms of creation $\hat{a}_{\sigma,i}^\dagger$ and annihilation $\hat{a}_{\sigma,i}$ operators for a boson on site i with spin $\sigma = \uparrow, \downarrow$: $\hat{S}_i^+ = \hat{a}_{\uparrow,i}^\dagger \hat{a}_{\downarrow,i}$, $\hat{S}_i^- = \hat{a}_{\downarrow,i}^\dagger \hat{a}_{\uparrow,i}$ and $\hat{S}_i^z = \left(\hat{n}_{\uparrow,i} - \hat{n}_{\downarrow,i}\right)/2$, where the number operators $\hat{n}_{\sigma,i}$ count the bosons of the respective spin states on each lattice site. The transverse coupling (the first term of equation (1)) corresponds to the spin exchange between two neighbouring sites and results in the propagation of spin excitations, or magnons[28,29]. The longitudinal coupling describes a nearest-neighbour interaction between the spins, which favours ferromagnetic order for $J_{\mathrm{ex}}\Delta > 0$. This term is the origin of the magnon bound states: two flipped spins can lower their energy when located on neighbouring sites. For our scattering parameters, the Heisenberg interactions are almost isotropic; that is, $\Delta \approx 1$ (see Methods). We note that the Heisenberg model above can be mapped onto a spinless Fermi system with nearest-neighbour attractive interactions via the Jordan–Wigner transformation. In the non-interacting case ($\Delta = 0$), magnons therefore behave as free fermions.

Starting from a general wavefunction for the case of two flipped spins of the form

$$|\Psi\rangle = \sum_{1 \le i < j \le N} a(i,j) |i,j\rangle \qquad (2)$$

with $|i,j\rangle = \hat{S}_i^- \hat{S}_j^- |\ldots\uparrow\uparrow\uparrow\uparrow\ldots\rangle$ and N denoting the length of the chain, we use the Bethe Ansatz to obtain the eigenvalues and eigenvectors of the system (see Extended Data Fig. 1). The bound states can be identified from the corresponding energy spectrum through their separation from the scattering states, and the spatial extension of each bound state is revealed in the spin–spin correlations $\sum_i |a(i,i+d)|^2$ as a function of the distance d between the two spins (see Extended Data Fig. 1). This analysis reveals that our initial state $|\ldots\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\ldots\rangle$ has a

[1]Max-Planck-Institut für Quantenoptik, 85748 Garching, Germany. [2]Laboratoire Charles Fabry, Institut d'Optique Graduate School — CNRS — Université Paris Sud, 91127 Palaiseau, France. [3]Fakultät für Physik, Ludwig-Maximilians-Universität München, 80799 München, Germany.

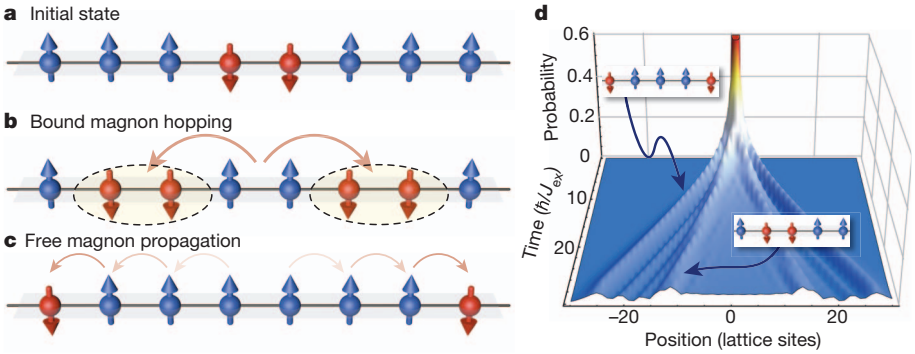


Figure 1 | Schematic representation of magnon propagation. **a**–**c**, Initially prepared state of 'up' spins (blue) with the two flipped 'down' spins shown in red (**a**), and its decomposition into bound magnons (**b**) and free magnons (**c**) propagating through the lattice. **d**, Numerical results obtained from exact diagonalization, showing the probability of finding a flipped spin at a given lattice site following the initial state preparation. Two different wavefronts corresponding to bound and free magnons can be identified (insets). Note that the maximum probability was clipped in the graph for clarity.

large overlap (~50%) with two-magnon bound states, the rest being shared among free magnon scattering states. We therefore expect both bound and free magnon dynamics to appear in the subsequent dynamical evolution after flipping two neighbouring spins (see Fig. 1).

The experiment started with the preparation of a two-dimensional degenerate gas of ^{87}Rb atoms in the $|\uparrow\rangle$ state in a single antinode of a vertical optical lattice (lattice spacing $a_{\mathrm{lat}} = 532$ nm). The spin degree of freedom was encoded in two hyperfine states with $|\uparrow\rangle \equiv |F = 1, m_F = -1\rangle$ and $|\downarrow\rangle \equiv |2, -2\rangle$. By ramping up one horizontal lattice, the gas was then split into approximately ten decoupled one-dimensional tubes of comparable length. The splitting was carried out in 120 ms with a final lattice depth of $30E_{\mathrm{r}}$, where $E_{\mathrm{r}} = h^2/(8ma_{\mathrm{lat}}^2)$ denotes the recoil energy and m is the atomic mass. Simultaneously, the lattice along the tubes was increased to $V = 20E_{\mathrm{r}}$, driving the system into the Mott insulating phase. In the next step, we applied a microwave-driven spin flip to the state $|\downarrow\rangle$ of two neighbouring atoms at the centre of each chain. For this spin flip, we used a line-shaped laser beam generated with a spatial light modulator that selectively shifted the addressed sites in resonance with the microwave radiation[28]. The addressing light was chosen at a wavelength and polarization such that the $|\uparrow\rangle$ states were unaffected while the $|\downarrow\rangle$ states were lowered in energy, and thus pinned at their positions. We then ramped down the lattice along the tubes to $V = 10E_{\mathrm{r}}$ in 50 ms and subsequently switched off the addressing beam within 1 ms. This marked the starting point of the dynamics. At this final lattice depth, the dynamics is sufficiently fast ($J_{\mathrm{ex}}/h = 2\pi \times 8.6$ Hz) compared to the typical heating time of several hundred milliseconds. After a variable evolution time, we rapidly ramped up all lattices to approximately $80E_{\mathrm{r}}$ to freeze out the dynamics. For state-selective detection, we applied a microwave sweep to invert the spin population, followed by a resonant laser pulse on the closed cycling transition to remove the $|2, -2\rangle$ majority component. We finally detected the atoms originally in the $|\downarrow\rangle$ state (now mapped to the remaining $|1, -1\rangle$ state) with single-site resolution[7].

We analysed the extracted atom positions in terms of a joint probability $P_{i,j}$ of simultaneously detecting atoms on lattice sites i and j along the tubes. Only data sets with exactly two atoms per tube were included. Approximately 50% of the data are discarded through this process, mainly because of the finite spin flip fidelity. We exclude $i = j$ from the analysis because the parity projection prevents the detection of two atoms on the same site[7]. In Fig. 2a, we show the resulting probability distributions. The bound-state population is directly reflected in the strong signal along the diagonals $j = i \pm 1$. The spread along this direction increases with evolution time, which is a signature of the correlated motion of the spin pair forming the bound state. To

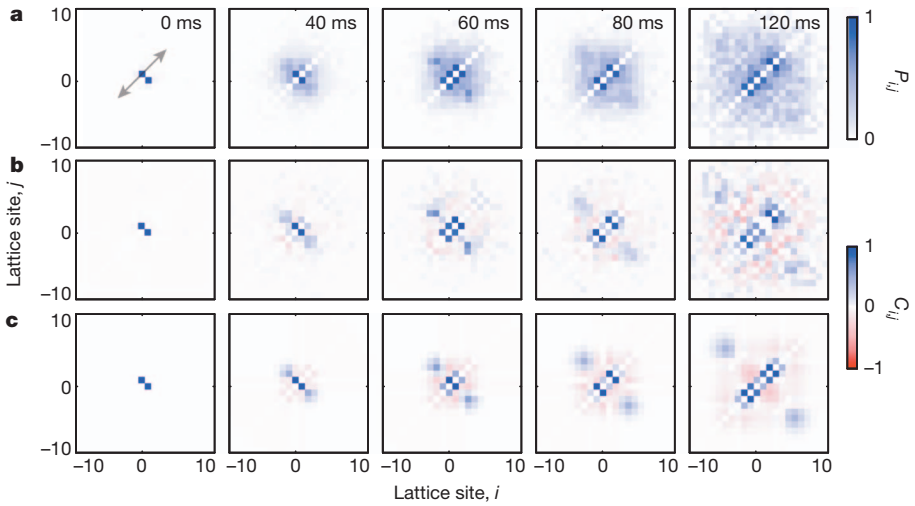


Figure 2 | Spatial correlations after dynamical evolution. **a**, Measured joint probability distributions $P_{i,j}$ of the position of the two spins for different evolution times, as indicated. The bound magnon signal and its spreading is visible on the diagonals $j = i \pm 1$ (arrow). In each image, the colour scale is normalized to the measured peak value. **b**, Corresponding correlation functions $C_{i,j} = P_{i,j} - P_iP_j$ of the measured data. The subtraction of uncorrelated detection events caused by finite-temperature effects and finite preparation fidelity gives better access to the correlation signal of the zero-temperature two-magnon evolution. For example, anti-bunching for the free magnons becomes visible, which is reflected in the outward propagating signal along the orthogonal diagonal. **c**, Numerical results for the correlations using exact diagonalization. Colour scales are normalized analogously to **a**. Note that the symmetry around the $j = i$ diagonal in all plots is caused by the indistinguishability of the two spins.

subtract uncorrelated detection events caused by finite-temperature effects and finite preparation fidelity (see Methods), we calculate the correlation function $C_{i,j} = P_{i,j} - P_i P_j$, where $P_i = \sum_j P_{i,j}$ is the probability of finding one atom on site i (see Fig. 2b). Next to the strong signal of bound magnons, a second feature is visible along the orthogonal diagonal. It corresponds to those free magnon states with which the prepared initial state has a finite overlap (see Extended Data Fig. 1). As we show below, the free magnon wavefront spreads at the predicted maximal free magnon velocity, $J_{ex}a_{lat}/\hbar$: the spins are at maximum separation. This anti-bunching behaviour can be understood intuitively from the above-mentioned mapping of the Heisenberg model to a fermionic Hamiltonian. Numerical results based on exact diagonalization of the Heisenberg chain (assuming zero temperature), shown in Fig. 2c, are in remarkable agreement with the experimental data. The bound magnon wavefront is also found to spread at the predicted maximal velocity, $J_{ex}a_{lat}/(2\hbar\Delta)$, owing to a singularity in the probability density of propagation velocities (see Methods and ref. 8).

To investigate the dynamics of the magnon bound states in more detail, we concentrate on the diagonals $j = i \pm 1$ in Fig. 2 and analyse the evolution of the normalized distribution $P_i^{\uparrow\uparrow} = P_{i,i+1} \big/ \sum_j P_{j,j+1}$ (Fig. 3a); that is, we use only data where two atoms on adjacent sites have been detected. We expect the magnon bound states to spread as compound objects almost freely across the lattice. We therefore extract the width w of the distributions $P_i^{\uparrow\uparrow}$ by fitting the data with Bessel functions of the first kind $[\mathcal{J}_i(w)]^2$ (see Methods and Extended Data Fig. 2). To measure the propagation velocity of the free magnon excitations, we analyse the correlations $\tilde{C}_d = \sum_i C_{i,i+d}$ as functions of distance, shown in Fig. 3b. Here, the correlation signal at $d = 1$ is due to the magnon bound states, while the second positive correlation signal, at a distance increasing with evolution time, is the free magnon contribution. We determine the position of the free magnons via Gaussian fits and define the wavefront as the centre plus one 1/e-width to take the dispersion into account (see Methods and Extended Data Fig. 3). Figure 3c shows the measured wavefront positions of both the free and the bound magnon states versus time. A linear fit yields velocity $v_f = 60(3)$ sites s^{-1} for the free magnons and $v_b = 26\left(^{+2+6}_{-2}\right)$ sites s^{-1} for the bound magnons, where the first uncertainty in v_b is due to the fit and the second takes into account a systematic underestimation of the bound-state velocity (see Methods). The ratio of the two velocities is $v_f/v_b = 2.3\left(^{+0.2}_{-0.7}\right)$, consistent with the predicted value $v_f/v_b = 2\Delta = 2$ for the isotropic case[8].

Above we analysed the data in the context of the isotropic Heisenberg chain and found good agreement with theoretical predictions. However, the experiment was not carried out at zero temperature, resulting in a finite density of particle or hole excitations (approximately 10%) in the atomic chain. We expect coupling of these thermal excitations to the magnon bound states to lead to a finite lifetime. To extract this lifetime, we plot the probability of finding two atoms on adjacent sites ($\sum_i P_{i,i+1}$) versus time in Fig. 4. For comparison, we show the zero-temperature prediction of the Heisenberg chain for the isotropic experimental case and for the case $\Delta = 0$, for which no bound states exist. Here, we take into account the finite preparation fidelity (87(1)%) for flipping the spin of two atoms at adjacent sites (see Methods). For long evolution times (Fig. 4 inset), the probability approaches zero for the non-interacting case ($\Delta = 0$), whereas it reaches a finite value of 38% in the isotropic model. This value is smaller than the overlap between our initial state and the magnon bound states because of the finite extension of the bound states beyond neighbouring sites (see Extended Data Fig. 1). We find the experimental data to lie in between the two scenarios (see Fig. 4). We fit the data with a heuristic model, which assumes the numerical prediction of the isotropic Heisenberg chain multiplied by an exponential decay. The extracted decay time of the bound magnon state is $\tau = 210(20)$ ms, where the uncertainty includes both the fitting error and the uncertainty in the numerical prediction. We believe this decay time to be determined by both thermal density fluctuations that are already present initially and technical heating during the evolution dynamics. It remains an interesting challenge for future theory to explain the lifetime due to the interaction of bound magnons with density fluctuations on the spin chain.

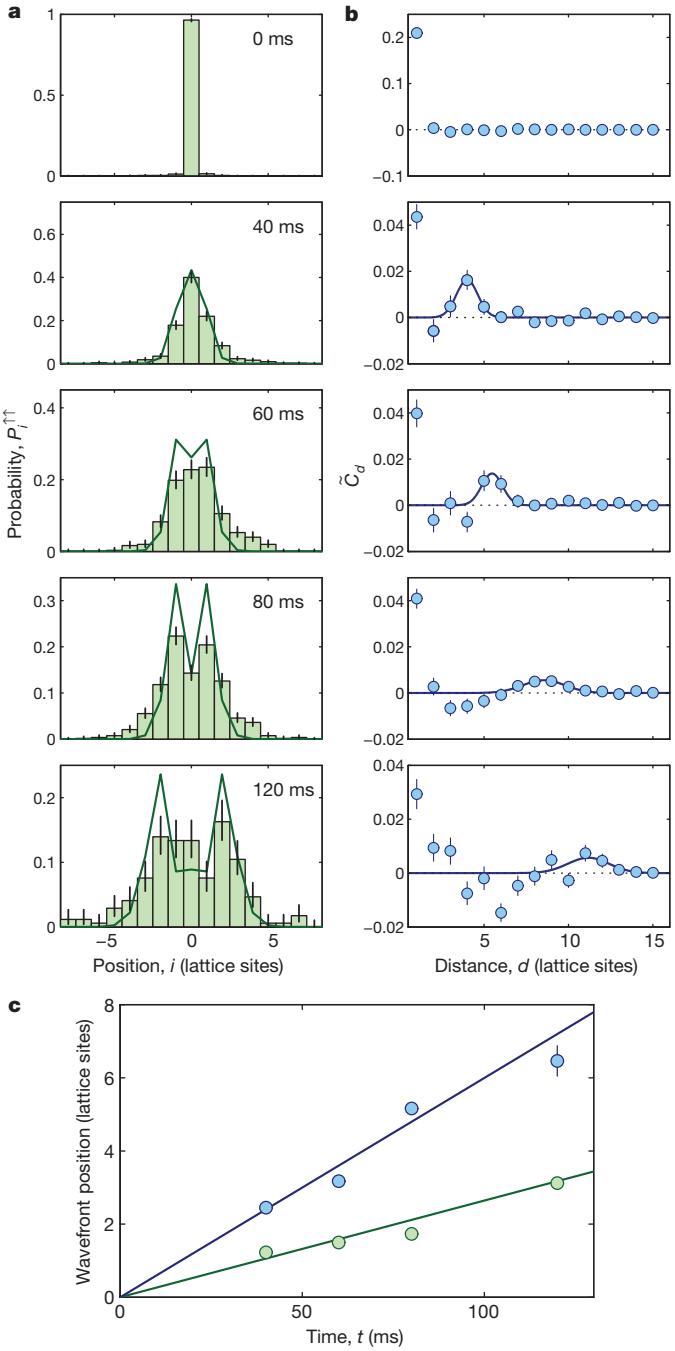


Figure 3 | Spreading wavefront velocity of bound and free magnons. **a**, Bound-state probability distributions $P_i^{\uparrow\uparrow}$ for different evolution times. The green bars show the experimental data. We extract the widths via Bessel function fits to the data (solid green lines). **b**, Propagation of the free magnons. The extracted correlation functions $\tilde{C}_d$ versus distance d are plotted for the same evolution times as used in **a** (blue circles). The signal at $d = 1$ is due to the bound states and the outward-moving peak stems from free magnons. The position and width of this peak are captured by Gaussian fits (dark blue lines). **c**, Comparing the propagation velocities. Linear regression of the extracted widths for the bound states (green) yields a velocity of $26\left(^{+2+6}_{-2}\right)$ sites s^{-1} compared to 60(3) sites s^{-1} for the wavefront of the free magnons (blue). Error bars, s.e.m.

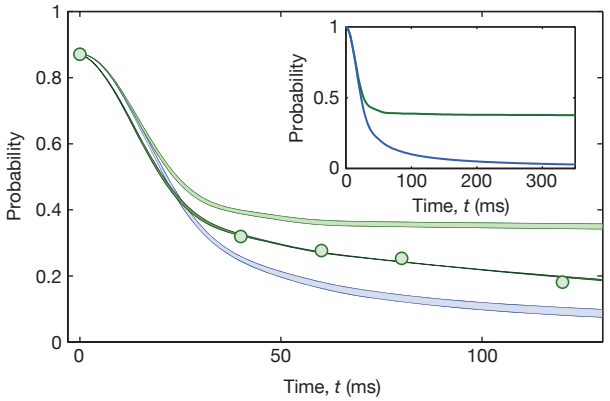


Figure 4 | Stability of the bound state. Probability of finding two spins at neighbouring sites as a function of the evolution time. Main figure, green circles are the experimental data and statistical error bars are smaller than the circles. We show numerical calculations (exact diagonalization) for the isotropic $\Delta = 1$ case (green shaded area) and for $\Delta = 0$ (blue shaded area), taking into account the preparation fidelity of 87% and the resulting uncertainty. The darker green line is a fit based on the isotropic numerical result multiplied by an exponential decay. Inset, numerical prediction for longer evolution times and without correcting for the preparation fidelity. The nearest-neighbour probability approaches zero in the $\Delta = 0$ case (blue line) whereas it converges to a finite value of 38% for $\Delta = 1$ (green line).

We have deterministically created a local excitation in a Heisenberg spin chain. We tracked the resulting dynamics microscopically and directly observed the spin correlations characteristic of magnon bound states and their evolution with time. This is, to our knowledge, the first realization of an interacting quantum walk in a magnetic spin chain. Future studies might address the question of the stability of magnon bound states in an environment containing thermal as well as stronger quantum fluctuations, or even the binding of two impurities in a superfluid environment, where a 'bipolaron' is expected to form. Other interesting extensions would be the study of universal Efimov physics using three magnons[30]. The reported results also pave the way towards the deterministic microscopic engineering of complex magnetic many-body states and the study of magnetic correlations in non-equilibrium situations.

METHODS SUMMARY

The general experimental procedure closely followed that of ref. 28. Additionally, for the long-evolution-time (120 ms) experiments, we used a vertically propagating, blue-detuned (~667 nm) beam to reduce the harmonic confinement in the horizontal plane. This enabled us to create larger Mott insulating plateaux with unity filling and thereby avoid reflections of the magnons from the boundaries of the atomic spin chains. We calculated the dynamics of the effective Heisenberg chain by directly diagonalizing the Hamiltonian. Because the number of magnons, or the total magnetization, is conserved, we considered only the Hilbert space containing two magnons. The numerical calculation was done for a superexchange coupling of $J_{ex}/h = 2\pi \times 8.6$ Hz, which has been estimated from $J_{ex} = 4J^2/U$. Here, the tunnelling matrix element J was obtained from the observation of the quantum walk of a single free atom, as shown in ref. 27. The on-site interaction energy U was obtained from an *ab initio* band-structure calculation using lattice depths, which were calibrated from amplitude modulation spectroscopy. More details, as well as the fits to extract the wavefront velocity, preparation fidelity, parameters of the Heisenberg model and probability density of propagation velocities, can be found in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 May; accepted 6 August 2013.
Published online 25 September 2013.

1. Bethe, H. A. Zur Theorie der Metalle. *Z. Phys.* **71,** 205–226 (1931).
2. Caux, J.-S. & Maillet, J. M. Computation of dynamical correlation functions of Heisenberg chains in a magnetic field. *Phys. Rev. Lett.* **95,** 077201 (2005).
3. Pereira, R. G., White, S. R. & Affleck, I. Exact edge singularities and dynamical correlations in spin-1/2 chains. *Phys. Rev. Lett.* **100,** 027206 (2008).
4. Kohno, M. Dynamically dominant excitations of string solutions in the spin-1/2 antiferromagnetic Heisenberg chain in a magnetic field. *Phys. Rev. Lett.* **102,** 037203 (2009).
5. Imambekov, A., Schmidt, T. L. & Glazman, L. I. One-dimensional quantum liquids: beyond the Luttinger liquid paradigm. *Rev. Mod. Phys.* **84,** 1253–1306 (2012).
6. Bakr, W. S. *et al.* Probing the superfluid-to-Mott insulator transition at the single-atom level. *Science* **329,** 547–550 (2010).
7. Sherson, J. F. *et al.* Single-atom resolved fluorescence imaging of an atomic Mott insulator. *Nature* **467,** 68–72 (2010).
8. Ganahl, M., Rabel, E., Essler, F. & Evertz, H. Observation of complex bound states in the spin-1/2 Heisenberg XXZ chain using local quantum quenches. *Phys. Rev. Lett.* **108,** 077206 (2012).
9. Wortis, M. Bound states of two spin waves in the Heisenberg ferromagnet. *Phys. Rev.* **132,** 85–97 (1963).
10. Takahashi, M. One-dimensional Heisenberg model at finite temperature. *Prog. Theor. Phys.* **46,** 401–415 (1971).
11. Hanus, J. Bound states in the Heisenberg ferromagnet. *Phys. Rev. Lett.* **11,** 336–338 (1963).
12. Fogedby, H. C. The spectrum of the continuous isotropic quantum Heisenberg chain: quantum solitons as magnon bound states. *J. Phys. C* **13,** L195–L200 (1980).
13. Schneider, T. Solitons and magnon bound states in ferromagnetic Heisenberg chains. *Phys. Rev. B* **24,** 5327–5339 (1981).
14. Schreiber, A. *et al.* A 2D quantum walk simulation of two-particle dynamics. *Science* **336,** 55–58 (2012).
15. Lahini, Y. *et al.* Quantum walk of two interacting bosons. *Phys. Rev. A* **86,** 011603(R) (2012).
16. Venegas-Andraca, S. E. Quantum walks: a comprehensive review. *Quant. Inf. Proc.* **11,** 1015–1106 (2012).
17. Bose, S. Quantum communication through spin chain dynamics: an introductory overview. *Contemp. Phys.* **48,** 13–30 (2007).
18. Subrahmanyam, V. Entanglement dynamics and quantum-state transport in spin chains. *Phys. Rev. A* **69,** 034304 (2004).
19. Batchelor, T. The Bethe ansatz after 75 years. *Phys. Today* **60,** 36–40 (2007).
20. Date, M. & Motokawa, M. Spin-cluster resonance in $CoCl_2 \cdot 2H_2O$. *Phys. Rev. Lett.* **16,** 1111–1114 (1966).
21. Torrance, J. B. & Tinkham, M. Excitation of multiple-magnon bound states in $CoCl_2 \cdot 2H_2O$. *Phys. Rev.* **187,** 595–606 (1969).
22. Hoogerbeets, R., van Duyneveldt, A. J., Phaff, A. C., Swüste, C. H. W. & de Jonge, W. J. M. Evidence for magnon bound-state excitations in the quantum chain system $(C_6H_{11}NH_3)CuCl_3$. *J. Phys. C* **17,** 2595–2608 (1984).
23. Winkler, K. *et al.* Repulsively bound atom pairs in an optical lattice. *Nature* **441,** 853–856 (2006).
24. Fölling, S. *et al.* Direct observation of second-order atom tunnelling. *Nature* **448,** 1029–1032 (2007).
25. Kuklov, A. & Svistunov, B. Counterflow superfluidity of two-species ultracold atoms in a commensurate optical lattice. *Phys. Rev. Lett.* **90,** 100401 (2003).
26. Duan, L.-M., Demler, E. & Lukin, M. Controlling spin exchange interactions of ultracold atoms in optical lattices. *Phys. Rev. Lett.* **91,** 090402 (2003).
27. Weitenberg, C. *et al.* Single-spin addressing in an atomic Mott insulator. *Nature* **471,** 319–324 (2011).
28. Fukuhara, T. *et al.* Quantum dynamics of a mobile spin impurity. *Nature Phys.* **9,** 235–241 (2013).
29. Trotzky, S. *et al.* Time-resolved observation and control of superexchange interactions with ultracold atoms in optical lattices. *Science* **319,** 295–299 (2008).
30. Nishida, Y., Kato, Y. & Batista, C. D. Efimov effect in quantum magnets. *Nature Phys.* **9,** 93–97 (2013).

Acknowledgements We thank H. G. Evertz, M. Haque, J.-S. Caux and W. Zwerger for discussions. We thank J. Zeiher for proofreading the manuscript. This work was supported by MPG, DFG, EU (NAMEQUAM, AQUTE, Marie Curie Fellowship to M.C.) and JSPS (Postdoctoral Fellowship for Research Abroad to T.F.).

Author Contributions All authors contributed extensively to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.F. (takeshi.fukuhara@mpq.mpg.de).

METHODS

Experimental procedure. The general experimental procedure closely followed that in ref. 28. Additionally, for the long-evolution-time (120 ms) experiments presented here, we used a vertically propagating, blue-detuned (~667 nm) beam to reduce the harmonic confinement in the horizontal plane. This enabled us to create larger Mott insulating plateaux with unity filling and thereby avoid reflections of the magnons from the boundaries of the atomic spin chains. The typical length of the chains was 20 sites with the deconfinement and 13 without. We generated this deconfinement beam by a broadband superluminescent diode to avoid possible interference of the beam with reflections from the vacuum window. The beam was successively amplified by two tapered amplifiers.

Extraction of the wavefront velocity from the fits. The validity of our method to extract the wavefront velocities was checked by analysing the results obtained from simulated data. To extract the velocity of the bound magnons, we used fits with Bessel functions. The Bessel function is not the exact distribution to describe the evolution of the bound magnons that we prepare, but, as we show below, it is suitable to capture the position of the wavefront in the distributions (see Extended Data Fig. 2a). The positions extracted from the simulated data for different times are shown in Extended Data Fig. 2b. In the long-time average, the resulting velocity agrees with the theoretical prediction ($J_{ex}a_{lat}/2\hbar$) for the bound magnons with $\Delta = 1$. For shorter, experimentally accessible times, the fits can underestimate the velocity by up to 20%. This is the reason for the systematic error in v_b given in the main text.

The free magnon velocity is extracted from the time evolution of the position and width of the outward moving peak in the correlation functions $\tilde{C}_d$ by using Gaussian fits: $A\exp[-(d-c)^2/s^2]$. To focus on the propagating peak, we exclude from the fit both the points at $d = 1$, which show the strong positive signal of the bound state, and the points with negative values. In Extended Data Fig. 3, the peak position c and the wavefront position $c + s$ are plotted. The wavefront velocity yields twice the velocity of a single free magnon because two free magnons propagate separately in opposite directions. The deviation of the extracted velocity from the maximal velocity $J_{ex}a_{lat}/\hbar$ of the single free magnon is found to be only 3%.

Preparation fidelity. The preparation fidelity for flipping the spin of two atoms on neighbouring sites is estimated to be 87%. This value is limited by two factors. First, the spin-flipping process might have addressed two spins initially separated by a larger distance. Second, the flipping process might have succeeded for one atom only, while the second atom observed is one from the majority component that was not removed during the push-out process because of its finite efficiency (98–99%). These two effects have different contributions to the probability of finding two atoms on neighbouring sites after the evolution time. In the first case, the effect can be calculated by solving the dynamics with the measured initial distributions ($P_{i,j}$ $(t = 0)$). For the second case, we can assume that falsely measured atoms are uniformly distributed over the chain (they were generated after the dynamics). The calculated probabilities shown in Fig. 4 of the main text take both these effects into account. The width of the shaded region displayed in Fig. 4 is due to the uncertainty of the ratio between the two effects.

Parameters of the Heisenberg model. The superexchange coupling J_{ex} and the anisotropy Δ are given by[25,26,31,32]:

$$J_{ex} = \frac{4J_\uparrow J_\downarrow}{U_{\uparrow\downarrow}}, \tag{3}$$

$$J_{ex}\Delta = \left(\frac{4J_\uparrow^2}{U_{\uparrow\uparrow}} + \frac{4J_\downarrow^2}{U_{\downarrow\downarrow}} - 2\frac{J_\uparrow^2 + J_\downarrow^2}{U_{\uparrow\downarrow}}\right). \tag{4}$$

Here J_σ are tunnelling matrix elements for a boson with spin $\sigma = \uparrow, \downarrow$ and $U_{\sigma\sigma'}$ are the on-site interaction energies between bosons with spins σ and σ'. In our case, the tunnelling matrix elements are spin independent ($J_\uparrow = J_\downarrow = J$), and the interaction energies are almost the same ($U_{\uparrow\uparrow} \approx U_{\downarrow\downarrow} \approx U_{\uparrow\downarrow} = U$). The anisotropy is $\Delta = 0.986$ for our ratios of the interaction energies ($U_{\uparrow\uparrow}$:$U_{\downarrow\downarrow}$:$U_{\uparrow\downarrow}$ = 100.4:99.0:99.0) that follow from the respective scattering lengths[33,34].

Numerical calculations using exact diagonalization. We calculated the dynamics of the effective Heisenberg chain by directly diagonalizing the Hamiltonian. Because the number of magnons, or the total magnetization, is conserved, we considered only the Hilbert space containing two magnons. The numerical calculation was done for a super-exchange coupling of $J_{ex}/\hbar = 2\pi \times 8.6$ Hz, which has been estimated from $J_{ex} = 4J^2/U$. Here, the tunnelling matrix element J was obtained from the observation of the quantum walk of a single free atom, as shown in ref. 27. The on-site interaction energy U was obtained from an *ab initio* band-structure calculation using lattice depths, which were calibrated from amplitude modulation spectroscopy. For the simulations, we used open boundary conditions and lattice sizes of 61 or 81 sites, depending on the evolution time, making sure that the magnons remain sufficiently far away from the edges to avoid spurious reflections.

Density of states and initial distribution of group velocities. Our initially prepared state $|\Psi_i\rangle$ can be decomposed into two parts:

$$|\Psi_i\rangle = |\Psi_b\rangle + |\Psi_f\rangle$$

that describe the overlap with magnon bound states ($|\Psi_b\rangle$) and free magnon scattering states ($|\Psi_f\rangle$). The bound magnon part is expanded in bound magnon eigenstates $|\psi_k\rangle$ with centre-of-mass wavevector k as:

$$|\Psi_b\rangle \propto \int_k \mathrm{d}k \langle\psi_k|\Psi_b\rangle|\psi_k\rangle \tag{5}$$

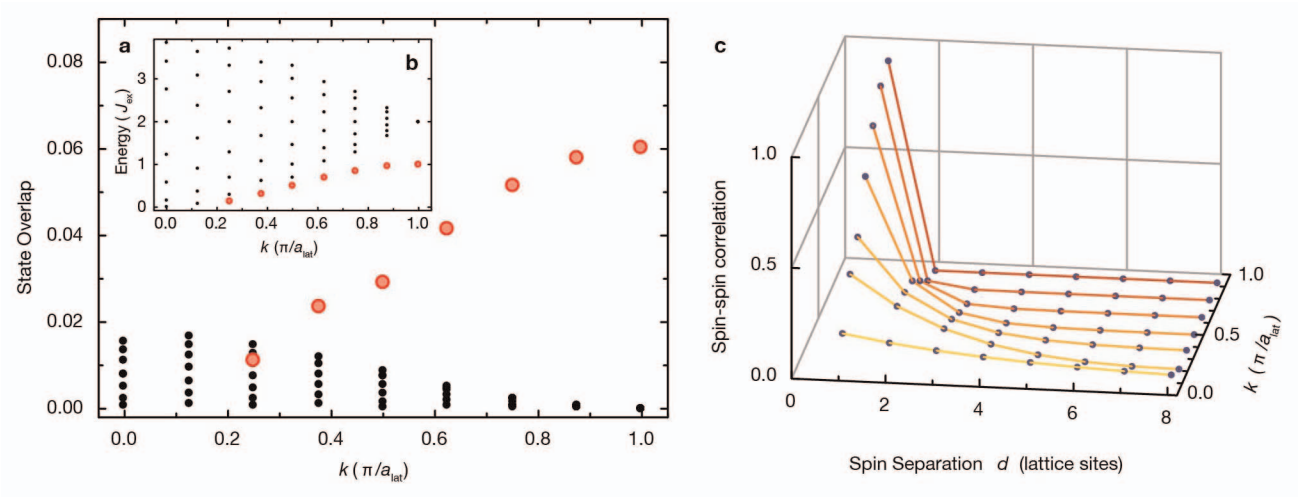
The probability density $P(v)$ of finding our initial state in magnon bound states that have a group velocity v can be written as:

$$\begin{aligned} P(v) &\propto \langle\Psi_b| \int_k \mathrm{d}k\delta\big(v - v_g(k)\big)|\psi_k\rangle\langle\psi_k||\Psi_b\rangle \\ &= \int_k \mathrm{d}k\delta\big(v - v_g(k)\big)|\langle\psi_k|\Psi_b\rangle|^2 \\ &= \frac{1}{\sqrt{v_{max}^2 - v^2}} \sum_{k_v} \big|\langle\psi_{k_v}|\Psi_b\rangle\big|^2 \end{aligned} \tag{6}$$

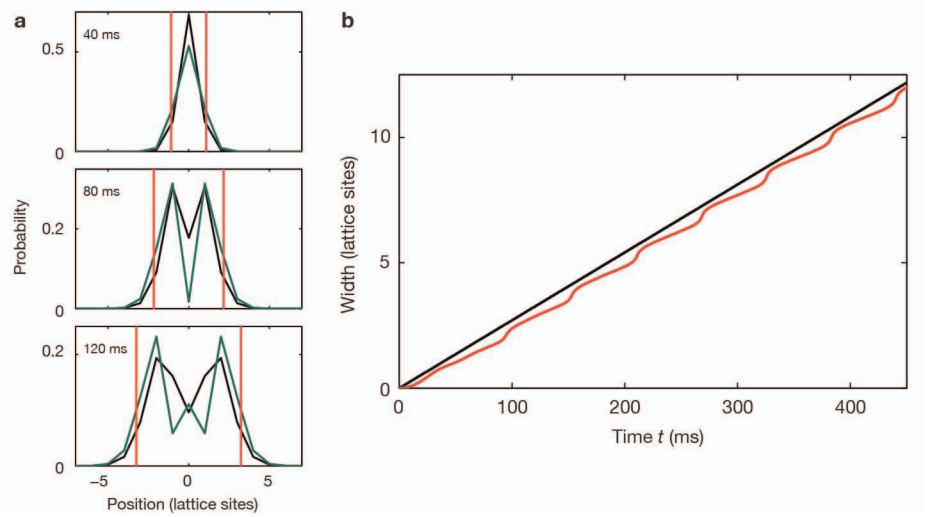
The group velocity is $v_g(k) = \frac{\mathrm{d}\varepsilon}{\hbar \mathrm{d}k} = \frac{J_{ex}a_{lat}}{2\hbar}\sin(ka_{lat})$ with the bound-state dispersion $\varepsilon = \frac{J_{ex}}{2}(1 - \cos(ka_{lat}))$ (refs 9, 35). The maximum group velocity is $v_{max} = J_{ex}a_{lat}/2\hbar$. The sum in the last line runs over all wavevectors k_v that yield a particular group velocity $v_g(k_v) = v$. The quantity $|\langle\psi_k|\Psi_b\rangle|^2$ describes the probability of finding the initial state in a magnon bound state with wavevector k and is plotted in Extended Data Fig. 1a. It is non-zero for the values $k = \pm\pi/2$ that yield the maximum group velocity $v_g(k) = \pm v_{max}$. Therefore, $P(v)$ shows a divergence for $v = \pm v_{max}$.

The quantity $\int_k \mathrm{d}k\ \delta\big(k - v_g(k)\big) \propto \frac{1}{\sqrt{v_{max}^2 - v^2}}$ essentially describes the density of states for a particular group velocity v (ref. 8). The singularity in this density of states is the origin of the singularity in $P(v)$.

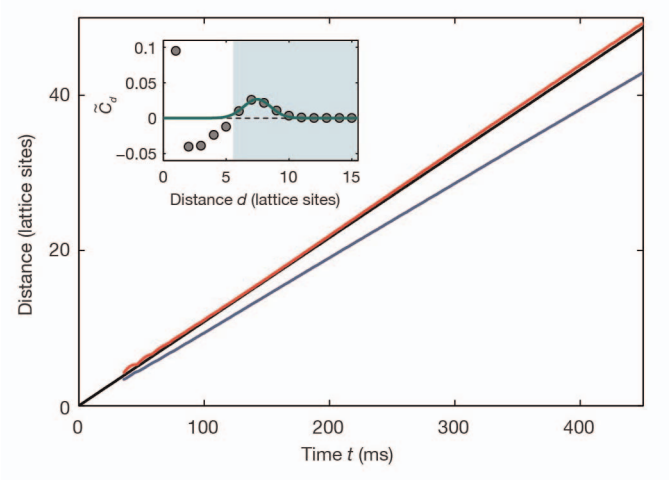
31. García-Ripoll, J. J. & Cirac, J. I. Spin dynamics for bosons in an optical lattice. *New J. Phys.* **5,** 76 (2003).
32. Altman, E., Hofstetter, W., Demler, E. & Lukin, M. D. Phase diagram of two-component bosons on an optical lattice. *New J. Phys.* **5,** 113 (2003).
33. Pertot, D., Gadway, B. & Schneble, D. Collinear four-wave mixing of two-component matter waves. *Phys. Rev. Lett.* **104,** 200402 (2010).
34. Hoefer, M. A., Chang, J. J., Hamner, C. & Engels, P. Dark-dark solitons and modulational instability in miscible two-component Bose-Einstein condensates. *Phys. Rev. A* **84,** 041605 (2011).
35. Karbach, M. & Müller, G. Introduction to the Bethe ansatz I. *Comput. Phys.* **11,** 36–43 (1997).



Extended Data Figure 1 | Quantum state analysis through the Bethe Ansatz. **a**, Overlap of the initial state with the bound (red circles) and free (black dots) magnon states calculated for $N = 16$ lattice sites[35]. For illustration, we show only the states with wavevectors within the interval $k \in [0,\pi/a_{\mathrm{lat}}]$. The inset **b** shows the corresponding energy spectrum. **c**, Spin–spin correlation of the magnon bound states as a function of the spin separation $d = |j - i|$ for different wavevectors k. For $k = \pi/a_{\mathrm{lat}}$ the wavefunction of the bound magnon state corresponds to tightly bound spins on neighbouring sites, giving the largest overlap with our initial state.



Extended Data Figure 2 | Propagation of bound magnons. **a**, Calculated probability distribution $P_i^{\uparrow\uparrow}$ (black lines) together with the Bessel function fit (green lines) for different evolution times (40, 80 and 120 ms). The red vertical lines show the width extracted from the fit. **b**, Determination of the velocity. The red line is the extracted width from the Bessel function fits for different evolution times. The black line corresponds to the expected maximum velocity ($J_{ex}a_{lat}/2\hbar$).



Extended Data Figure 3 | Propagation of free magnons. Main figure, the blue and red lines show the Gaussian centre c and the centre plus the width, $c + s$. Note that the centre moves more slowly than the maximum wavefront velocity. The black line, almost overlapping with the red line, corresponds to twice the expected single magnon maximum velocity ($2J_{ex}a_{lat}/\hbar$). Inset, an example of the Gaussian fit (green line). The grey circles represent the calculated correlation function $\tilde{C}_d$ for the evolution time of 80 ms. The blue shading highlights the region used for the fit.

doi:10.1038/nature12469

Three-dimensional imaging of localized surface plasmon resonances of metal nanoparticles

Olivia Nicoletti[1]*, Francisco de la Peña[1]*, Rowan K. Leary[1], Daniel J. Holland[2], Caterina Ducati[1] & Paul A. Midgley[1]

The remarkable optical properties of metal nanoparticles are governed by the excitation of localized surface plasmon resonances (LSPRs). The sensitivity of each LSPR mode, whose spatial distribution and resonant energy depend on the nanoparticle structure, composition and environment, has given rise to many potential photonic, optoelectronic, catalytic, photovoltaic, and gas- and bio-sensing applications[1–3]. However, the precise interplay between the three-dimensional (3D) nanoparticle structure and the LSPRs is not always fully understood and a spectrally sensitive 3D imaging technique is needed to visualize the excitation on the nanometre scale. Here we show that 3D images related to LSPRs of an individual silver nanocube can be reconstructed through the application of electron energy-loss spectrum imaging[4], mapping the excitation across a range of orientations, with a novel combination of non-negative matrix factorization[5,6], compressed sensing[7,8] and electron tomography[9]. Our results extend the idea of substrate-mediated hybridization of dipolar and quadrupolar modes predicted by theory, simulations, and electron and optical spectroscopy[10–12], and provide experimental evidence of higher-energy mode hybridization. This work represents an advance both in the understanding of the optical response of noble-metal nanoparticles and in the probing, analysis and visualization of LSPRs.

Applications of the optical properties of metal nanoparticles, including waveguides, light concentrators and resonators, single-molecule sensors, near-field scanning optical microscopy and surface-enhanced Raman spectroscopy[2,3,13], rely on local enhancement of the electromagnetic field by the plasmonic response. The spatial distribution and resonant energy of LSPR modes depend on the size, shape, composition and environment of the nanoparticle[3,11], and so far have been studied predominantly by optical spectroscopy and using optically based microscopies[3]. However, to reveal the intricate relationship between nanoparticle structure and LSPR, electron energy-loss spectroscopy (EELS) in a monochromated scanning transmission electron microscope (STEM) offers an unrivalled combination of spatial and energy resolution. The ability to provide nanometre resolution and to excite all LSPR modes strongly, including 'dark' modes in optical spectroscopy, has led to a greater understanding of the plasmonic response of a range of nanoparticles[14,15]. However, these EELS studies have yielded only two-dimensional maps of the intensity of the LSPRs, even from 3D nanostructures. We refer to these two-dimensional maps as EELS-LSPR maps. Here we retrieve key information pertaining to the often critical third dimension, by probing the nanoparticle at different orientations through the combination of EELS with electron tomography.

Metal nanocubes (metallic cubes with edge lengths varying from a few to several hundred nanometres) have particular optical properties that have stimulated studies by optical spectroscopy, simulations and EELS[10–12,16]. It has been observed[16] that LSPR modes of silver nanocubes supported on dielectric substrates hybridize[11,12,16]. Such hybridized modes are especially sensitive to changes in the nanoparticle structure and environment, and supported silver nanocubes are therefore seen as ideal candidates for use in, for example, sensor technology[11]. Here we study a silver nanocube[17] (Methods Summary) that is approximately 100 nm in size, has rounded corners of radii ~5 nm and has a (100) face in contact with a silicon nitride substrate (Fig. 1, insets). The amorphous silicon nitride substrate membrane, which is 30 nm thick, has a band gap of ~4 eV and therefore does not contribute spectral features in the range 1–4 eV that may mask the nanoparticle EELS-LSPRs. (Cerenkov radiation, and similar radiative losses, will be minimal for silicon nitride films of this thickness.) A tilt series of EELS spectrum images, 214 nm × 201 nm (152 pixels × 143 pixels) in area (Methods Summary), was acquired using a 300-kV electron beam at 0° and then at every 15°, from −60° to −15° and +15° to +60°, relative to a [100] cube axis perpendicular to the electron beam. The tilt increment was chosen as a compromise between obtaining sufficient sampling of the 3D object and limiting possible beam damage effects, build-up of carbonaceous contamination or both. Although we have acquired spectrum images across a ±60° tilt range, only the spectrum images on one side of 0° need to be considered in the analysis, because the C_{4v} ($4mm$) symmetry of the cube–substrate system ensures that those on the other side of 0° are related by mirror symmetry. By using only the first half of the recorded tilt series, effects of damage or contamination on the analysis are minimized. In Fig. 1a, we show representative EELS spectra from the silver nanocube, recorded at tilts of 0° (top) and −60° (bottom).

EELS-LSPR maps are determined conventionally by integrating the signal in the energy-loss range of interest[18,19] or by fitting peaks (normally Gaussians or Lorentzians) within a spectral range[14] and plotting the area under the peak. However, in most cases the EELS spectrum in the low-loss range consists of a superposition of partly overlapping peaks, whose width and asymmetry increases with retardation[20], and so an approach that better separates the spectral components is needed. If the EELS-LSPR signal can be described as a linear combination of spectral components (Methods), blind source separation methods can offer a better means of obtaining EELS-LSPR maps[21,22].

For our data, which has a significant noise level, we found that non-negative matrix factorization[5,6] (NMF) provided more distinct EELS-LSPR maps than did conventional methods (Methods). NMF performs an approximate factorization of a positive matrix, the experimental data set, into two positive matrices: in our case spectral components and spatial distribution maps. We determined that a factorization using eight components was sufficient to model accurately the complete data set in the 1–4 eV energy range of interest. In our case, five components are related to surface plasmon excitations (Fig. 1b, coloured areas labelled α, β, γ, δ and ε), one is related to bulk excitation at 3.8 eV (black area) and two (grey line) are related to a combination of the zero-loss peak (ZLP) tail, low-energy LSPRs and the effects of contamination[12]. For completeness, we include energy-filtered maps in Extended Data Fig. 2.

An example of the NMF spectral components, weighted according to the value for a single pixel from the spatial distribution maps, is shown in Fig. 1b (coloured areas α–ε). Further details of the spectral analysis are given in Methods. Each surface plasmon component

[1]Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK. [2]Department of Chemical Engineering and Biotechnology, University of Cambridge, Pembroke Street, Cambridge CB2 3RA, UK.
*These authors contributed equally to this work.

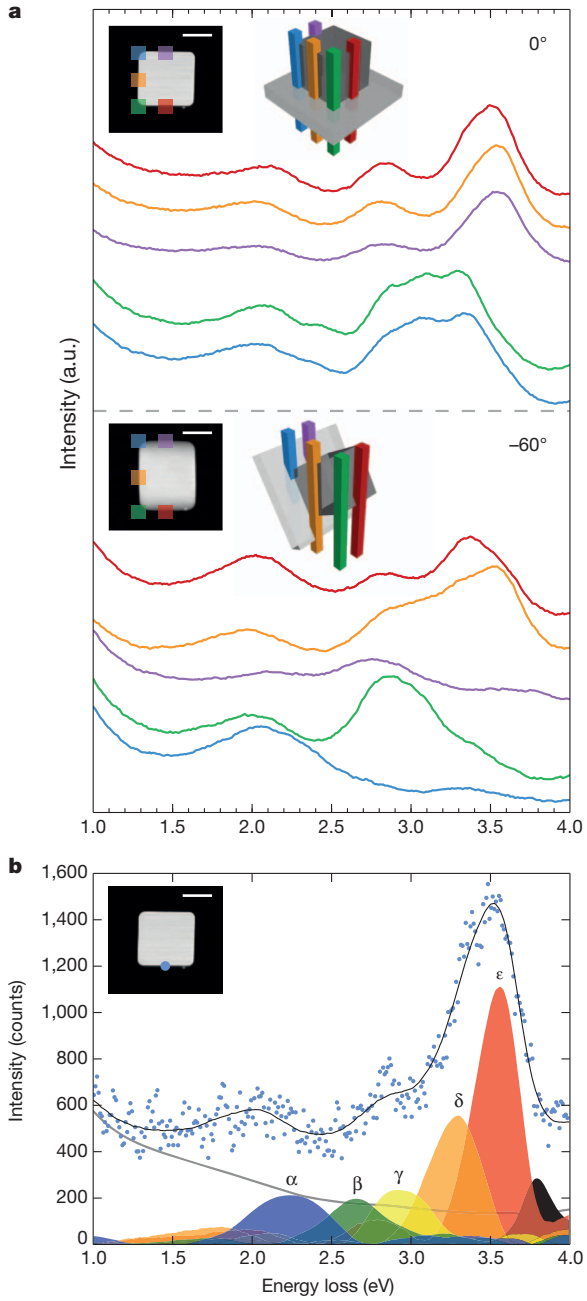


Figure 1 | Selected-area spectra of a silver nanocube and NMF model. **a**, Selected-area spectra (unprocessed) corresponding to the areas highlighted in the insets, acquired at 0° tilt (top) and at −60° tilt (bottom). The selected-area squares, and square prisms, in the schematics, highlight 25.4 nm × 25.4 nm (18 × 18 pixels) areas from which individual spectra have been extracted and summed. Insets, high-angle annular dark-field STEM images of a silver nanocube resting on a 30-nm-thick silicon nitride membrane at 0° tilt (top) and at −60° tilt (bottom). a.u., arbitrary units. **b**, NMF spectral components weighted according to the value for a single pixel from the spatial distributions maps, at the position indicated by the large blue marker in the inset. The unprocessed spectrum (blue dots) is shown together with the five spectral components, labelled α, β, γ, δ and ε, related to the LSPRs (areas in blue, green, yellow, orange and red), the one spectral component related to the silver volume plasmon at 3.8 eV (black area) and a combination of spectral components related to the ZLP tail, low-energy dipole loss and effects of contamination (grey line). The sum of the eight spectral components is also shown (black line). Scale bars, 50 nm.

(Fig. 2a) is dominated by a single peak with weak subsidiary features likely to originate in residual mixing of the components, lower-energy LSPRs and spectrometer imperfections. The corresponding normalized spatial distributions (the EELS-LSPR maps) thus obtained are shown in Fig. 2b.

In principle, an infinite number of LSPR modes exist for a cube (consistent with the underlying point group symmetry[23,24]) but their energy and/or spatial degeneracy, natural linewidth and intensity, and the limited signal-to-noise ratio, restrict the number of peaks that can be observed in an experimental spectrum. Indeed, most far-field optical spectroscopy studies of nanocubes supported on substrates[10,16] observe only two major peaks, the 'proximal' and 'distal' modes, so called[10] because simulations have shown[10–12,16] that the induced field is concentrated at the cube corners adjacent to, and far from, the substrate, respectively. In Fig. 2b, we observe that, at 0° tilt, components α, β and γ are localized at the four corners of the cube projection. Tilting reveals that α is especially intense at the bottom corners (near the substrate) and that the same is true of β and γ at the top corners; this is direct experimental visualization of a substrate-induced top–bottom splitting of modes. In addition, the tilt series suggests that δ and ε are strong at the edges and faces, respectively, in agreement with simulations[12]. We note also that additional weak excitation is apparent at the bottom edges in the β tilt series and at the bottom face in the γ series. Although the tilt series is instructive, a more powerful way to represent the data would be valuable as a means of better interpreting the intensity distribution of the EELS-LSPR maps.

The electron energy-loss probability due to plasmon resonances for a fast electron travelling with constant velocity v along a straight-line trajectory $r = (R_0, z)$, where $R_0 = (x_0, y_0)$ defines the impact parameter of the trajectory and z is the direction of travel, can be written as[20]

$$\Gamma_{\mathrm{EELS}}(R_0,\omega) = \frac{e}{\pi\hbar\omega|v|}\int_{-\infty}^{\infty} \mathrm{d}z\,\mathrm{Re}\left[\mathrm{e}^{-i\omega z/|v|}\, v \bullet E_{\mathrm{ind}}(r,\omega)\right] \qquad (1)$$

where $E_{\mathrm{ind}}(r, \omega)$ represents the induced electric field vector acting back on the electron and $\hbar\omega$ is the energy loss ($\hbar$, Planck's constant divided by 2π; ω, angular frequency). Directly reconstructing a (vector) electric field from multiple tilt series of EELS-LSPR maps may be theoretically possible[25], but it is experimentally challenging. We note that because the induced field depends on the trajectory, for a more conventional vector tomography approach to be used, the excitation and probing processes should be uncoupled, as is the case in electron energy-gain spectroscopy[15]. Here, for this proof-of-principle experiment, we instead use a single tilt series but exploit the C_{4v} ($4mm$) symmetry of the cube–substrate system, and the energy degeneracy of symmetry-equivalent modes, to compensate approximately for the effects of the trajectory dependency of the induced electric field vector in equation (1). As explained further in the Methods section, for singly degenerate modes this compensation is absent, resulting in tomographic reconstructions with negative intensity (see mode C_3 in Extended Data Fig. 3). In addition, for localized, low-energy modes of isolated, optically isotropic nanoparticles, such as in this case, we can make the approximation that the charge distribution of the electron beam takes the form of a 'charged wire'[26], and that the change in the phase term, $\mathrm{e}^{-i\omega z/|v|}$, in equation (1) across the spatial extent of the mode is small enough that its effect on the reconstruction is minimal (Extended Data Fig. 3). Furthermore, it has been shown recently[27] that, in the quasistatic limit, the LSPR loss probability, $\Gamma^{\mathrm{SP}}(R_0,\omega)$, can be written as a linear combination of eigenspectra, each corresponding to a mode i, composed of a spectral weighting factor $\mathrm{Im}(-g_i(\omega))$ and spatially varying basis functions $|\phi_i(R_0,q)|^2$ (Methods):

$$\Gamma^{\mathrm{SP}}(R_0,\omega) = C\sum_i \mathrm{Im}(-g_i(\omega))|\phi_i(R_0,q)|^2 \qquad (2)$$

Here $\phi_i(R_0,q)$ is the Fourier transform of the induced potential associated with mode i, C is a constant, $q = \omega/|v|$ and $g_i(\omega)$ is the response function.

Assuming that a modal decomposition of the EELS-LSPR signal is possible in the retarded case, as is true for spheres[28], using NMF we

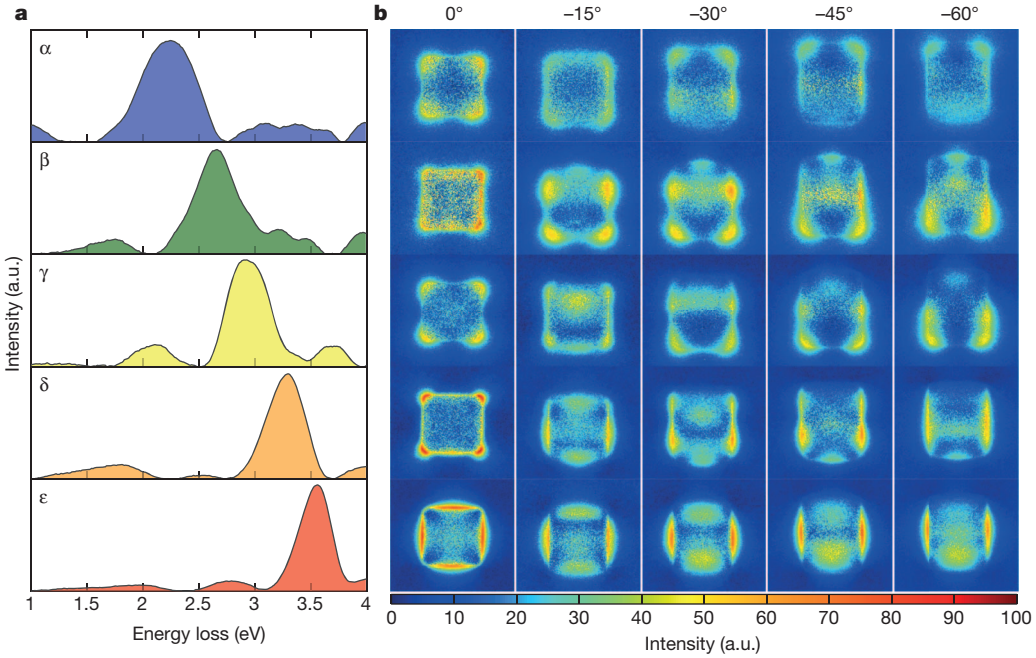


Figure 2 | EELS maps of LSPR components of a silver nanocube. **a**, LSPR spectral components (α, β, γ, δ and ε) resulting from applying NMF to the spectrum images at different tilt angles, in order of increasing energy loss. **b**, Normalized EELS maps corresponding to the five NMF components shown on the left. In the tilted images, the substrate is closer to the observer in the top half of the image. See Methods for details of the normalization process.

undertake an analogous decomposition of the EELS spectra to derive 'spectral components' and 'spatial distribution' EELS-LSPR maps. Each EELS-LSPR map may then be regarded as if it were a projection of a scalar quantity, and it is possible for a conventional tomographic reconstruction approach to be taken. Within these approximations, this scalar quantity may be related to the magnitude of the potential as seen by the electron that induces it. Using simulated EELS spectra, we have verified that these approximations hold remarkably well for our cube system (Methods).

For a 3D visualization of the LSPR modes, we used a compressed sensing[7,8] tomographic algorithm that can produce high-fidelity electron tomography reconstructions even with relatively few measurements[9]. Compressed sensing theory asserts that when a signal, or its representation in some transform domain, is sparse, or is well approximated as being sparse, relatively few incoherent measurements are needed to effect its recovery. Here we used the EELS-LSPR maps (Fig. 2b) as the input to the compressed sensing electron tomography algorithm, and enforcement of sparsity of the reconstructed 3D LSPR intensity in a wavelet domain during the reconstruction process. The C_{4v} (4*mm*) symmetry of the cube–substrate system was imposed during the reconstruction process. For the reconstructions (Fig. 3), we estimate the spatial resolution to be ~15 nm in each dimension. A data set with more images may provide improved resolution, but this will ultimately be limited by the delocalized nature of the surface plasmon excitation. Further details of the reconstruction process can be found in Methods. Supplementary Videos 1 and 2 show a 3D visualization of the LSPRs.

Figure 3 shows 3D reconstructions of the EELS-LSPR maps of Fig. 2b. In agreement with our discussion of the two-dimensional EELS-LSPR maps, the α component can be described as a bottom-corner LSPR, β as a mixture of top-corner and bottom-edge LSPRs, γ as a mixture of top-corner and bottom-face LSPRs, δ as top- and side-edge LSPRs and ε as top- and side-face LSPRs. It is worth noting the complementarity of the features: the bottom-edge and -face LSPRs found in β and γ are the 'missing' resonances in δ and ε, respectively, damped and redshifted by ~0.7 eV. It is also notable that the spatial distribution of the side edges in δ seems to be shifted slightly away from the midpoint of the cube edge towards the substrate.

To verify the validity of the NMF analysis, and assist in the interpretation of the 3D EELS-LSPR visualization, we simulated EELS spectra for the cube–substrate system using discrete dipole approximation (DDA) code[29] (Methods). Spectra (Fig. 4) were simulated as a function of electron probe position with trajectories 5 nm away from, and parallel to, the top and side faces of a 100-nm silver cube with rounded corners. Figure 4a shows the results for a cube *in vacuo* and Fig. 4b show those for a cube on a 30-nm-thick silicon nitride substrate. Our simulations are consistent with EELS spectra calculated previously[12]. For the cube in Fig. 4a, the series of spectra shows four main peaks labelled C_1, C_{2+3}, E and F. Of these, C_1 is a first-order dipole corner LSPR[23,24] with an energy-loss peak redshifted and broadened as a result of retardation effects[11,20]. The intensity distribution of C_{2+3} is consistent with corner LSPRs composed of a mix of (near-degenerate) dipole, quadrupole and octupole modes[11,12], and those of E and F with edge and face LSPRs, respectively (Fig. 4c). In Fig. 4b, we see that the introduction of a substrate breaks the top–bottom symmetry of the resonances. The corner dipole C_1 is redshifted (~0.3 eV) by the addition of the substrate, as expected. The C_{2+3} LSPR splits into two distal (labelled with a superscript 'd') components (C_2^d and C_3^d) and two proximal (labelled with a superscript 'p') components (C_2^p and C_3^p). These four components arise from the substrate-mediated hybridization of the (near-degenerate) dipole, quadrupole and octupole modes[23,24] giving rise to bonding (proximal) and anti-bonding (distal) modes. In a similar fashion, the E LSPR splits into distal and proximal components (E^d and E^p). The excitations of the face components on the top and bottom of the cube are too weak to be seen in the simulations for the trajectories used. The strong excitation at the side edges and faces becomes slightly asymmetric with respect to the cube centre, with the peak excitation nearer the substrate, as seen in the experiment. The energies of the peaks present in Fig. 4a, b are reported in Extended Data Table 1.

The simulations allow the features seen in the reconstructed LSPRs in Fig. 3 to be identified unambiguously. Starting at high energies, the ε component can be identified as the distal-face component, F^d; δ can be identified as the distal-edge component, E^d; and the bottom edges of β can be identified as E^p. The bottom face excitation of γ is the proximal-face component, F^p, although it does not appear in the simulation as

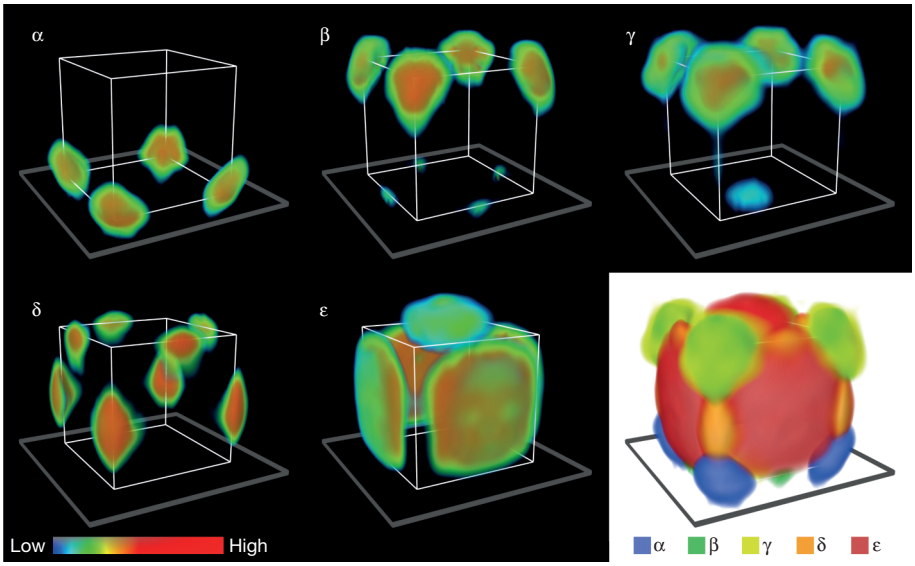


Figure 3 | 3D visualization of the LSPR components of a silver nanocube. The 3D images (α–ε) were obtained by tomographic reconstruction of the EELS maps of the respective LSPR components in Fig. 2. The visualizations are voxel projections of the reconstructed 3D volumes. The colour bar indicates the LSPR intensity. The image in the bottom right of the figure shows a combined 3D rendering of all the components. See also Supplementary Videos 1 and 2.

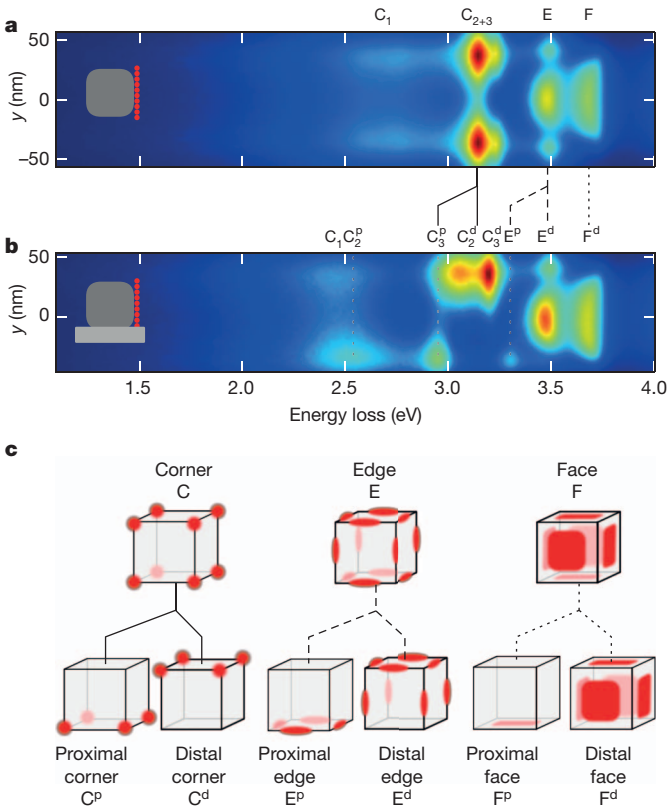


Figure 4 | DDA EELS simulations of a silver nanocube. a, **b**, DDA EELS simulations of line spectra as functions of probe position (red dotted line) 5 nm from the side face of a 100-nm cube with rounded corners (radius, ~19 nm) in vacuum (**a**) and on 30-nm-thick silicon nitride substrate (**b**). **c**, Schematic representation of substrate-induced hybridization probed in **b**: corner, edge and face modes of an isolated cube (top), and splitting of the corner, edge and face modes of an isolated cube in a proximal component (close to substrate) and a distal component (away from substrate) (bottom). The DDA EELS simulations were obtained using the code in ref. 29. The modelling dielectric functions were taken from ref. 33 for silver and from ref. 34 for silicon nitride.

discussed above. In general, and especially at lower energies, experimental LSPR peak energies are lower than predicted by simulation (Methods), as seen previously[12,29]. This is likely to result from a combination of factors including parameterization of dielectric functions, the degree of roundedness of the cube and the effect of contamination, native silver oxide or both changing the local dielectric environment. The dipole C_1 was not identified in the NMF analysis because experimentally the resonance is broad, spatially diffuse and at low energy, all of which leads to a signal-to-noise ratio too low to allow C_1 to be decomposed into a single component. As a result, it was incorporated into the 'background' curve shown in grey in Fig. 1b. Finally, to achieve consistency between simulation and experiment, we can associate the α component with C_3^p and the top-corner excitations of β and γ with C_2^d and C_3^d, respectively.

We have experimentally studied the 3D spatial distribution of LSPRs of a silver nanocube supported on a dielectric substrate. The application of NMF to a tilt series of spectrum images, acquired using monochromated STEM EELS, has enabled the identification of distinct LSPR components. Through a series of approximations, shown to be valid for our cube system, compressed sensing electron tomography was used to reconstruct high-fidelity 3D images of EELS-LSPRs, and allowed a direct visualization of substrate-mediated hybridization of modes with corner, edge and face excitations. This first step into visualizing LSPRs in three dimensions should motivate further 3D imaging studies, the inclusion of relativistic effects and the development of vector electron tomography methods allowing reconstructions from arbitrary-shaped nanoparticles. These techniques will enable a deeper understanding of the optical properties of many metallic nanostructures, including colloidal nanoparticles, lithographically produced structures and self-assembled metamaterials, and should lead to the exploitation of 3D LSPR morphology in future applications.

METHODS SUMMARY

Sample preparation. Transmission electron microscope samples of silver nanocubes, made by polyol synthesis[17] (Nano Research Facility, Washington University in St Louis), were prepared by depositing drops of a silver nanoparticle solution on a 30-nm-thick silicon nitride membrane (Agar Scientific).

Data acquisition. An FEI Titan 60-300 X-FEG (S)TEM was operated at 300 kV with the monochromator in decelerating mode at an excitation of 0.8. The beam convergence semi-angle was 8 mrad. EELS spectra were recorded on a Gatan GIF Quantum ERS energy-loss spectrometer with a collection angle of 32 mrad, a

dispersion of 0.01 eV per channel and a dwell time of 5 ms, and using a 260 × 2,048 pixel section of a 2,048 × 2,048 pixel charge-coupled device camera, with vertical binning of 1 × 5. The energy resolution (full-width at half-maximum of the ZLP) of the EELS spectra was 170 meV.

Spectral processing. Spectral processing was performed using the open-source software HYPERSPY[22] (formerly EELSLAB). After X-ray spikes were removed and the ZLP aligned, the spectrum images were scaled to normalize the Poisson noise[30] and factorized using a projected gradient method NMF algorithm[31]. The NMF was repeated for different numbers of components, ranging from four to twelve, and showed that eight components were optimal. EELS maps were divided by the ZLP intensity pixel by pixel to obtain a magnitude proportional to the excitation probability. The scaling factor of each spectral component was chosen by normalizing to unity the area of the spectrum of each component.

Simulation. EELS DDA simulations were performed using DDEELS v1.07alpha[29]. We used 4,701 dipoles to model a 100-nm, rounded silver cube (3,645 dipoles for the 150 nm × 150 nm × 30 nm Si_3N_4 substrate).

3D reconstruction. The compressed sensing electron tomography reconstruction algorithm is described elsewhere[9] and only minor modifications were made for the reconstructions performed here. For the 'sparsifying' transform, a Coiflet wavelet transform was used. The reconstruction was performed in MATLAB, using the conjugate gradient descent algorithm of ref. 32.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 April; accepted 15 July 2013.

1. Maier, S. A. & Atwater, H. A. Plasmonics: localization and guiding of electromagnetic energy in metal/dielectric structures. *J. Appl. Phys.* **98,** 011101 (2005).
2. Schuller, J. A. *et al.* Plasmonics for extreme light concentration and manipulation. *Nature Mater.* **9,** 193–204 (2010).
3. Anker, J. N. *et al.* Biosensing with plasmonic nanosensors. *Nature Mater.* **7,** 442–453 (2008).
4. Jeanguillaume, C. & Colliex, C. Spectrum-image: the next step in EELS digital acquisition and processing. *Ultramicroscopy* **28,** 252–257 (1989).
5. Paatero, P. & Tapper, U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5,** 111–126 (1994).
6. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).
7. Donoho, D. L. Compressed sensing. *IEEE Trans. Inf. Theory* **52,** 1289–1306 (2006).
8. Candes, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52,** 489–509 (2006).
9. Leary, R., Saghi, Z., Midgley, P. A. & Holland, D. J. Compressed sensing electron tomography. *Ultramicroscopy* **131,** 70–91 (2013).
10. Ringe, E. *et al.* Unraveling the effects of size, composition, and substrate on the localized surface plasmon resonance frequencies of gold and silver nanocubes: a systematic single-particle approach. *J. Phys. Chem. C* **114,** 12511–12516 (2010).
11. Zhang, S., Bao, K., Halas, N. J., Xu, H. & Nordlander, P. Substrate-induced Fano resonances of a plasmonic nanocube: a route to increased-sensitivity localized surface plasmon resonance sensors revealed. *Nano Lett.* **11,** 1657–1663 (2011).
12. Mazzucco, S. *et al.* Ultralocal modification of surface plasmons properties in silver nanocubes. *Nano Lett.* **12,** 1288–1294 (2012).
13. Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R. R. & Feld, M. S. Surface-enhanced Raman scattering and biophysics. *J. Phys. Condens. Matter* **14,** R597 (2002).
14. Nelayah, J. *et al.* Mapping surface plasmons on a single metallic nanoparticle. *Nature Phys.* **3,** 348–353 (2007).
15. Yurtsever, A., van der Veen, R. M. & Zewail, A. H. Subparticle ultrafast spectrum imaging in 4D electron microscopy. *Science* **335,** 59–64 (2012).
16. Sherry, L. J. *et al.* Localized surface plasmon resonance spectroscopy of single silver nanocubes. *Nano Lett.* **5,** 2034–2038 (2005).
17. Korte, K. E., Skrabalak, S. E. & Xia, Y. Rapid synthesis of silver nanowires through a CuCl-or $CuCl_2$-mediated polyol process. *J. Mater. Chem.* **18,** 437–441 (2008).
18. Nelayah, J. *et al.* Direct imaging of surface plasmon resonances on single triangular silver nanoprisms at optical wavelength using low-loss EFTEM imaging. *Opt. Lett.* **34,** 1003–1005 (2009).
19. Nicoletti, O. *et al.* Surface plasmon modes of a single silver nanorod: an electron energy loss study. *Opt. Express* **19,** 15371–15379 (2011).
20. García de Abajo, F. J. Optical excitations in electron microscopy. *Rev. Mod. Phys.* **82,** 209–275 (2010).
21. Comon, P. *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (Academic, 2010).
22. de la Peña, F. *et al.* Mapping titanium and tin oxide phases using EELS: an application of independent component analysis. *Ultramicroscopy* **111,** 169–176 (2011).
23. Langbein, D. Normal modes at small cubes and rectangular particles. *J. Phys. Math. Gen.* **9,** 627–644 (1976).
24. Fuchs, R. Theory of the optical properties of ionic crystal cubes. *Phys. Rev. B* **11,** 1732–1740 (1975).
25. Lade, S. J., Paganin, D. & Morgan, M. J. Electron tomography of electromagnetic fields, potentials and sources. *Opt. Commun.* **253,** 392–400 (2005).
26. Hohenester, U., Ditlbacher, H. & Krenn, J. R. Electron-energy-loss spectra of plasmonic nanoparticles. *Phys. Rev. Lett.* **103,** 106801 (2009).
27. Boudarham, G. & Kociak, M. Modal decompositions of the local electromagnetic density of states and spatially resolved electron energy loss probability in terms of geometric modes. *Phys. Rev. B* **85,** 245447 (2012).
28. García de Abajo, F. J. Relativistic energy loss and induced photon emission in the interaction of a dielectric sphere with an external electron beam. *Phys. Rev. B* **59,** 3095–3107 (1999).
29. Geuquet, N. & Henrard, L. EELS and optical response of a noble metal nanoparticle in the frame of a discrete dipole approximation. *Ultramicroscopy* **110,** 1075–1080 (2010).
30. Keenan, M. R. & Kotula, P. G. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf. Interface Anal.* **36,** 203–212 (2004).
31. Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **19,** 2756–2779 (2007).
32. Lustig, M., Donoho, D. & Pauly, J. M. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* **58,** 1182–1195 (2007).
33. Johnson, P. B. & Christy, R. W. Optical constants of the noble metals. *Phys. Rev. B* **6,** 4370–4379 (1972).
34. Johnson, W. L., Kim, S. A., Utegulov, Z. N., Shaw, J. M. & Draine, B. T. Optimization of arrays of gold nanodisks for plasmon-mediated Brillouin light scattering. *J. Phys. Chem. C* **113,** 14651–14657 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the Nano Research Facility (NRF), School of Engineering and Applied Science, Washington University, St Louis, USA for the nanoparticle synthesis. The NRF is a member of the US National Nanotechnology Infrastructure Network, supported by the US National Science Foundation under NSF award no. ECS-0335765. F.d.I.P. and C.D. acknowledge funding from the ERC under grant no. 259619 PHOTO EM. C.D. acknowledges the Royal Society for funding. P.A.M. and O.N. acknowledge financial support from the European Union's Seventh Framework Programme (FP/2007-2013) under the European Research Council ERC Grant Agreement 291522-3DIMAGE and under Grant Agreement 312483-ESTEEM2 (Integrated Infrastructure Initiative – I3). D.J.H. acknowledges Microsoft Research Connections and the EPSRC (grants nos EP/K008218/1 and EP/K039318/1) for financial support. We thank A. Howie, M. Kociak, E. Ringe and S. M. Collins for discussions and FEI (in particular E. Yucelen and S. Lazar) for access to an FEI Titan Microscope at the FEI Nanoport in Eindhoven.

Author Contributions P.A.M. and O.N. designed the experiment. O.N. performed the electron microscopy. F.d.I.P. performed simulations and data analysis. R.K.L. and D.J.H. performed compressed sensing electron tomography. All authors interpreted and discussed the experimental results and wrote and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.A.M. (pam33@cam.ac.uk).

METHODS

Spectral processing. The five spectrum images corresponding to data sets acquired at each tilt angle were first combined into a single array, to enable simultaneous processing. The large area of the spectrometer charge-coupled device camera and the pipelined acquisition cause a significant number of X-ray spikes (high-intensity pixels), which must be removed before performing the NMF. This was undertaken via a routine that identifies the spikes and performs an interpolation in the area where they are located. Spectra were then aligned to subpixel accuracy, via a correlation-based routine using the ZLP. The spectrum-image array was then cropped in the energy dimension to the 1–4 eV region of interest, and unfolded in its spatial dimension, to obtain a two-dimensional (2D) data object of size $300 \times 108{,}680$. The data were subsequently scaled to normalize the Poissonian noise[30], and factorized using a projected gradient method NMF algorithm[31]. The NMF was repeated for different numbers of components, ranging from four to twelve. The residuals of the factorization were used to judge the optimal number of components, which was determined to be eight. As an alternative method to estimate the optimal number of components, we used a principal-component analysis scree plot, obtaining the same result. The data from the -15° tilt spectrum image (the last recorded in the half-set) was not used for the factorization because it was judged to be modified by the effects of contamination. Contamination produces a redshift in the LSPR and would require a factorization with an increased number of components, unnecessarily increasing the complexity of the model. Nevertheless, EELS maps corresponding to the -15° tilt data were obtained by fitting the eight NMF spectral components to the -15° tilt spectrum image. Following these steps, the scaling was reversed and all the spatial components folded back to the original dimensions.

To a good approximation, the recorded EELS signal (which follows a Poissonian distribution) corresponds to the single scattering distribution, in which the intensity is proportional to the product of the probability of LSPR scattering and the area of the ZLP. Hence, we obtained the probability of LSPR scattering by dividing the EELS maps by the ZLP signal, pixel by pixel. This was an essential part of the data processing because without it a valid tomographic reconstruction would not be possible. This methodology is limited by multiple scattering, and is therefore applicable only to samples whose projected thickness throughout the tilt series remains sufficiently small relative to the mean free path.

Blind source separation methods in general, and NMF in particular, can estimate the original sources, except for an arbitrary scaling factor for each mode. This means that the intensity of the EELS map and the spectrum of each component are not calibrated, and that the relative probability of each LSPR excitation therefore cannot be directly obtained from the NMF EELS maps. In our case, we chose each scale factor by multiplying the EELS maps of each mode by the integral of the spectral component. In this way, the constant of proportionality, which links the intensity of the EELS maps with the probability of scattering, was made the same for all the modes, for ease of comparison. Unlike the division by the ZLP area, this normalization was not essential for the tomographic reconstruction process.

3D reconstruction. In recent work[9] involving high-angle annular dark-field STEM electron tomography, we showed that a compressed sensing approach to reconstruction can yield high-fidelity 3D reconstructions from far fewer measurements than are traditionally thought necessary. To make this paper self-contained, we describe the salient aspects of the compressed sensing electron tomography reconstruction methodology, as well as specific adaptations made to the algorithm for application to 3D reconstruction of the LSPRs of the silver nanocube.

Compressed sensing[7,8] can enable robust (and possibly even exact) recovery of highly undersampled signals by using the principle that many signals are sparse or compressible. Here 'compressible' means that the important details of the signal can be recovered from far less information than would be required to represent the entire signal. These ideas are also exploited in image compression algorithms, such as those used in the well-known JPEG and JPEG 2000 standards[35]. In the case of image compression, the image is transformed to a domain in which the important information can be represented by a relatively small number of large coefficients. This means that the minor coefficients can be discarded, and the amount of data needed to store the image is reduced. The original image can be reconstructed from the compressed representation, with minimal loss of information, by applying the inverse transform. Compressed sensing harnesses these principles during the initial acquisition of a signal. In compressed sensing, a small number of 'incoherent' initial samples are acquired. These samples capture just enough information to describe the signal accurately. The desired signal can be recovered via a nonlinear optimization process that incorporates the prior knowledge that the signal can be represented, or well approximated, sparsely in some transform domain.

As a preceding step to the tomographic reconstruction, we performed a normalization of the EELS maps, to compensate for the variation in the probability of excitation arising from a small amount of contamination in the irradiated area increasing through the tilt series. This was achieved by bringing the respective integrated intensities of the maps for each mode to the same level as the map with the highest integrated intensity among the maps for that mode.

To carry out the compressed sensing electron tomography reconstruction of the LSPR, we first applied a 2D Fourier transform to the EELS maps at each tilt angle, to obtain radially sampled data in the Fourier domain. Here we exploited the C_{4v} ($4mm$) symmetry of the cube–substrate system and the low contamination levels of the first half of the tilt series of EELS spectrum images, to attain a high-quality sampling of Fourier space over the full $\pm 60^\circ$ tilt range. This was achieved by using the first half-set of tilt-series EELS maps at their respective angles, as well as the same maps transformed by mirror symmetry in the plane of the tilt axis for the positive angles.

The incoherence condition for compressed sensing reconstruction is most readily interpreted as saying that artefacts resulting from the limited number of samples should appear 'noise-like' and be distributed throughout the reconstruction. Although the theoretical proofs of compressed sensing have focused on random data acquisition, radial sampling of Fourier space has been shown to be sufficiently incoherent to allow the application of compressed sensing[8,32]. To provide an initial reconstruction from the radial Fourier data, we applied a 3D non-uniform Fourier transform, adapted from the 2D approach of ref. 36. We then used the conjugate gradient descent algorithm of ref. 32 to solve the unconstrained optimization problem defined by

$$\min_x \left(\|Fx - y\|_2^2 + \lambda \|\Psi x\|_1 \right) \tag{3}$$

where F is the undersampled Fourier operator, x is the vector describing the 3D reconstruction, y is the vector describing the Fourier transform of the EELS maps and Ψ is the transform that maps the reconstruction to a domain in which it can be represented sparsely. In equation (3), the l_2-norm is defined as $\|x\|_2 = (|\sum_i |x_i|^2)^{1/2}$ and the term $\|\Psi x\|_1$ is the l_1-norm of the coefficients in the transform domain, defined as

$$\|x\|_1 = \sum_i |x_i|$$

The l_1-norm acts as a promoter of sparsity[7]. In equation (3), λ is a Lagrange multiplier that determines the level of importance of sparsity in the reconstruction. The optimum λ value for this work was determined empirically. In words, the above optimization seeks the sparsest solution for the 3D reconstruction in the transform domain, subject to consistency with the original data, in this case the EELS maps. Because the LSPRs can be considered localized, gradually varying functions, they can be well approximated in the wavelet domain. Here we used a separable Coiflet transform with eight vanishing moments as the sparsifying transform, implemented via the WAVELAB software package[37] for MATLAB.

The symmetry-based enhancement of the reconstruction was effected by combining, at each iteration in the reconstruction processes, the current reconstructed 3D volume with the same volume rotated by 90° in the plane of the substrate (akin to a dual-axis reconstruction approach[38,39] used in conventional electron tomography), followed by enforcement of symmetry of the wavelet-transformed reconstruction in the {110} planes parallel to the beam direction. This approach was found to yield high-quality reconstructions in good agreement with the EELS maps, as confirmed by re-projecting the reconstructed volumes encompassing the nanocube and LSPRs at the same angles as the experimental data acquisition (Extended Data Fig. 1).

DDA EELS simulations. EELS surface plasmon spectra in the 1.1–4.0-eV spectral range were simulated using the DDA code DDEELS v1.07alpha[29] with a 0.0145-eV energy step. The rounded silver cube was parameterized using a supersphere given by

$$\frac{x^{2/p} + y^{2/p} + z^{2/p}}{a} \leq 1$$

where a determines the size and p the roundness. To obtain a 100 nm cube with rounded corners and edges (radius, ~19 nm), we set $a = 100$ nm and $p = 0.4$. The high roundness (almost four times that of the sample under investigation) enabled us to use a coarser discretization, the only drawback being a blueshift of the spectral features that make quantitative comparisons of the energies difficult between simulation and experiment. The function was evaluated using a 5.5-nm-step 3D grid to obtain the position of 4,701 dipoles. The tabulated silver dielectric function was taken from ref. 33. The substrate was parameterized using a square prism of dimensions $150\ \text{nm} \times 150\ \text{nm} \times 30\ \text{nm}$ and was discretized using the same 5.5-nm-step 3D grid to obtain the position of 3,645 dipoles. The tabulated silicon nitride dielectric function was taken from ref. 34.

3D visualization. The voxel projection visualizations of the reconstructed 3D volumes shown in Fig. 3 were generated using the volume-rendering module in AVIZO FIRE (Visualization Sciences Group). This form of volume rendering enables objective visualization of the 3D information by 2D projection of the

colour-coded semi-transparent reconstructed volume. To exclude false regions of localized intensity that arose primarily at the periphery of some of the reconstruction volumes (owing to the imperfections involved in seeking a tomographic reconstruction from so few tilt-series images), the voxel projection views have been limited to the volume immediately surrounding the nanocube and LSPRs. Similarly, to enable clear visualization of the regions of maximal LSPR intensity, each voxel projection shows about the top 40–70% most intense voxels (the particular lower bound being chosen according to the prominence of background intensities, or those due to artefacts, in each reconstruction), with linearly increasing opacity from fully transparent at the minimum to fully opaque at the maximum. Some regions of weaker localized intensity potentially corresponding to weaker LSPR are present in the 3D reconstructions of some components, but because their intensities are appreciably lower than the maximal LSPR intensity, they are not displayed in the visualizations of Fig. 3. The outline of the cube superimposed on the voxel projections is based on the high-angle annular dark-field images.

Approximations for a valid tomographic reconstruction and interpretation. (1) Thin-particle approximation. In its most general form, the loss probability due to plasmon resonances of a fast electron travelling with constant velocity v, at time t, along a straight line trajectory $r=(R_0,z)$ can be written as[20]

$$\Gamma_{\mathrm{EELS}}(R_0,\omega)=\frac{e}{\pi\hbar\omega|v|}\int_{-\infty}^{\infty}\mathrm{d}z\ \mathrm{Re}\left[\mathrm{e}^{-i\omega z/|v|}v\bullet E_{\mathrm{ind}}(r,\omega)\right] \tag{4}$$

where $E_{\mathrm{ind}}(r,\omega)$ is the induced electric field acting back on the electron.

However, it is useful to write the loss probability in the quasistatic limit in terms of the screened interaction, $W(R_0,z,R_0,z',\omega)$ (ref. 20), as

$$\Gamma_{\mathrm{EELS}}(R_0,\omega)=\frac{e^2}{\pi\hbar|v|^2}\int_{-\infty}^{\infty}\mathrm{d}z\mathrm{d}z'\cos\left(\frac{\omega(z-z')}{|v|}\right)\mathrm{Im}[-W(R_0,z,R_0,z',\omega)]$$

where the induced potential ϕ can be expressed in terms of W:

$$\phi(R_0,z,\omega)=-\frac{e}{|v|}\int_{-\infty}^{\infty}\mathrm{d}z'\ W(R_0,z,R_0,z',\omega)\mathrm{e}^{iz'\omega/|v|}$$

We note that if $z-z'\ll|v|/\omega$, that is, in the so-called 'thin-particle approximation'[27], the loss probability is a projection of the imaginary part of the potential, ϕ, over the electron trajectory:

$$\lim_{z-z'\ll|v|/\omega}\Gamma_{\mathrm{EELS}}(R_0,\omega)=\frac{e}{\pi\hbar|v|}\int_{-\infty}^{\infty}\mathrm{d}z\ \mathrm{Im}[\phi(R_0,z,\omega)]$$

The difference $z-z'$ is usually taken as the size of the nanoparticle[26], but this may be an overestimate in some cases. We note that $\lim_{z-z'\to\infty}W(R_0,z,R_0,z',\omega)\to 0$, and it is therefore sufficient that $z-z'$ is large enough that $\mathrm{Im}[-W(R_0,z,R_0,z',\omega)]$ becomes negligible for a given frequency. As an example to illustrate this, for a cube on a substrate a trajectory parallel and close to a vertical edge may excite LSPRs at the bottom corner, top corner and edge of the cube. However, these LSPRs are excited at different frequencies, and W is therefore non-negligible only in the region over which the electron can efficiently excite each of the modes, a fraction of the edge length. Given the size of the cube, the very low energy losses and the energy and trajectory of the electrons, our experiment can be considered to be in the domain of this approximation. More generally, we note that when the frequency tends to zero, a spatial Fourier transform is simply a projection.

(2) Modal decomposition. From the analysis above, it does not follow directly that the EELS-LSPR signal can be described as a linear combination of spectral components—which we assume is possible for our NMF analysis. However, recently it has been shown[27] that, in the quasistatic limit, the LSPR loss probability, $\Gamma^{\mathrm{SP}}(R_0,\omega)$, can always be written as a linear combination of eigenspectra, each corresponding to a mode i and composed of a spectral weighting factor, $\mathrm{Im}(-g_i(\omega))$, and spatially varying basis functions, $|\phi_i(R_0,q)|^2$, such that

$$\Gamma^{\mathrm{SP}}(R_0,\omega)=C\sum_i\mathrm{Im}(-g_i(\omega))|\phi_i(R_0,q)|^2 \tag{5}$$

(equivalent to equation (2)), where $\phi_i(R_0,q)$ is the Fourier transform of $\phi_i(R_0,z)$, the induced potential associated with mode i:

$$\phi_i(R_0,q)=\int_{-\infty}^{\infty}\mathrm{d}z\ \phi_i(R_0,z)\mathrm{e}^{-iz\omega/|v|} \tag{6}$$

Furthermore, C is a constant and $q=\omega/|v|$.

With NMF, we undertake an analogous decomposition of the EELS spectra to derive 'spectral components' and 'spatial distribution' maps. The similarity between our NMF decomposition and that of equation (5) (equivalent to equation (2)) is striking. However, the question of whether the positivity constraint is sufficient to determine fully the decomposition deserves further study and is, indeed, an active topic of research[40]. It is, however, possible to make an ad hoc judgement of the quality of the decomposition in view of the NMF result for a given data set. Even if the NMF decomposition were ideal, the EELS-LSPR maps would be unlikely, in general, to correspond one to one with the basis functions in the modal decomposition, primarily because, owing to experimental limitations, we are unable to discern individual eigenmodes. Each experimental spectral component may be analogous to a single eigenspectrum, given by the spectral weighting factor $\mathrm{Im}(-g_i(\omega))$, but is more likely, at best, to be an incoherent sum of near-degenerate eigenspectra. In fact, in our experimental data sets from the cube–substrate system, we can separate and identify only five surface plasmon components, and it is only through the tomographic reconstruction that we can discern that some of them are composed of linear combinations of eigenmodes that are near degenerate in energy.

If we assume that the quantity $z\omega/|v|$ is small, equivalent to the 'thin-particle approximation'[27] mentioned above, then the Fourier transform of equation (6) approximates to a simple projection. In the modal decomposition, the basis functions for each mode are then equal approximately to the square modulus of a potential, associated with that mode, projected parallel to the beam trajectory. Within this approximation, the analogous term in our NMF decomposition, the spatial distribution, or EELS-LSPR, map for each component then resembles a quantity related to the basis functions in the modal decomposition.

In principle, there is no reason why, in general, a modal decomposition could not be possible also in the retarded domain and, importantly, a modal decomposition does indeed exist in the retarded domain for spheres[20,28]. It is also worth noting that, in general, retardation affects the energy and amplitude of the mode but not its projected symmetry[11].

(3) Vectorial nature of the excitation. In the domain of the thin-particle approximation we can now relate the EELS-LSPR maps to an induced potential, a scalar quantity, but one that, in general, depends on the trajectory of the electron because of the anisotropy of the crystalline (atomic) structure and/or the nanoparticle shape. Indeed, in general, the dielectric function of an anisotropic material, $\varepsilon(r)$, is a tensor and so will lead to variation in the induced potential with orientation. However, for the cubic crystals of interest here (nanoparticles of silver) the dielectric function is scalar and will thus be invariant with respect to crystal orientation. The geometry of the particle, which determines the geometry of the plasmonic excitation[27], imposes symmetry constraints on the modes for given trajectories and, in general, mode excitations will therefore show a variation with the angle of incidence similar to that of a dipole or higher multipoles. In principle, this would seem immediately to preclude any valid scalar tomographic reconstruction of modes (or components), but we show in an example described below that in some cases a scalar tomographic reconstruction is valid and interpretable.

We undertook a series of tomographic reconstructions using EELS spectra simulated using DDA software including retardation effects[29] (Extended Data Fig. 3). The software allowed for non-penetrating trajectories only, and energy loss spectra were therefore simulated as a function of position across the top face of a silver nanocube *in vacuo*. The cube was then tilted incrementally about an axis perpendicular to the substrate with another set of spectra simulated at each tilt increment. Five peaks, C_1, C_2, C_3, E and F, can be identified in the spectra. To obtain the intensity of the peaks we fitted Lorentzian functions to all except C_1, whose intensity was obtained by integrating the signal at low energy.

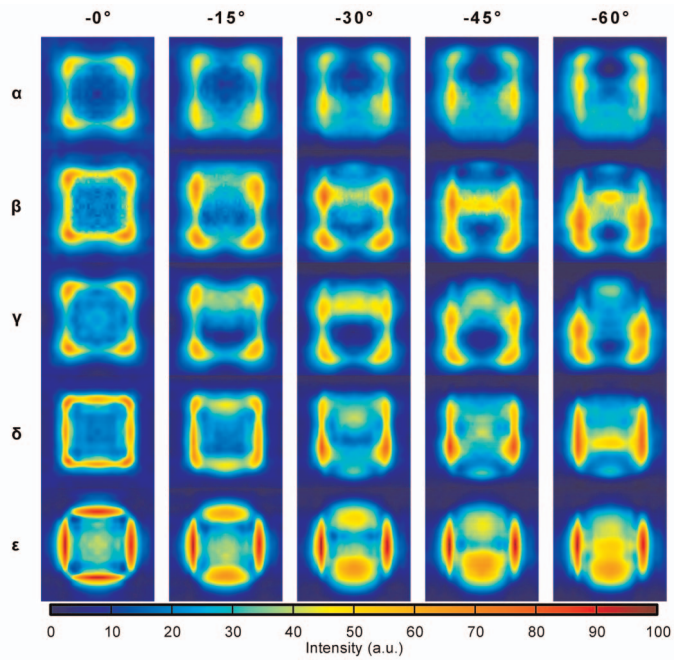
Tomographic reconstructions were made from this simulated tilt series using a conventional SIRT (simultaneous iterative reconstruction technique) algorithm[41]. Successful reconstructions were achieved for four of the five components; their spatial distributions showed significant excitation at the corners, but significantly delocalized over the whole cube face (C_1); at the corners (C_2); at the edges (E); and at the face centres (F). The only component that reconstructed poorly was C_3. This component corresponds to a single octupole mode[23,24] that has 2D quadrupolar symmetry on the top face of the cube.

We can now understand, using the symmetry of the excited multipolar fields, why the four major components reconstructed well but C_3 did not. The modes strongly excited in the simulations, which correspond to the four major components seen in the simulations, have dipole or quadrupole symmetry, and each is therefore triply degenerate[23,24]. For each dipole and quadrupole, there are two degenerate modes with dipolar symmetry at the top face in perpendicular directions. Consider a cube, with principal axes parallel to the x, y and z directions, tilted about the y axis, perpendicular to the beam. As the cube is tilted, the reduction in the contribution to the energy loss from a dipolar field parallel to the x axis (a reduction of the $v\bullet E$ term in equation (4) and an equivalent reduction in the

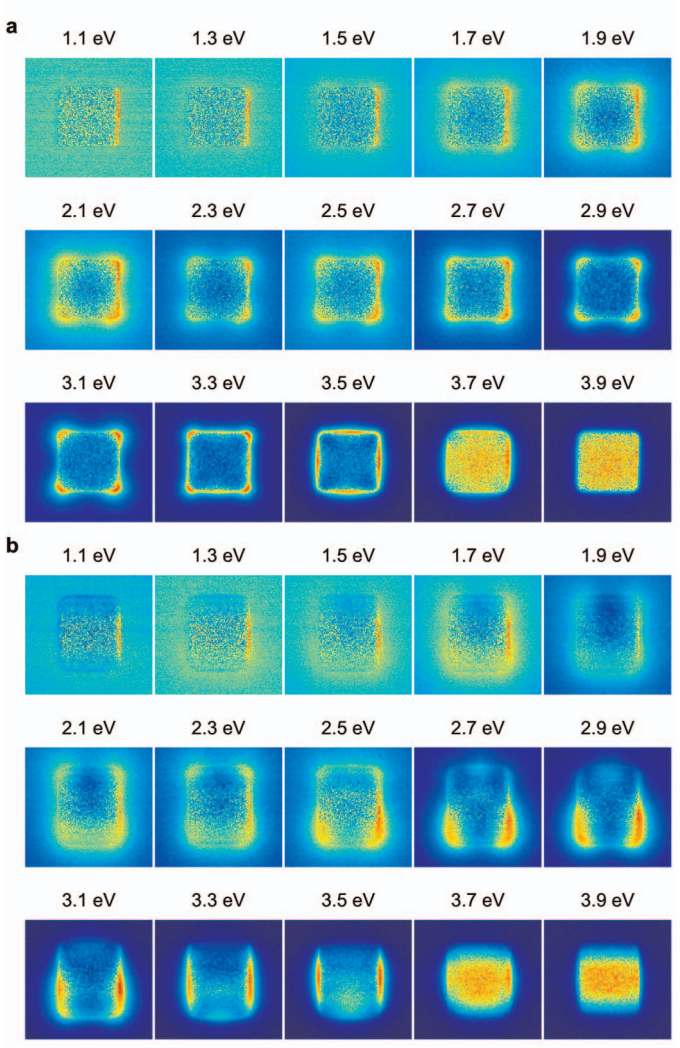
excitation) is compensated by the increase in the contribution of a (perpendicular) dipolar field parallel to the z direction. The net contribution from these two degenerate (or near-degenerate) dipolar fields, which, experimentally, would be reconstructed as a single component, gives an approximately uniform contribution to the energy loss across all tilt angles, and the component can therefore be reconstructed approximately using conventional scalar tomographic methods.

At the top face, a singly degenerate octupole mode[23,24] gives rise to a field with quadrupolar symmetry. As the cube is tilted, the quadrupolar field contribution varies sinusoidally but, unlike the dipolar fields described above, there is no other degenerate (or near-degenerate) mode with the same symmetry available to compensate, leading to a tomographic reconstruction of this mode with some significant negative values (Extended Data Fig. 3c). However, in our case the reconstruction intensities should be positive and the negative values seen for component C_3 indicate a breakdown in the simple scalar approximation. Although with small modifications the present approach can be adapted to other geometries, for objects of much lower symmetry a vector tomography method may be needed[25]. Interestingly, in our experimental geometry the substrate breaks the top–bottom (z) symmetry of the cube, and dipole and quadrupole modes will thus not be triply degenerate but will be split into two degenerate (x and y) modes and a singly degenerate (z) mode. Furthermore, substrate-mediated hybridization of the near-degenerate dipole, quadrupole and octupole modes leads to a more complex situation than in the simulation described above, but the 'compensation effect' of degenerate and near-degenerate modes is still applicable, allowing for example what seems to be a valid reconstruction of C_3 from the experimental data. However, the mismatches seen between the re-projections of Extended Data Fig. 1 and those of the original images in Fig. 2, which seem to be larger as trajectories become more parallel to the substrate, may, in part, be indicative of a breakdown of this compensation.

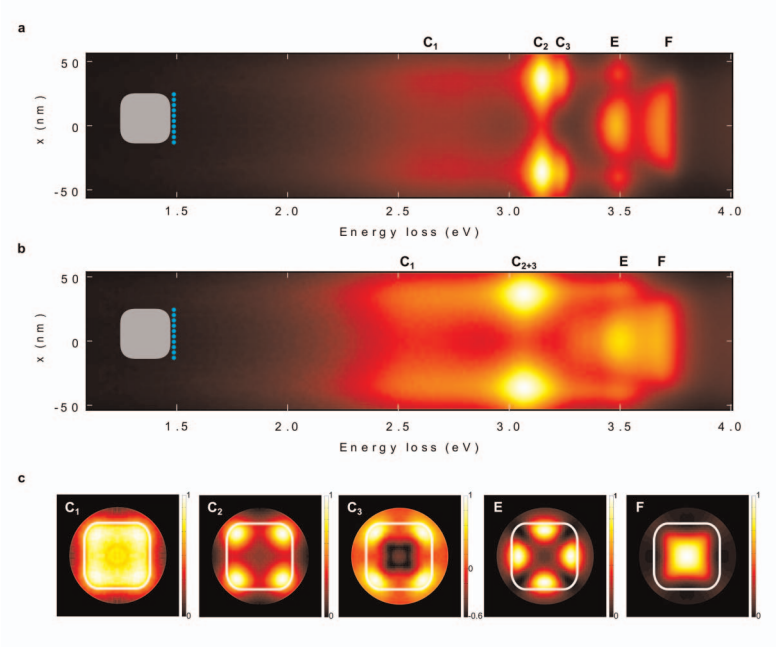
35. Taubman, D. S. & Marcellin, M. W. JPEG2000: standard for interactive imaging. *Proc. IEEE* **90,** 1336–1357 (2002).
36. Fessler, J. A. & Sutton, B. P. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Trans. Signal Process.* **51,** 560–574 (2003).
37. Buckheit, J. B. & Donoho, D. L. Wavelab and reproducible research. *Lect. Notes Stat.* **103,** 55–81 (1995).
38. Arslan, I., Tong, J. R. & Midgley, P. A. Reducing the missing wedge: high-resolution dual axis tomography of inorganic materials. *Ultramicroscopy* **106,** 994–1000 (2006).
39. Tong, J., Arslan, I. & Midgley, P. A novel dual-axis iterative algorithm for electron tomography. *J. Struct. Biol.* **153,** 55–63 (2006).
40. Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K. & Jensen, S. H. Theorems on positive data: on the uniqueness of NMF. *Comput. Intell. Neurosci.* **2008,** 764206 (2008).
41. Gilbert, P. Iterative methods for the three-dimensional reconstruction of an object from projections. *J. Theor. Biol.* **36,** 105–117 (1972).
42. Palik, E. D. *Handbook of Optical Constants of Solids* Vol. 1, 350–357 (Academic, 1998).
43. Agulleiro, J. I. & Fernandez, J. J. Fast tomographic reconstruction on multicore computers. *Bioinformatics* **27,** 582–583 (2011).



Extended Data Figure 1 | Re-projections of the LSPR components. Re-projections of the reconstructed 3D volumes of the five LSPR components, at the same tilt angles as the experimental data acquisition. Letters α–ε correspond to those of Figs 1–3. The orientation is as in Fig. 1, with the substrate–cube interface towards the top of the image at negative tilts. To exclude false regions of localized intensity that arose at the periphery of some of the reconstruction volumes (owing to the imperfections involved in seeking a tomographic reconstruction from so few tilt-series images), the re-projected volume was restricted to that immediately surrounding the nanocube and LSPRs.



Extended Data Figure 2 | Energy-filtered series of the STEM EELS spectrum images. Energy filtered series for the 0° tilt (**a**) and the −60° (**b**) STEM EELS spectrum images. The images are limited to the range of interest between 1 and 4 eV for LSPRs of the silver nanocube. The EELS signal in the images is integrated over an energy window 0.2 eV wide.



Extended Data Figure 3 | DDA EELS simulations of a silver nanocube. **a**, **b**, DDA EELS simulated as functions of position (as shown by the blue dotted line) for a 100-nm silver cube with rounded corners in vacuum with dielectric function tabulated in ref. 33 (**a**), and for a 100-nm silver cube with rounded corners in vacuum with dielectric function tabulated in ref. 42 (**b**). **c**, 2D tomographic reconstructions of the five spectral features identified in **a** and **b** (C_1, C_2, C_3, E and F), made from a tilt series of simulated spectra in which the cube is rotated about a cube axis perpendicular to the beam. The five spectral features correspond to lowest-energy (dipole) corner (C_1), higher-energy corner (C_2 and C_3), edge (E) and face (F) components. We note that the individual components C_2 and C_3 are not separated if the dielectric tabulation in ref. 42 is used, as in **b**. Each reconstruction is scaled independently. There are significant negative values only in the reconstruction of the C_3 component, which corresponds to a singly degenerate octupole mode. Tomographic reconstructions made from this tilt series were obtained using SIRT[41], implemented using the TOMO3D software package[43].

Extended Data Table 1 | Peak energies of LSPRs

Spectral feature	in vacuo	on 30 nm thick silicon nitride	
		distal	***proximal***
C_1	2.7 (-)	2.4 (2.0*)	
C_2	3.1 (-)	3.1 (2.7)	2.6 (-)
C_3	3.2 (-)	3.2 (2.9)	2.9 (2.2)
E	3.5 (-)	3.5 (3.3)	3.3 (2.7)
F	3.7 (-)	3.7 (3.6)	- (2.9)

Summary of the LSPRs of a 100-nm silver nanocube both *in vacuo* and resting on a 30-nm-thick silicon nitride substrate, as calculated from DDEELS simulations and as measured by STEM EELS spectrum images (values in brackets). Shown is the peak energy of each spectral feature (component) calculated with DDEELS (Fig. 4a, b) and each of the experimental NMF spectral components (Fig. 2a).
* Energy value measured directly from the peak maximum in EELS spectra. It is presented to facilitate comparison between simulations and experimental data and with other related work.

doi:10.1038/nature12532

Enhanced reversibility and unusual microstructure of a phase-transforming material

Yintao Song[1]*, Xian Chen[1]*, Vivekanand Dabade[1], Thomas W. Shield[1] & Richard D. James[1]

Materials undergoing reversible solid-to-solid martensitic phase transformations are desirable for applications in medical sensors and actuators[1], eco-friendly refrigerators[2,3] and energy conversion devices[4]. The ability to pass back and forth through the phase transformation many times without degradation of properties (termed 'reversibility') is critical for these applications. Materials tuned to satisfy a certain geometric compatibility condition have been shown[2,5–14] to exhibit high reversibility, measured by low hysteresis and small migration of transformation temperature under cycling[6,9,12,15]. Recently, stronger compatibility conditions called the 'cofactor conditions'[5,15] have been proposed theoretically to achieve even better reversibility. Here we report the enhanced reversibility and unusual microstructure of the first martensitic material, $Zn_{45}Au_{30}Cu_{25}$, that closely satisfies the cofactor conditions. We observe four striking properties of this material. (1) Despite a transformation strain of 8%, the transformation temperature shifts less than 0.5 °C after more than 16,000 thermal cycles. For comparison, the transformation temperature of the ubiquitous NiTi alloy shifts up to 20 °C in the first 20 cycles[9,16]. (2) The hysteresis remains approximately 2 °C during this cycling. For comparison, the hysteresis of the NiTi alloy is up to 70 °C (refs 9, 12). (3) The alloy exhibits an unusual riverine microstructure of martensite not seen in other martensites. (4) Unlike that of typical polycrystal martensites, its microstructure changes drastically in consecutive transformation cycles, whereas macroscopic properties such as transformation temperature and latent heat are nearly reproducible. These results promise a concrete strategy for seeking ultra-reliable martensitic materials.

Martensitic transformations are diffusionless, solid-to-solid phase transformations characterized by a change of crystal structure[8,14]. Accompanying this structural change, the mechanical (such as shape memory[12]), electromagnetic (such as magneto- and electro-caloric[2,3,11]), and transport (such as conductivity[17]) properties of the material can also change abruptly, which is useful in practical applications. During cyclic phase transformation, the desired functionality of martensitic materials degrades[2,9,16,18]. It is generally believed that the degradation of properties originates from the stressed transition layer between the two phases[13,14,19]. The same transition layer gives rise to an energy barrier that causes thermal hysteresis[7]. During the martensitic phase transformation, the stress in the transition layer drives irreversible processes, such as the formation of dislocations and the nucleation of microcracks[19]. These irreversible processes in turn lead to functional degradation and failure. Hence, high functional stability (that is, reversibility) can be achieved by reducing or even eliminating the elastic transition layer, which leads to the study of the geometric compatibility of the two phases.

A successful strategy[5–7,9,10,13] for eliminating this transition layer has been found by using the crystallographic theory of martensite[13,14,20,21]. According to this theory, if certain mild conditions are satisfied, each pair of twinned variants (a 'twin system') can form a laminated microstructure that meets austenite at a low-elastic-energy transition layer. The theory generically has four solutions per twin system, yielding four austenite–martensite interfaces, but corresponding to only two twinning volume fractions, $f^\star$ and $1-f^\star$. Figure 1a shows a typical solution of the crystallographic theory. The special cases $f^\star = 0$ and $f^\star = 1$ can occur and correspond to transition-layer-free interfaces between austenite and single variant martensite (Fig. 1b). This degeneracy occurs if and only if the condition $\lambda_2 = 1$ is satisfied[13], where λ_2 is the middle eigenvalue of the 3×3 'transformation stretch matrix' U, which is obtained from X-ray measurements of lattice parameters and knowledge of the space groups of the two phases[8,14]. Thus, the strategy for elimination of the stressed transition layer is to make $\lambda_2 \rightarrow 1$ by systematically tuning the composition of alloys. This strategy has been successfully applied to shape memory alloys[6,7,12], magnetocaloric materials[2] and energy materials[22,23].

The cofactor conditions represent a further degeneracy of the crystallographic theory of martensite[5,15]. The cofactor conditions are necessary and sufficient conditions on the lattice parameters and the twin system such that the crystallographic theory has solutions for every volume fraction $0 \leq f \leq 1$. This is in contrast with the above cases where there are only two volume fractions per twin system. When the cofactor conditions are satisfied, one can continuously vary the volume fraction of the twin variants while retaining the low-elastic-energy interface with austenite. The cofactor conditions consist of three sub-conditions that restrict the distortion and twin system. The first is $\lambda_2 = 1$. The second is $|U^{-1}\hat{e}| = 1$ for type I twins and $|U\hat{e}| = 1$ for type II twins, where $\hat{e}$ is a unit vector aligned with the twofold axes associated to these twins. Here, types I and II refer to certain classic symmetry relations[8,14] that hold for the lattices on each side of a twin plane. Physically, these conditions imply the presence of certain unstretched directions for the distortion and for the inverse distortion. The third is a mild condition that is only relevant for compound twins. A detailed description of these concepts is given in the Supplementary Information.

A detailed theoretical analysis[15] of cofactor conditions, summarized also here in the Supplementary Information, yields further unexpected implications of cofactor conditions. First, if cofactor conditions are satisfied for a twin system, they are typically satisfied also by a large family of twin systems. Second, if the twin system is type I or type II[8], half of these solutions of the crystallographic theory require no transition layer at all. These yield zero-elastic-energy interfaces with austenite for every volume fraction f. Finally, if cofactor conditions are satisfied simultaneously by twin systems of both types I and II, numerous further zero-elastic-energy microstructures can be constructed from triple junctions formed by austenite and a pair of type I twinned martensite variants (A and B in Fig. 1c, d), and quad junctions formed by four pairwise twinned variants (Fig. 1e, f). These two simple junctions can be combined to form the 'riverine' zero-elastic-energy microstructure seen in Fig. 1g. Figure 1c, e and g are drawn accurately using the measured lattice parameters of $Zn_{45}Au_{30}Cu_{25}$, but perturbed very slightly to satisfy the cofactor conditions exactly.

This plethora of zero-elastic-energy deformations implies that the material has a great many ways of accommodating non-transforming

[1]Department of Aerospace Engineering and Mechanics, University of Minnesota, Minneapolis, Minnesota 55455, USA.
*These authors contributed equally to this work.

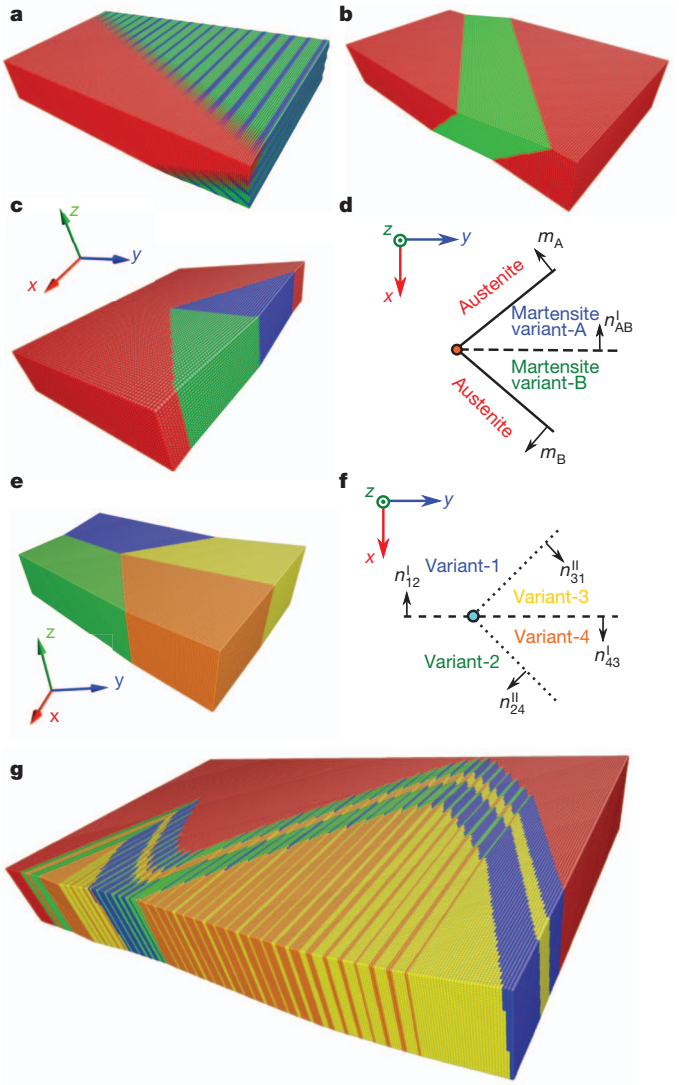


Figure 1 | Various austenite–martensite boundaries and special junctions. **a**, Planar phase boundary with transition layer. **b**, Planar phase boundary without transition layer. **c**, A triple junction formed by austenite and a type I twin pair, and its two-dimensional projection (**d**). **e**, A quad junction formed by four variants, and its two-dimensional projection (**f**). In **d** and **f**, solid lines are austenite–martensite interfaces with normals m_A and m_B, whereas dashed and dotted lines are type I and type II twin walls respectively, with normals given by n^I and n^{II}, with subscripts indicating the neighbouring variants. **g**, Curved phase boundary and riverine microstructure. In **a**–**c**, **e** and **g**, the red lattice represents austenite, and other colours are variants of martensite.

inclusions, defects and precipitates during transformation. The benefit of having such large classes of low-energy deformations is a recurring theme in the literature on phase transformations in polycrystals[12,24]. The cofactor conditions combine the advantages for hysteresis of having no transition layer with the existence of a great many low-energy deformations and the implications for reversibility.

Literature values of lattice parameters[25,26] suggested that the Heusler-type system Zn_2AuCu was a suitable candidate to tune so as to satisfy the cofactor conditions. We prepared a set of seven $Zn_{45}Au_xCu_{55-x}$ alloys in the composition range $20 \leq x \leq 30$ for preliminary study. After this preliminary screening, a set of three alloys, with $x = 25$ (Au25), $x = 27$ (Au27) and $x = 30$ (Au30) respectively, were prepared by arc-melting high-purity elements in vacuum. Their functional stability properties were characterized by X-ray diffraction and calorimetry (Supplementary Methods). For this alloy system, the austenite is face-centred cubic ($L2_1$ ordering)[25], whereas the martensite phase is M18R monoclinic[26]. Following the 'recipe' provided in the Supplementary Information, we use $X_I = |U^{-1}\hat{e}|$ and $X_{II} = |U\hat{e}|$ to quantify the cofactor conditions for twins of types I and II, respectively. The values $X_{I/II} = \lambda_2 = 1$ represents exact satisfaction of the cofactor conditions in each case. The values of λ_2, X_I and X_{II} (Table 1) show that (1) all three samples have λ_2 close to 1, and Au30 is the closest; (2) by changing the composition from Au25 to Au30, both X_I and X_{II} approach 1 simultaneously, and both are closest to 1 in Au30. In theory, X_I and X_{II} need not approach 1 simultaneously. Thus, the coincidental satisfaction of the cofactor conditions for twins of both type I and type II is apparently an accident, or else arises for reasons that are currently unknown.

Thermal cycling was done by the combination of differential scanning calorimetry (DSC) and a thermal cycling apparatus designed by us involving a thin-film heater competing against a liquid-nitrogen-cooled sample holder (Supplementary Fig. 3). For each specimen DSC measurements were made for the first 64 cycles. For each of the subsequent 2^n DSC cycles (where $n = 7, 8, ...$), the sample was removed from the cycling apparatus and a DSC measurement was made. The sample was then returned to the apparatus for further cycling. During cycling in the apparatus, the surface morphology of the specimen was observed *in situ* by optical microscopy.

The results of DSC measurements are shown in Fig. 2. Figure 2a–c shows the calorimetric curves of the DSC cycles. All three samples have sharp transformation peaks, from which the austenite start (A_s), austenite finish (A_f), martensite start (M_s) and martensite finish (M_f) temperatures are determined by the conventional onset-point construction. Hysteresis, calculated by $(A_s + A_f - M_s - M_f)/2$ for the first cycle, is given in Fig. 2a–c. The DSC curves clearly shift in Au25 and Au27, but no significant shift is observed in Au30. The data are summarized in Fig. 2d, which shows the shift of transformation temperatures versus the cycle number on a log scale. We see that the transformation temperatures migrate downwards significantly in Au25 and Au27, whereas in Au30, the transformation temperature oscillates slightly around the initial value. In Au25 the size of the hysteresis increases significantly with cycling, but the average transformation temperature migrates downward. These behaviours suggest that significant damage occurs in both phases due to the transformation process, but that the average free energy of martensite is more strongly increased. Also seen in Fig. 2a–c, and most clearly demonstrated by Au25, is that the area under the transformation peak, corresponding to the latent heat, shrinks during cycling. This is summarized in Fig. 2e. Again, as the composition is changed from Au25 to Au30, the shrinkage of latent heat decreases, and it almost disappears in Au30. We extended the cycling test on Au30 to $2^{14} = 16{,}384$ cycles. The shift of the transformation temperatures and the shrinkage of latent heat during this long test are plotted in Fig. 2f. We see only small changes of these values in Au30 even after such a large number of thermal cycles. This is remarkable given that Zn_2AuCu is a soft alloy with a relatively high homologous temperature (the transformation temperature to melting temperature ratio) of about 0.22. Taken together, these observations cast significant doubt on the standard explanations for hysteresis based on pinning of interfaces by defects or thermal activation.

Figure 3a,c and e shows the surface morphology of each specimen in the phase in which it was originally polished (austenite for Au30 and martensite for Au25 and Au27), after 64 cycles. Figure 3b, d and f shows the microstructure of the other phase in several consecutive

Table 1 | Geometric compatibility conditions in three specimens

	Au25	Au27	Au30
λ_2	1.0092 ± 0.0002	1.0086 ± 0.0001	1.0006 ± 0.0002
X_I	0.9899 ± 0.0034	1.0090 ± 0.0001	1.0081 ± 0.0008
	1.0179 ± 0.0028	1.0222 ± 0.0024	1.0339 ± 0.0005
X_{II}	1.0256 ± 0.0050	1.0056 ± 0.0007	0.9996 ± 0.0008
	0.9893 ± 0.0017	0.9884 ± 0.0005	0.9695 ± 0.0004

The reason that X_I and X_{II} have two values for each material is given in Supplementary Information III.

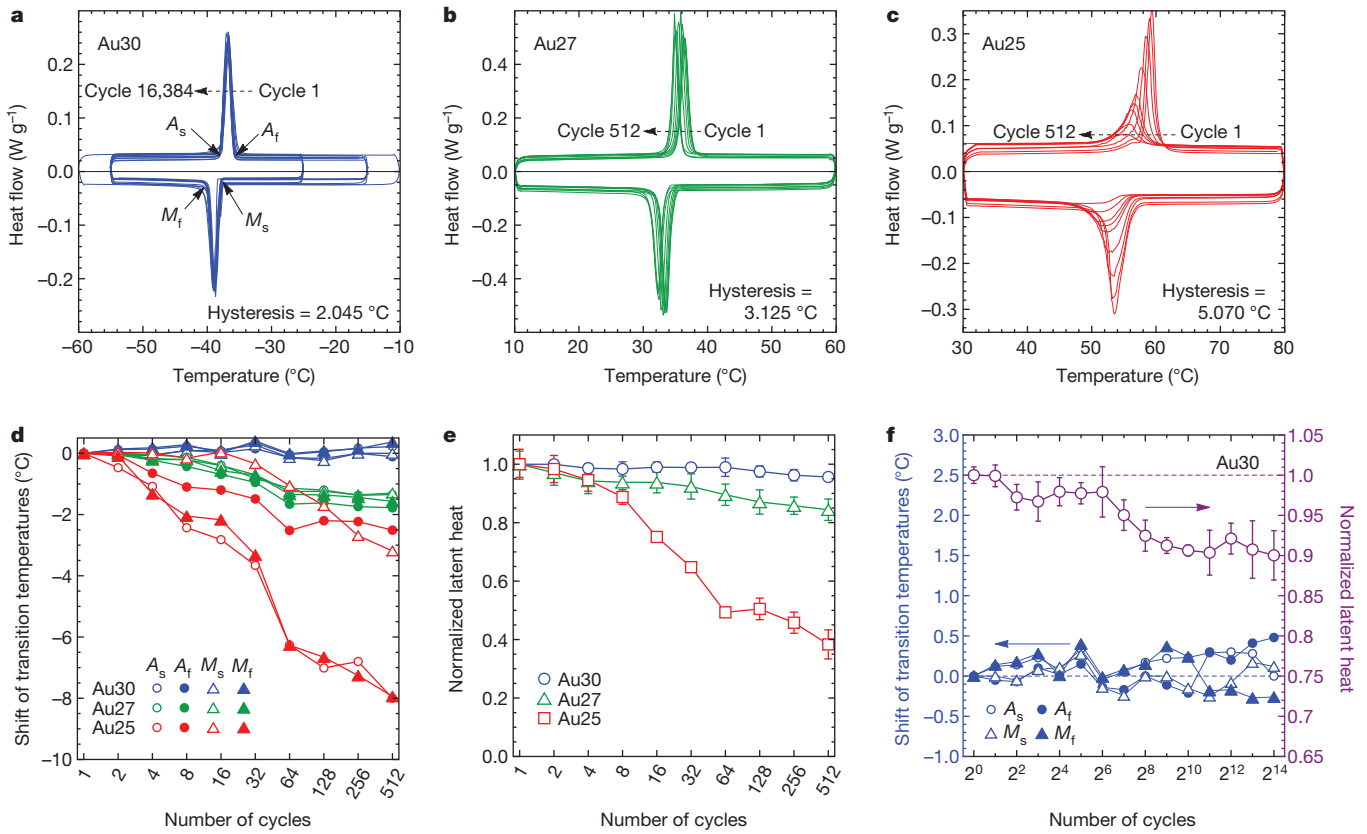


Figure 2 | Functional stability of $Au_xCu_{55-x}Zn_{45}$ alloys where $x = 25$ (Au25), $x = 27$ (Au27) and $x = 30$ (Au30) during thermal cycling. **a**–**c**, DSC data of three specimens. The values of hysteresis, $(A_s + A_f - M_s - M_f)/2$, are calculated for the virgin cycle. **d**, The shift of austenite start (A_s), finish (A_f) and martensite start (M_s), finish (M_f) temperatures. **e**, Latent heat measured in different cycles normalized by the value of the virgin cycle. Data points represent the average values of latent heat upon heating and cooling, and the error bars represent the differences between them. **f**, Functional stability of Au30 extended to $2^{14} = 16{,}384$ cycles.

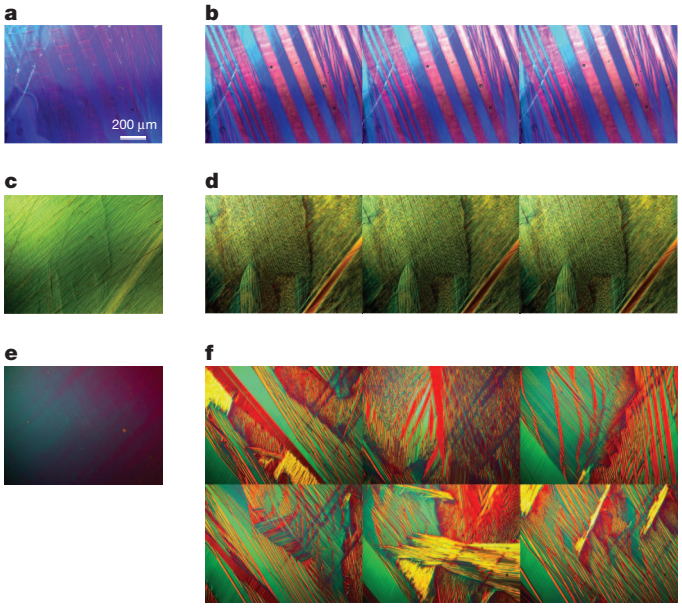


Figure 3 | Microstructures in consecutive cycles. **a**, Polished martensite surface of Au25 after 64 cycles. **b**, Austenite (inverse) microstructure of Au25 for three consecutive cycles immediately after taking the micrograph in **a**. **c**, Polished martensite surface of Au27 after 64 cycles. **d**, Austenite (inverse) microstructure of Au27 for three consecutive cycles immediately after taking the micrograph in **b**. **e**, Polished austenite surface of Au30 after 64 cycles. **f**, Martensite microstructure of Au30 for six consecutive cycles immediately after taking the micrograph in **e**.

cycles immediately following those of Fig. 3a, c and e. We see that Au25 and Au27 show the same microstructure in all three cycles. However, the microstructure in Au30 is completely different in each of the six cycles (see the Supplementary Video), which is repeated throughout the cycling process.

Figure 4 shows the morphology of Au30 in a single grain (about 1 mm), obtained by stitching a dozen micrographs together. We can see various hierarchical microstructures that resemble those predicted by the above theory. Particularly, the riverine microstructure shown on the left of Fig. 4, also seen frequently during cycling, has to our knowledge not been seen in any martensitic material. An enlarged view of the edge of this riverine microstructure is provided in Supplementary Fig. 4. In forthcoming work involving electron backscatter diffraction and transmission electron microscopy we will investigate the connection between Fig. 4 and theoretical predictions more precisely.

Of the three samples, Au30, which is the one that most closely satisfies cofactor conditions for both type I and type II twin systems, exhibits the lowest hysteresis and the highest functional stability. Also, its microstructure is completely unlike any other martensite we have seen. For example, the repeating microstructures upon phase transformation cycles in Au25 and Au27 are consistent with the common observation[27–30] that polycrystal martensitic materials exhibit detailed reproducibility of microstructure and acoustic emission traces, related to 'return-point memory'[28]. Au30 clearly loses this memory. We conjecture that this observation is linked to the large number of ways of constructing low- and zero-elastic-energy austenite–martensite interfaces in materials satisfying the cofactor conditions. The vast number of low-energy states possible in this material implies that minor variations of conditions from cycle to cycle lead to diverse microstructures.

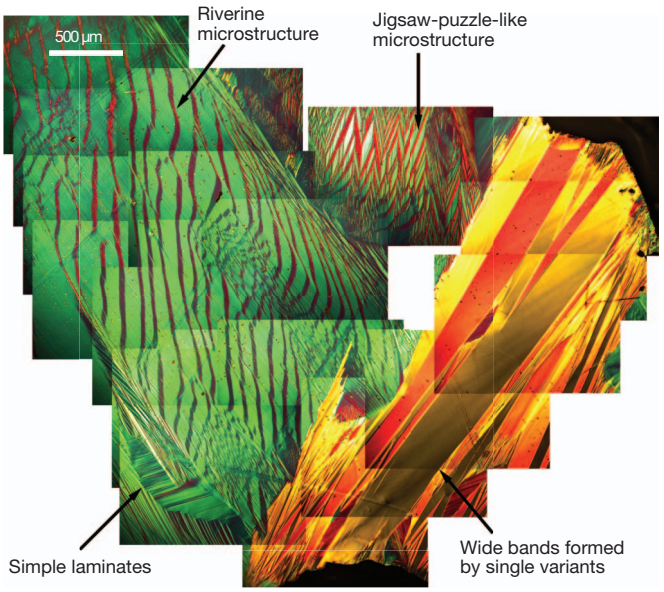


Figure 4 | Various hierarchical microstructures in Au30, the alloy most closely satisfying the cofactor conditions for both type I and type II twin systems.

We have thus found the first martensitic material that closely satisfies the cofactor conditions, and it happens to satisfy them for both type I and type II twin systems. This material exhibits ultrahigh reversibility and unusual microstructure. The theory for its fabrication can readily be adapted to other alloy systems, because it depends only on lattice parameters that can be finely tuned by changing the composition. Our result suggests a universal strategy for developing ultra-reliable martensitic materials particularly suited to medical, microelectronic and energy applications.

METHODS SUMMARY

Polycrystal ingots with nominal composition $Au_xCu_{55-x}Zn_{45}$ ($x = 25$, 27 and 30) were prepared by melting Cu (99.9999%), Au (99.999%) and Zn (99.9999%) pellets in an evacuated (10^{-5} mbar) and sealed silica capsule. The capsule was kept in the furnace, the temperature of which was varied as follows: 600 °C for 2 h, 800 °C for 8 h and finally 1,200 °C for 24 h. To promote homogeneity, the ingots were again reheated to 1,200 °C for 24 h while the silica capsules were rotated about the axis of a tube furnace at 30 r.p.m. The net weight losses were less than 0.01%. Finally, the ingots were annealed at 650 °C for 24 h and quenched in ice water.

A TA Q1000 machine calibrated by indium was used for DSC measurements at the rate of ±10 °C per minute. Specimens were finely polished on both sides at the beginning to ensure good thermal contact. For each sample, the first two cycles were scanned from −100 °C to 180 °C to identify the transformation temperatures. The following DSC cycles were then scanned over a temperature range of about 50 °C covering the identified transformation temperatures.

The cycling on the thermal stage that we designed was performed over a small temperature range determined by the stabilization of microstructure upon transformation, which was about 10 °C. The cycling frequency was about 0.1 Hz. Microstructure was observed by optical microscopy with differential interference contrast technology. The colour code was not calibrated.

X-ray diffraction was done using a Bruker AXS microdiffractometer (Cu Kα radiation source) with a temperature-controlled stage. Data was collected by general area detector diffraction system (GADDS). The sample surface was polished before being mounted to the stage at room temperature. The peak positions were refined using the JADE v7.0 software for the precise determination of lattice parameters.

Received 15 May; accepted 6 August 2013.

1. Walia, H., Brantley, W. A. & Gerstein, H. An initial investigation of the bending and torsional properties of Nitinol root canal files. *J. Endod.* **14,** 346–351 (1988).
2. Liu, J., Gottschall, T., Skokov, K. P., Moore, J. D. & Gutfleisch, M. O. Giant magnetocaloric effect driven by structural transitions. *Nature Mater.* **11,** 620–626 (2012).
3. Moya, X. *et al.* Giant electrocaloric strength in single-crystal $BaTiO_3$. *Adv. Mater.* **25,** 1360–1365 (2013).
4. Srivastava, V., Song, Y., Bhatti, K. & James, R. D. The direct conversion of heat to electricity using multiferroic alloys. *Adv. Energy Mater.* **1,** 97–104 (2011).
5. James, R. D. & Zhang, Z. In *Magnetism and Structure in Functional Materials* (eds Planes, A., Mañosa, L. & Saxena, A.) 159–175 (Springer, 2005).
6. Cui, J. *et al.* Combinatorial search of thermoelastic shape-memory alloys with extremely small hysteresis width. *Nature Mater.* **5,** 286–290 (2006).
7. Zhang, Z., James, R. D. & Müller, S. Energy barriers and hysteresis in martensitic phase transformations. *Acta Mater.* **57,** 4332–4352 (2009).
8. Pitteri, M. & Zanzotto, G. *Continuum Models for Phase Transitions and Twinning in Crystals* (Chapman and Hall/CRC, 2010).
9. Zarnetta, R. *et al.* Identification of quaternary shape memory alloys with near-zero thermal hysteresis and unprecedented functional stability. *Adv. Funct. Mater.* **20,** 1917–1923 (2010).
10. Delville, R. *et al.* Transmission electron microscopy study of phase compatibility in low hysteresis shape memory alloys. *Phil. Mag.* **90,** 177–195 (2010).
11. Srivastava, V., Chen, X. & James, R. D. Hysteresis and unusual magnetic properties in the singular heusler alloy $Ni_{45}Co_5Mn_{40}Sn_{10}$. *Appl. Phys. Lett.* **97,** 014101 (2010).
12. Bechtold, C., Chluba, C., de Miranda, R. L. & Quandt, E. High cyclic stability of the elastocaloric effect in sputtered TiNiCu shape memory films. *Appl. Phys. Lett.* **101,** 091903 (2012).
13. Ball, J. M. & James, R. D. Fine phase mixtures as minimizers of energy. *Arch. Ration. Mech. Anal.* **100,** 13–52 (1987).
14. Bhattacharya, K. *Microstructure of Martensite: Why It Forms and How It Gives Rise to the Shape-Memory Effect* (Oxford Univ. Press, 2003).
15. Chen, X., Srivastava, V., Dabade, V. & James, R. D. Study of the cofactor conditions: conditions of supercompatibility between phases. *J. Mech. Phys. Solids* http://dx.doi.org/10.1016/j.jmps.2013.08.004 (2013).
16. Tadaki, T., Nakata, Y. & Shimizu, K. Thermal cycling effects in an aged Ni-rich Ti-Ni shape memory alloy. *Trans. Jpn Inst. Metals* **28,** 883–890 (1987).
17. Mott, N. F. *Metal-Insulator Transitions* Ch. 5 (Taylor & Francis, 1990).
18. Eggeler, G., Hornbogen, E., Yawny, A., Heckmann, A. & Wagner, M. Structural and functional fatigue of NiTi shape memory alloys. *Mater. Sci. Eng. A* **378,** 24–33 (2004).
19. Norfleet, D. M. *et al.* Transformation-induced plasticity during pseudoelastic deformation in Ni-Ti microcrystals. *Acta Mater.* **57,** 3549–3561 (2009).
20. Wechsler, M. S., Lieberman, D. S. & Read, T. A. On the theory of the formation of martensite. *J. Metall./Trans. AIME* **197,** 1503–1515 (1953).
21. Bowles, J. S. & Mackenzie, J. K. The crystallography of martensite transformations I/II. *Acta Metall.* **2,** 129–137 (1954).
22. Meethong, N., Huang, H.-Y., Speakman, S., Carter, W. & Chiang, Y.-M. Strain accommodation during phase transformations in olivine-based cathodes as a materials selection criterion for high-power rechargeable batteries. *Adv. Funct. Mater.* **17,** 1115–1123 (2007).
23. Louie, M. W., Kislitsyn, M., Bhattacharya, K. & Haile, S. M. Phase transformation and hysteresis behavior in $Cs_{1-x}Rb_xH_2PO_4$. *Solid State Ion.* **181,** 173–179 (2010).
24. Bhattacharya, K. & Kohn, R. V. Symmetry, texture and the recoverable strain of shape-memory poly-crystals. *Acta Mater.* **44,** 529–542 (1996).
25. Tadaki, T., Okazaki, H., Yoshiyuki, N. & Shimizu, K. Atomic configuration determined by ALCHEMI and X-ray diffraction of the $L2_1$ type parent phase in a Cu-Au-Zn shape memory alloy. *Mater. Trans. JIM* **31,** 935–940 (1990).
26. Tadaki, T., Okazaki, H., Yoshiyuki, N. & Shimizu, K. Atomic configuration determined by ALCHEMI and X-ray diffraction of a stabilized M18R martensite in a β phase Cu-Au-Zn alloy. *Mater. Trans. JIM* **31,** 941–947 (1990).
27. Amengual, A. *et al.* Systematic study of the martensitic transformation in a Cu-Zn-Al alloy. Reversibility versus irreversibility via acoustic emission. *Thermochim. Acta* **116,** 195–208 (1987).
28. Sethna, J. P. *et al.* Hysteresis and hierarchies: dynamics of disorder-driven first-order phase transformations. *Phys. Rev. Lett.* **70,** 3347–3350 (1993).
29. Sethna, J. P., Dahmen, K. A. & Myers, C. R. Crackling noise. *Nature* **410,** 242–250 (2001).
30. Vives, E., Soto-Parra, D., Mañosa, L., Romero, R. & Planes, A. Imaging the dynamics of martensitic transitions using acoustic emission. *Phys. Rev. B* **84,** 060101 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the financial support of MURI projects FA9550-12-1-0458 (administered by AFOSR) and W911NF-07-1-0410 (administered by ARO). This research also benefited from the support of NSF-PIRE grant number OISE-0967140. Y.S. thanks the Graduate School of the University of Minnesota for support through a Doctoral Dissertation Fellowship.

Author Contributions R.D.J. is the Principal Investigator and initiated and supervised the work. Y.S. designed the thermal cycling apparatus and carried out optical and calorimetric experiments. X.C. performed X-ray diffraction measurements and theoretical calculations of microstructure. V.D. synthesized all the specimens used in the study. T.W.S. provided expertise in the experimental design and data acquisition. All authors discussed the results and approved the manuscript. Y.S., X.C. and R.D.J. interpreted the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.D.J. (james@umn.edu).

doi:10.1038/nature12567

Calving fluxes and basal melt rates of Antarctic ice shelves

M. A. Depoorter[1], J. L. Bamber[1], J. A. Griggs[1], J. T. M. Lenaerts[2], S. R. M. Ligtenberg[2], M. R. van den Broeke[2] & G. Moholdt[3]

Iceberg calving has been assumed to be the dominant cause of mass loss for the Antarctic ice sheet, with previous estimates of the calving flux exceeding 2,000 gigatonnes per year[1,2]. More recently, the importance of melting by the ocean has been demonstrated close to the grounding line and near the calving front[3–5]. So far, however, no study has reliably quantified the calving flux and the basal mass balance (the balance between accretion and ablation at the ice-shelf base) for the whole of Antarctica. The distribution of fresh water in the Southern Ocean and its partitioning between the liquid and solid phases is therefore poorly constrained. Here we estimate the mass balance components for all ice shelves in Antarctica, using satellite measurements of calving flux and grounding-line flux, modelled ice-shelf snow accumulation rates[6] and a regional scaling that accounts for unsurveyed areas. We obtain a total calving flux of 1,321 ± 144 gigatonnes per year and a total basal mass balance of −1,454 ± 174 gigatonnes per year. This means that about half of the ice-sheet surface mass gain is lost through oceanic erosion before reaching the ice front, and the calving flux is about 34 per cent less than previous estimates derived from iceberg tracking[1,2,7]. In addition, the fraction of mass loss due to basal processes varies from about 10 to 90 per cent between ice shelves. We find a significant positive correlation between basal mass loss and surface elevation change for ice shelves experiencing surface lowering[8] and enhanced discharge[9]. We suggest that basal mass loss is a valuable metric for predicting future ice-shelf vulnerability to oceanic forcing.

Antarctica gains mass from snow accumulation in its interior and loses mass through ice discharge across the grounding line and into the ocean, where ice shelves form. These floating shelves are crucial to the stability of the ice sheet because they buttress the grounded ice upstream[10]. Loss of buttressing from ice-shelf thinning or removal leads to enhanced discharge of inland ice[11] and may be triggered by oceanic[8] and atmospheric[12] warming.

Calving fluxes for the whole of Antarctica have been inferred from temporally and spatially limited ship-based campaigns and satellite tracking from the US National Ice Center[1,2,7]. These calculations relied on many assumptions about the volume, density and lifetime of icebergs[1]. In 1992, the total calving flux was calculated to be 2,016 ± 672 Gt yr^{-1} (ref. 1), in agreement with the mean of 12 previous estimates from the 1970s and 1980s. Combining estimates of snow accumulation at the surface of the ice sheet and subshelf melt rates, this led to the conclusion that Antarctica was losing more than 1.3 mm yr^{-1} in sea-level equivalent as a result of enhanced iceberg calving[1,2]. In the absence of better estimates, recent studies of the hydrographic effects[7] of, and iron fluxes[13] from, icebergs in the Southern Ocean have used these figures (Supplementary Discussion 1).

Melting of ice shelves in Antarctica is caused by three different modes of relatively warm-water circulation[1]. The first mode is related to sea-ice formation and production of high-salinity shelf water that reaches the grounding line and forms ice-shelf water (ISW), a mix of high-salinity shelf water and fresh water. The second mode is due to the incursion of circumpolar deep water into the ice-shelf cavity, and the third mode is due to tidal and wind-induced mixing near the ice-shelf edge. Between the grounding line and the ice-shelf edge, refreezing takes place as the rising plume of ISW becomes supercooled and precipitates frazil ice (mode 1). This results in melting under ice shelves being most prevalent at the grounding line and close to the calving front[3–5,14]. Melt rates under Antarctic ice shelves have been inferred from glaciological studies[3], water measurements underneath[15] and in front of ice shelves, modelling studies[16], and compilations of different approaches[1]. Whereas glaciological studies are limited to a few ice-shelf locations, oceanographic studies lack temporal and spatial resolution. Modelling studies have provided important knowledge on ice–ocean interactions[16] but still have considerable uncertainties owing to a paucity of oceanographic and sub-ice-shelf geometry data[1,17], poorly resolved basal accretion[18], and grid resolution limitations[18].

Detailed studies of ice-shelf mass fluxes have provided improved estimates of the mass balance for a few targeted ice shelves[5,14], but so far no study has undertaken this rigorously across the whole of Antarctica. Here we calculate calving fluxes from ice thickness, derived directly from either our analysis of satellite radar altimeter measurements of freeboard[19] (ice-shelf elevation above mean sea level) or from ice-penetrating radar (IPR) data (32% by the first technique and 68% by the second). The altimetry is combined with corrections for changes in elevation between 1995 and 2009, and for firn air content and compaction obtained from our regional climate model[6] (Supplementary Discussion 2). We do this for all ice shelves exceeding 100 km^2 in size and combine the derived thicknesses with ice surface velocity from synthetic-aperture radar interferometry[20] (InSAR). The grounding-line fluxes (GLFs) are then obtained in combination with a new grounding-line data set. Although we do not focus here on the mass balance of the grounded ice sheet, it is interesting to note that the difference between our GLF and grounded surface mass balance (SMB) is −66 Gt yr^{-1}. This is, unlike previous mass budget estimates, very close to a recent assessment, from the GRACE satellite mission, of −69 Gt yr^{-1} for the period 2002–2010[21].

The basal mass balance (BMB) is determined, assuming conservation of mass, from the difference between the GLF and the SMB, and the calving flux (CF). Ice-shelf thinning rates[8] are added to BMB to account for non-steady-state behaviour (Table 1). We account for unsurveyed shelves using a physically based regional upscaling of our results (Supplementary Discussion 3). We find that for Antarctica as a whole, mass loss is roughly equally split between basal mass loss (the sum of total melt and accretion) and calving. Locally, however, the melt ratio ($MR = |BMB|/[CF + |BMB|]$) varies considerably, from ~10% to ~90% (Fig. 1). For the fringing ice shelves of West Antarctica, it is 74% (Table 1). Thus, for the Bellingshausen Sea and Amundsen Sea sectors, about two-thirds of the mass loss is via BMB. In contrast, the average melt ratio for the rest of Antarctica is 40%, and for the two largest ice shelves, the Filchner-Ronne and the Ross, the ratio is just 17%. These two ice shelves are consequently responsible for one-third of the iceberg production in Antarctica.

[1]Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK. [2]Institute for Marine and Atmospheric Research Utrecht, Utrecht University, 3584 CC Utrecht, The Netherlands. [3]Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA.

Table 1 | Mass balance of Antarctic ice shelves by oceanic sector

Ocean sector	Ice shelves	GLF (Gt yr^{-1})	SMB (Gt yr^{-1})	CF (Gt yr^{-1})	dh/dt (Gt yr^{-1})	BMB (Gt yr^{-1})	Ice-shelf area (10^3 km^2)	SBMB (m yr^{-1})	MR (%)
West Indian Ocean	AR, NE, AIS, W*	235 ± 30	49 ± 8	155 ± 22	−11 ± 8	−140 ± 38	174	−0.80 ± 0.22	47
West Indian Ocean+		324 ± 31	—	204 ± 29	—	−179 ± 43	—	—	47
East Indian Ocean	SHA*, VAN, TOT*,	333 ± 16	48 ± 7	213 ± 44	−51 ± 20	−219 ± 48	65	−3.35 ± 0.73	51
East Indian Ocean+	MU, POR*, ADE*,MER, NIN, COO, REN*	508 ± 26	—	306 ± 75	—	−300 ± 80	—	—	50
Ross Sea	DRY, RIS, SUL,	149 ± 16	71 ± 17	153 ± 10	0 ± 0	−67 ± 26	492	−0.14 ± 0.05	30
Ross Sea+		175 ± 16	—	167 ± 15	—	−79 ± 28	—	—	32
Amundsen Sea	LAN*, GET*, CD*,	383 ± 19	55 ± 11	198 ± 43	−156 ± 13	−395 ± 48	56	−7.11 ± 0.87	67
Amundsen Sea+	THW*, PI*, COS	505 ± 24	—	232 ± 50	—	−484 ± 57	—	—	68
Bellingshausen Sea	ABB*, VEN*, GEO*,	139 ± 11	82 ± 16	31 ± 10	−65 ± 43	−255 ± 22	86	−2.98 ± 0.26	89
Bellingshausen Sea+	WOR	174 ± 12	—	41 ± 13	—	−281 ± 23	—	—	87
Weddell Sea	LBC, FRIS, BRL, JFL	334 ± 35	139 ± 23	355 ± 31	0 ± 0	−118 ± 52	608	−0.19 ± 0.09	25
Weddell Sea+		363 ± 35	—	371 ± 33	—	−131 ± 53	—	—	26
Fringing West Antarctica	SUL, LAN*, GET*,	542 ± 23	147 ± 19	232 ± 54	−221 ± 45	−678 ± 53	154	−4.40 ± 0.35	74
Fringing West Antarctica+	CD*, THW*, PI*, COS*, ABB*, VEN*, GEO*, WOR	700 ± 27	—	275 ± 63	—	−792 ± 62	—	—	74
Total surveyed	—	1,573 ± 56	444 ± 36	1,106 ± 141	−282 ± 50	−1,193 ± 163	1,481	−0.81 ± 0.11	52
Total upscaling	—	476 ± 67	—	216 ± 33	—	-261 ± 34	74	−3.53 ± 0.47	55
Total Antarctica	—	2,049 ± 87	—	1,321 ± 44	—	−1,454 ± 174	1,555	−0.94 ± 0.11	52

See Fig. 1 for ice-shelf names and Supplementary Table 1 for the data here as tabulated for individual ice shelves. A plus sign indicates that regional upscaling is included. dh/dt, non-steady-state mass change. Uncertainty estimates, 1 s.d.
* Corrected for imbalance using ICESat (NASA's Ice, Cloud, and land Elevation Satellite) elevation rates.

The greatest basal mass loss does not come from the largest ice shelves, but from medium- to small-sized ones such as George VI, Getz, Totten and Pine Island (Fig. 1). Representing 91% of the ice-shelf area, the ten largest ice shelves produce only ~50% of the basal mass loss for Antarctica. Studies focusing on a small number of large ice shelves (four to ten) are therefore not representative of the continent as a whole[1,17]. Our total BMB, of −1,454 ± 174 Gt yr^{-1}, is of the same order of magnitude as estimates from oceanographic measurements and modelling[1,17,18,22] (~500–1,600 Gt yr^{-1}), but we find large regional differences. For example, a finite-element mesh ice–ocean model yields larger numbers for most of the ice shelves considered, especially the largest ones[18] (Supplementary Table 2). This is despite not taking into account the tidal effect on melting at the calving front, which is an important factor in the overall ice-shelf mass balance[3,5]. This overestimation seems partly to stem from very low accretion rates[18] (an order of magnitude lower than estimates based on observations for most parts of the Filchner-Ronne Ice Shelf accretion zone[5]). Different atmospheric forcings explain most of the large discrepancies between different model simulations[17,18,22] (H. Hellmer, personal communication). However, our results for the BMB agree well with previous estimates using a similar methodology[14,23] (Supplementary Discussion 4).

The mean specific BMB (SBMB = BMB per unit area) for all ice shelves is −0.81 ± 0.11 m yr^{-1} (water equivalent), but it varies from −0.07 to −15.96 m yr^{-1} between ice shelves (Fig. 2 and Supplementary Table 1). The SBMB is strongly negative (more than 2.00 m yr^{-1} in magnitude) for all ice shelves fringing West Antarctica (SUL, LAN, GET, CD, THW, PI, COS, ABB, VEN, GEO and WOR) and in clustered parts of East Antarctica, that is, Wilkes Land (VAN, TOT, MU, POR and ADE) and Enderby Land (NE and SHA). The SBMB is relatively small (less than 1.00 m yr^{-1} in magnitude) mainly for the large ice shelves (LBC, FRIS, BRL, JF, AR, AIS and RIS), and this could be due to substantial bottom-ice accretion compensating for strong melting near the grounding line. In forming and depositing buoyant frazil ice crystals, the rising ISW benefits from troughs and cavities in the subshelf topography, which are formed downstream of peninsulas and ice rises for the Filchner-Ronne[5], Amery[24], Larsen B and Larsen C[25]

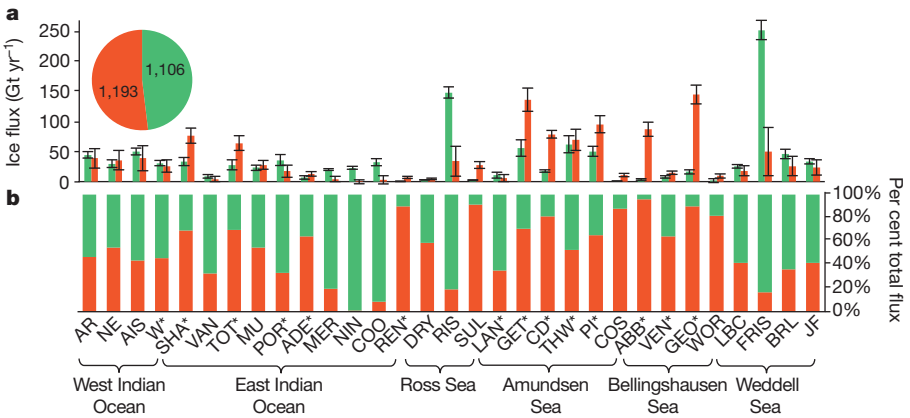


Figure 1 | Basal mass loss and calving fluxes of Antarctic ice shelves. **a**, Calving fluxes (green) and basal mass loss (−BMB; red). Pie chart shows numbers for surveyed ice shelves only. Errors, 1 s.d. **b**, Ratio between calving flux (green) and BMB (red), in per cent of total flux. Ice shelves are ordered clockwise geographically, starting from longitude 0°. Ice-shelf names: AR, Astrid-Ragnhild; NE, Northeast; AIS, Amery; W, West; SHA, Shackleton; VAN, Vanderford; TOT, Totten; MU, Moscow University; POR, Porpoise; ADE, Adélie; MER, Mertz; NIN, Ninnis; COO, Cook; REN, Rennick; DRY, Drygalski; RIS, Ross; SUL, Sulzberger; LAN, Land; GET, Getz; CD, Crosson and Dotson; THW, Thwaites; PI, Pine Island; COS, Cosgrove; ABB, Abbot; VEN, Venable; GEO, George VI; WOR, Wordie; LBC, Larsen B and Larsen C; FRIS, Filchner-Ronne; BRL, Brunt and Riiser-Larsen; JF, Jelbart and Fimbul. Asterisks indicate basins experiencing dynamic thinning[9] and ice shelves experiencing thinning[8]. See Supplementary Fig. 1 for mapped melt ratios. The brackets underneath indicate oceanic sectors.

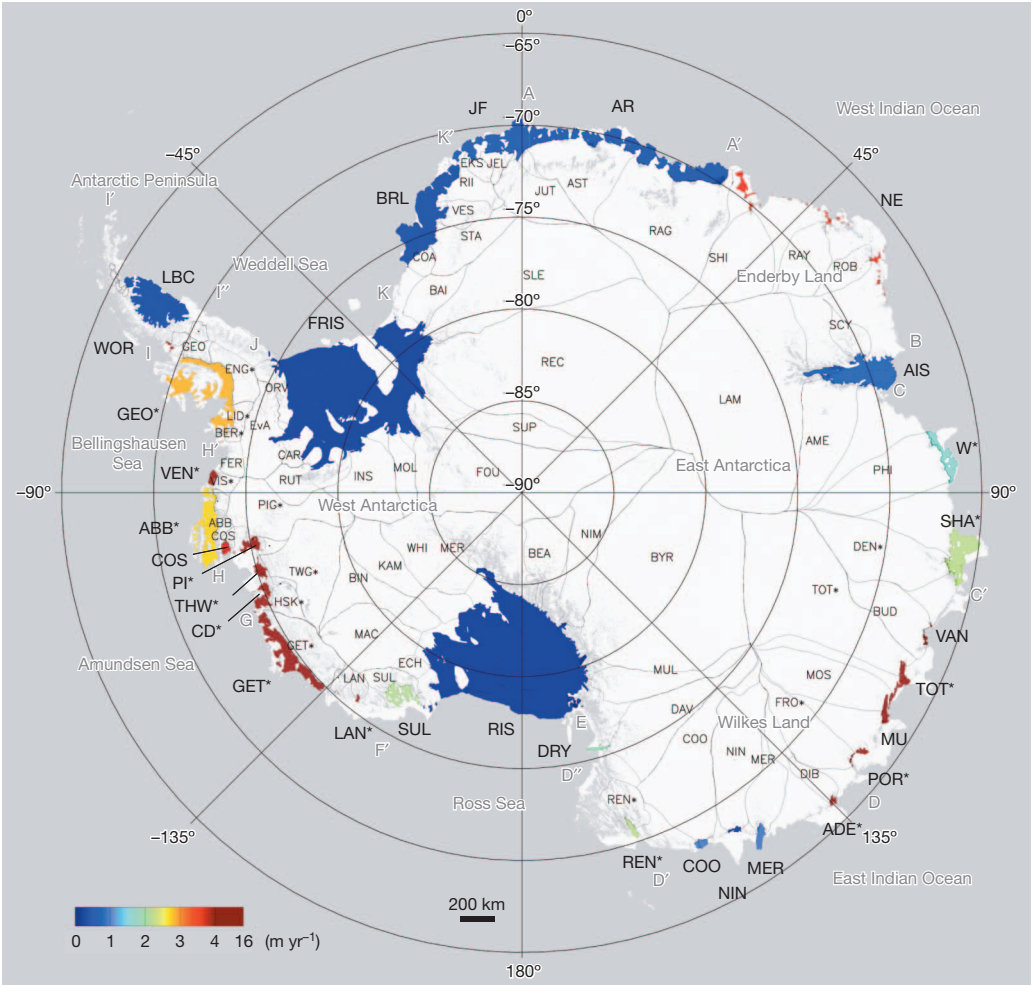


Figure 2 | Mean basal mass-loss rates of Antarctic ice shelves. Ice shelves are colour-coded for area averaged basal mass loss. Drainage basins feeding the respective ice shelves are indicated using thin black lines. Grey labels indicate oceanic sectors and major basins.

ice shelves. For smaller ice shelves, the ISW plume will not be able to precipitate and deposit marine ice to the same degree, because the subshelf topography is more homogenous and the grounding-line melt zone is closer to the ice front.

The Filchner-Ronne and Ross ice shelves have similar areas of continental shelf[3] where high-salinity shelf water forms, grounding-line lengths of ~5,100 km and respective areas of 423×10^3 and 477×10^3 km^2, integrated SMBs of 70 and 61 Gt yr^{-1} and SBMBs of −0.12 and −0.07 m yr^{-1}. Similar SBMBs indicate that the higher grounding-line melt rate on Filchner-Ronne relative to Ross, because of a deeper grounding line[4], is balanced by higher marine-ice accretion. Thus, there is relatively little marine-ice accumulation found underneath Ross[26] by comparison with the large volume under Filchner-Ronne[5].

Basal mass loss is spread quite evenly between the six oceanic sectors (131–300 Gt yr^{-1}) except for the Ross Sea and Amundsen Sea regions (79 and 484 Gt yr^{-1}, respectively) (Fig. 1 and Table 1). About 30% of all Antarctic icebergs, by mass, are formed in the Weddell Sea sector (371 Gt yr^{-1}) and only 3% are formed in the Bellingshausen Sea (41 Gt yr^{-1}). This large volume of icebergs is exported from the Weddell Sea, along the Antarctic Peninsula and into the Scotia Sea (forming the 'iceberg alley'). This may explain the increased concentration of iron maintained in the Scotia Sea as well as its high productivity compared with the mainly high-nutrient, low-chlorophyll Southern Ocean[27]. The western coastal current in combination with the geometry of the Antarctic Peninsula provides unique conditions for efficient iceberg export away from the Antarctic coast[7].

Comparing our SBMB with surface lowering rates[8] from ICESat (NASA's Ice, Cloud, and land Elevation Satellite), we find that a large negative SBMB seems to be a good indicator of ice-shelf vulnerability to oceanic forcing (Fig. 3): in this case, the incursion of warm circumpolar deep water through deep troughs. This implies that other fringing ice shelves with large negative SBMB (more than 2.00 m yr^{-1} in

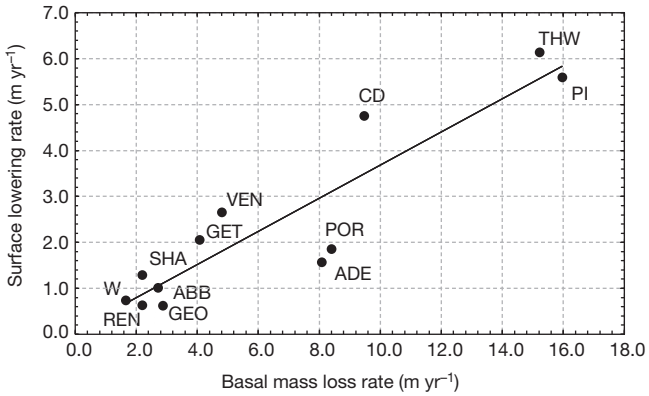


Figure 3 | Ice-shelf surface lowering rates versus mean basal mass-loss rates. We find a significant correlation ($R^2 = 0.84$ (coefficient of determination); $P = 3.13 \times 10^{-5}$; F-test) between surface lowering rates[8] and our mean basal mass-loss rates (−SBMB) for thinning ice shelves. All 12 ice shelves with a significant and extensive surface lowering rate are included.

magnitude), such as the Northeast, Vanderford, Moscow University, Totten, Sulzberger, Land and Cosgrove ice shelves (Fig. 2), could have a similar vulnerability to oceanic forcing.

Freshwater fluxes enter the Southern Ocean by different paths. Whereas basal melt water is distributed over the upper few hundred metres of the coastal water column, icebergs drift and melt farther away from the continent. Having good constraints on these fluxes and their distribution will improve our understanding of Antarctic deep-water formation and of the hydrography of the Southern Ocean. Our results will also help constrain the controls on the primary productivity of the Southern Ocean via iron fertilization[13], because bottom melt water has been linked to phytoplankton blooms[28] and melting icebergs are considered hotspots for marine life[29]. Quantifying the relative importance of bottom melt and iceberg calving is also crucial for accurately modelling the formation of sea ice. Indeed, ISW has a stabilizing effect on the water column in front of ice shelves and favours the formation of sea ice[17], whereas icebergs promote convection and mixing[7]. The poor agreement, regionally, with ice–ocean–atmosphere models indicates that further work is required before these can faithfully reproduce observed patterns of BMB and freshwater production.

METHODS SUMMARY

We use standard budget methods to calculate the BMB for each ice shelf in Antarctica: $0 = \mathrm{BMB} + \mathrm{SMB} + \mathrm{GLF} - \mathrm{CF}$. We determine BMB as the remaining unknown in the mass balance equation, assuming a steady-state front position but accounting for ice-shelf imbalance using surface elevation changes from ICESat. Calving flux is found by integrating ice-shelf thickness and ice velocity along the calving front (Supplementary Discussion 6 and Supplementary Fig. 2). Our ice-shelf thickness is based on altimetry data from the European Remote-sensing Satellite (ERS-1) for 1994–1995, and is supplemented by ICESat data for latitudes south of the ERS-1 limit for 2003–2009[19]. Elevation data are corrected to the year 2009 using elevation rates from ERS-1[30] (1994–2002) and from ICESat[8] (2003–2009) to fit velocity and IPR data sets. Ice thickness is found from freeboard elevation assuming hydrostatic equilibrium and using a correction term for the firn air content (Supplementary Discussion 8 and Supplementary Fig. 5). This correction comes from a semi-empirical model using our regional climate model, RACMO2, and depth–density observations[31]. The velocity data are an InSAR mosaic over the 2007–2009 period[20]. The GLF is obtained using InSAR velocities and ice thicknesses at, or near, the grounding line. For 68% of the GLF, thicknesses from IPR data are used. Our grounding line is a new compilation to provide the most complete and accurate coverage (Supplementary Discussion 5). Our SMB is an average over 32 years (1979–2010) from RACMO2[6] (Supplementary Discussion 7 and Supplementary Figs 3 and 4). To include the 10% of ice-sheet area unsurveyed in our total calving and melt estimates, we use the SMB of these areas and apply a regionally differentiated melting and calving ratio to them (Supplementary Discussion 3 and Supplementary Fig. 6).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 April; accepted 7 August 2013.
Published online 15 September 2013.

1. Jacobs, S. S., Helmer, H. H., Doake, C. S. M., Jenkins, A. & Frolich, R. M. Melting of ice shelves and the mass balance of Antarctica. *J. Glaciol.* **38,** 375–387 (1992).
2. Orheim, O. in *Glaciers, Ice Sheets and Sea Level: Effect of a CO2-Induced Climatic Change* 210–215 (National Academic, 1985).
3. Jenkins, A. & Doake, C. S. M. Ice-ocean interaction on Ronne Ice Shelf, Antarctica. *J. Geophys. Res.* **96,** 791–813 (1991).
4. Rignot, E. & Jacobs, S. S. Rapid bottom melting widespread near Antarctic ice sheet grounding lines. *Science* **296,** 2020–2023 (2002).
5. Joughin, I. & Padman, L. Melting and freezing beneath Filchner-Ronne Ice Shelf, Antarctica. *Geophys. Res. Lett.* **30,** 1477 (2003).
6. Lenaerts, J. T. M., van den Broeke, M. R., van de Berg, W. J., van Meijgaard, E. & Kuipers Munneke, P. A new, high-resolution surface mass balance map of Antarctica (1979–2010) based on regional atmospheric climate modeling. *Geophys. Res. Lett.* **39,** L04501 (2012).
7. Silva, T. A. M., Bigg, G. R. & Nicholls, K. W. Contribution of giant icebergs to the Southern Ocean freshwater flux. *J. Geophys. Res.* **111,** C03004 (2006).
8. Pritchard, H. D. *et al.* Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature* **484,** 502–505 (2012).
9. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461,** 971–975 (2009).
10. Dupont, T. K. & Alley, R. B. Assessment of the importance of ice-shelf buttressing to ice-sheet flow. *Geophys. Res. Lett.* **32,** L04503 (2005).
11. De Angelis, H. & Skvarca, P. Glacier surge after ice shelf collapse. *Science* **299,** 1560–1562 (2003).
12. Doake, C. S. M. & Vaughan, D. G. Rapid disintegration of the Wordie Ice Shelf in response to atmospheric warming. *Nature* **350,** 328–330 (1991).
13. Raiswell, R., Benning, L., Tranter, M. & Tulaczyk, S. Bioavailable iron in the Southern Ocean: the significance of the iceberg conveyor belt. *Geochem. Trans.* **9,** 7 (2008).
14. Yu, J., Liu, H., Jezek, K. C., Warner, R. C. & Wen, J. Analysis of velocity field, mass balance, and basal melt of the Lambert Glacier-Amery Ice Shelf system by incorporating Radarsat SAR interferometry and ICESat laser altimetry measurements. *J. Geophys. Res.* **115,** B11102 (2010).
15. Nicholls, K. W., Makinson, K. & Johnson, M. R. New oceanographic data from beneath Ronne Ice Shelf, Antarctica. *Geophys. Res. Lett.* **24,** 167–170 (1997).
16. Williams, M. J. M., Jenkins, A. & Determann, J. in *Ocean, Ice, and Atmosphere: Interactions at the Antarctic Continental Margin* 285–299 (Antarct. Res. Ser. 75, AGU, 1998).
17. Hellmer, H. H. Impact of Antarctic ice shelf basal melting on sea ice and deep ocean properties. *Geophys. Res. Lett.* **31,** L10307 (2004).
18. Timmermann, R., Wang, Q. & Hellmer, H. H. Ice-shelf basal melting in a global finite-element sea-ice/ice-shelf/ocean model. *Ann. Glaciol.* **53** (2012).
19. Griggs, J. A. & Bamber, J. L. Antarctic ice-shelf thickness from satellite radar altimetry. *J. Glaciol.* **57,** 485–498 (2011).
20. Rignot, E., Mouginot, J. & Scheuchl, B. Ice flow of the Antarctic ice sheet. *Science* **333,** 1427–1430 (2011).
21. King, M. A. *et al.* Lower satellite-gravimetry estimates of Antarctic sea-level contribution. *Nature* **491,** 586–589 (2012).
22. Hellmer, H. H., Kauker, F., Timmermann, R., Determann, J. & Rae, J. Twenty-first-century warming of a large Antarctic ice-shelf cavity by a redirected coastal current. *Nature* **485,** 225–228 (2012).
23. Potter, J. R., Paren, J. G. & Loynes, J. Glaciological and oceanographic calculations of the mass balance and oxygen isotope ratio of a melting ice shelf. *J. Glaciol.* **30,** 161–170 (1984).
24. Fricker, H. A., Popov, S., Allison, I. & Young, N. Distribution of marine ice beneath the Amery Ice Shelf. *Geophys. Res. Lett.* **28,** 2241–2244 (2001).
25. Holland, P. R., Corr, H. F. J., Vaughan, D. G., Jenkins, A. & Skvarca, P. Marine ice in Larsen ice shelf. *Geophys. Res. Lett.* **36,** L11604 (2009).
26. Zotikov, I. A., Zagorodnov, V. S. & Raikovsky, J. V. Core drilling through the Ross ice shelf (Antarctica) confirmed basal freezing. *Science* **207,** 1463–1465 (1980).
27. Whitehouse, M. J. *et al.* Substantial primary production in the land-remote region of the central and northern Scotia Sea. *Deep-Sea Res. II* **59–60,** 47–56 (2012).
28. Alderkamp, A.-C. *et al.* Iron from melting glaciers fuels phytoplankton blooms in the Amundsen Sea (Southern Ocean): phytoplankton characteristics and productivity. *Deep-Sea Res. II* **71–76,** 32–48 (2012).
29. Smith, K. L. *et al.* Free-drifting icebergs: hot spots of chemical and biological enrichment in the Weddell Sea. *Science* **317,** 478–482 (2007).
30. Zwally, H. J. *et al.* Mass changes of the Greenland and Antarctic ice sheets and shelves and contributions to sea-level rise: 1992–2002. *J. Glaciol.* **51,** 509–527 (2005).
31. Ligtenberg, S. R. M., Helsen, M. M. & van den Broeke, M. R. An improved semi-empirical model for the densification of Antarctic firn. *Cryosphere* **5,** 809–819 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by funding from the ice2sea programme of the European Union Seventh Framework Programme, grant number 226375. This work is ice2sea contribution number 139. M.R.v.d.B., J.T.M.L. and S.R.M.L. acknowledge funding from the Netherlands Polar Programme. J.L.B. was supported by NERC grant NE/I027401/1.

Author Contributions M.A.D. produced the results, led the development of the study and wrote the manuscript. J.L.B. had the idea for the study and contributed to the development of the methods, to the discussion of results and, extensively, to writing the manuscript. J.A.G. produced the calving-front elevation error and provided the ice-shelf elevation. J.T.M.L. and M.R.v.d.B. provided the SMB data and error analysis. S.R.M.L. and M.R.v.d.B. provided the firn data and error analysis. G.M. provided the grounding-line and ice-shelf mask data and discussion. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.D. (mathieu.depoorter@bristol.ac.uk) or J.L.B. (j.bamber@bristol.ac.uk).

METHODS

Budget analysis. We use standard mass budget methods to calculate the basal mass balance for each ice shelf in Antarctica, following $0 = \mathrm{BMB} + \mathrm{SMB} + \mathrm{GLF} - \mathrm{CF}$. We determine BMB as the remaining unknown of the mass balance equation, assuming a steady state. Ice-shelf thinning rates[8] are added to BMB to account for non-steady-state behaviour. Our ice-shelf thickness data set does not include shelves smaller than 100 km^2 (ref. 19). Therefore, no GLF or CF is calculated for those shelves. To include this unsurveyed 10% of the total ice-sheet area in our total calving and melting numbers, we use the surface mass balance[6] of these areas and apply a physically based regional melting and calving ratio to them (Supplementary Discussion 3 and Table 6).

Calving fluxes. The calving flux is found by integrating ice-shelf thickness and ice velocity along the calving front. The calving front has been tracked following the coastline close to the ice front, but 2–10 km inland to avoid interpolation artefacts at the ice–ocean boundary (Supplementary Fig. 2), using a combined 1-km-resolution mask of both ice-shelf thickness[19] and velocity data[20]. Our 1-km-gridded ice-shelf thickness is based on 1994–1995 altimetry data from ERS-1 and supplemented by ICESat data from south of the ERS-1 limit[19]. Elevations are corrected to the year 2009 using elevation rates from ERS-1[30] (1994–2002) and from ICESat[8] (2003–2009) to be consistent with velocity and IPR data. Thickness is found from freeboard elevation assuming hydrostatic equilibrium and using a correction term for the firn layer. The firn correction stems from a semi-empirical model using the regional atmospheric climate model RACMO2 and depth–density observations[31]. The velocity data set is an InSAR mosaic over the 2007–2009 period[20]. A sensitivity analysis for the calving-flux gate placement is provided in Supplementary Discussion 6.

Grounding lines and grounding-line fluxes. Our grounding-line data set is a compilation of published grounding lines (mainly from InSAR, but also complemented with imagery and ICESat) to achieve complete coverage and the most accurate and up-to-date delineations (Supplementary Discussion 5). The GLF is obtained using InSAR velocities and ice thicknesses close to the grounding line. We use surface-elevation-derived thicknesses for 32% of the GLF. For the remainder, we use IPR-derived thicknesses from various campaigns for the years 2009–2012 (Supplementary Table 1). IPR tracks are chosen just upstream of the grounding line (rather than downstream) to avoid the strong melting at the grounding line, and SMB is used to correct for the small area in between.

Surface mass balance. For our SMB, we use the average of 32 years (1979–2010) of SMB from RACMO2[6] run at a resolution of 27 km (Supplementary Fig. 3). This model takes into account drifting snow processes and is forced by the ERA-Interim reanalyses from the European Centre for Medium-Range Weather Forecasts. Islands and ice rises located within ice shelves are included in the ice-shelf mask[19] for SMB calculation under the assumption that net surface accumulation equals GLF for those features. There is no statistically significant temporal trend in SMB over the ice shelves[6].

Error assessment. The error assessment is done separately for each ice shelf (Fig. 1 and Supplementary Table 1). ICESat points within a zone extending 10 km upstream of the calving front are used to assess the calving-flux elevation error at the calving front of each ice shelf. To account for the time difference between the ERS-1 (1994–95) and the ICESat data (2003–2009), a dh/dt correction is applied following dh/dt trends in ERS-1 and ERS-2 for the period 1992–2001[30]. The grounding-line thickness error is assumed to be 10 m for IPR tracks. We estimate a 28% error for ice-shelf SMB and a 10% error for the firn air content correction (Supplementary Discussion 7 and 8 and Supplementary Figs 4 and 5). Errors in ice thickness derived from ice-penetrating radar, ice surface velocity, SMB and firn correction for each shelf are assumed to be uncorrelated. For SMB, this is supported by the spread of points around the least-squares linear fit shown in Supplementary Fig. 4. A random error of 3% is included in the calving-flux error to account for gate placement (Supplementary Discussion 6). In determining the total calving-flux error, we assumed that the error in surface elevation (which affects the ice thickness error) is correlated between ice shelves, because it seems to be systematic at a regional level[19]. For the GLF derived from surface elevation and the assumption of hydrostatic equilibrium, the error in thickness is found to be 10–15% in the vicinity of the grounding line[19], with both positive and negative differences. We assume an uncertainty in GLF of 20% for these areas. The error in GLF introduced from the interpolation of unsurveyed areas is determined from the root mean squared SMB error and a 10% deviation from balance for the grounded ice sheet. This is supported by the fact that elevation rates from altimetry for these unsurveyed sectors are small and they are in areas of slow flow, where changes in ice dynamics are expected to be limited[9]. The standard deviation of the differences in melt ratio between shelves experiencing similar oceanic conditions (Supplementary Discussion 3) is used as a measure of the uncertainty in partitioning the interpolated GLF between calving and BMB for the unsurveyed sectors.

Data description. The data produced for this paper are ice-shelf ice thickness and a continent-wide grounding line. These data sets can be found at http://pangaea.de/. The other data sets used in this study can be found at the following websites: http://nsidc.org/data/docs/measures/nsidc0484_rignot/ (ice velocity field), http://nsidc.org/data/icebridge/ and https://data.cresis.ku.edu/ (IPR ice thickness), http://nsidc.org/data/atlas/news/antarctic_coastlines.html (MOA grounding line and coastline), http://nsidc.org/data/docs/agdc/nsidc0469_brunt/ (ICESat grounding-line points), http://nsidc.org/data/docs/agdc/nsidc0489_bindschadler/ (ASAID project grounding line), http://nsidc.org/data/docs/measures/nsidc0498_rignot/ (DInSAR grounding line), http://nsidc.org/data/nsidc-0280.html (MOA mosaic), http://nsidc.org/data/radarsat/ramp_basics/mosaic_5kmw.html (RAMP mosaic) and http://lima.nasa.gov/ (LIMA mosaic).

doi:10.1038/nature12489

Life history trade-offs at a single locus maintain sexually selected genetic variation

Susan E. Johnston[1,2]†, Jacob Gratten[1,3]†, Camillo Berenos[2], Jill G. Pilkington[2], Tim H. Clutton-Brock[4], Josephine M. Pemberton[2] & Jon Slate[1]

Sexual selection, through intra-male competition or female choice, is assumed to be a source of strong and sustained directional selection in the wild[1,2]. In the presence of such strong directional selection, alleles enhancing a particular trait are predicted to become fixed within a population, leading to a decrease in the underlying genetic variation[3]. However, there is often considerable genetic variation underlying sexually selected traits in wild populations, and consequently, this phenomenon has become a long-discussed issue in the field of evolutionary biology[1,4,5]. In wild Soay sheep, large horns confer an advantage in strong intra-sexual competition, yet males show an inherited polymorphism for horn type and have substantial genetic variation in their horn size[6]. Here we show that most genetic variation in this trait is maintained by a trade-off between natural and sexual selection at a single gene, relaxin-like receptor 2 (*RXFP2*). We found that an allele conferring larger horns, Ho^{+}, is associated with higher reproductive success, whereas a smaller horn allele, Ho^{P}, confers increased survival, resulting in a net effect of overdominance (that is, heterozygote advantage) for fitness at *RXFP2*. The nature of this trade-off is simple relative to commonly proposed explanations for the maintenance of sexually selected traits, such as genic capture[7,8] ('good genes') and sexually antagonistic selection[5,9]. Our results demonstrate that by identifying the genetic architecture of trait variation, we can determine the principal mechanisms maintaining genetic variation in traits under strong selection and explain apparently counter-evolutionary observations.

The persistence of genetic variation in traits under sustained sexual selection is a fundamental paradox in evolutionary biology, for which several explanations have been proposed. First, sexually dimorphic characters could be an honest signal of male quality or condition, in which the best-condition males develop the largest traits[7,8]. Under this 'genic capture' model, many loci contribute to male condition, creating a large mutational target for the sexually selected trait. As a result, genetic variation will persist despite strong directional selection[7]. Second, variation may be maintained by genetic trade-offs that constrain evolution of the focal trait. For example, sexually selected traits often exceed the point at which they would be optimal for survival[10], indicating that trade-offs exist between sexual and non-sexual fitness. A variant on this model is intra-locus sexual conflict, driven by sexually antagonistic selection[5,9]; in this scenario, alleles that increase male fitness are associated with decreased female fitness (and vice versa).

Testing and disentangling which of the theories explains empirical patterns remains difficult owing to the relatively limited knowledge of the genetic architecture (that is, the number of genes and the magnitude of their effects) of relevant traits[11,12], as there are remarkably few systems where the genes responsible for sexually selected trait variation are known. However, the advent of affordable genomic technologies now provides an unparalleled opportunity to identify genes responsible for fitness-related variation in wild populations[13–15].

One wild population where the genetic architecture of a sexually selected trait has been characterized is the Soay sheep of St Kilda (*Ovis aries*), a population of primitive domestic sheep that has existed completely unmanaged for around 4,000 years and has been intensively studied since 1985. During the mating season (rut), there is strong competition between males for access to oestrous females. Most males develop normal horns, but around 13% develop vestigial horns (scurs) conferring reduced reproductive success[16] (Fig. 1). The horn size of normal-horned males is positively correlated with mating success[17], yet there is substantial heritable variation underlying this trait ($h^2 = 0.37$)[6]. Females develop much smaller horns, and are either normal-horned (32%), scurred (40%), or 'polled' (lacking horns or scurs, 28%; Supplementary Fig. 1). Recent studies have revealed that a single gene, relaxin-like receptor 2 (*RXFP2*), explains most of the genetic variation in horn morphology in Soay sheep[6] and domestic sheep[18,19]. Two *RXFP2* alleles, Ho^{+} and Ho^{P}, have been identified in Soay sheep: Ho^{+} confers larger, normal horns, whereas Ho^{P} confers smaller horns, with around half of $Ho^{P}Ho^{P}$ males developing scurs[6] (Fig. 1). Furthermore, *RXFP2* contributes ~76% of the additive genetic variation in horn size in normal-horned males, including normal-horned $Ho^{P}Ho^{P}$ males[6] (Fig. 1 and Supplementary Note 1). This discovery provides a critical opportunity to understand the relative importance of sexual and natural selection in maintaining genetic variation underlying horn development.

Figure 1 | Horn morphology variation with *RXFP2* genotype. Examples of adult male horn morphology with their corresponding *RXFP2* genotypes. **a**, Four-year-old normal-horned $Ho^{+}Ho^{+}$. **b**, Five-year-old normal-horned $Ho^{+}Ho^{P}$. **c**, Five-year-old normal-horned $Ho^{P}Ho^{P}$. **d**, Three-year-old scurred $Ho^{P}Ho^{P}$.

[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK. [2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. [3]Queensland Brain Institute, University of Queensland, Brisbane 4072, Australia. [4]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. †Present addresses: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK (S.E.J.); Queensland Brain Institute, University of Queensland, Brisbane 4072, Australia (J.G.).

We used genetic and phenotypic data from 1,750 sheep sampled over a 21-year period to understand how genetic variation at *RXFP2* is maintained and to determine whether microevolution of horns is occurring. To achieve this, we: (1) examined differences in annual reproductive success, survival and overall fitness between *RXFP2* genotypes; (2) determined selection coefficients and equilibrium frequencies of the two *RXFP2* alleles; and (3) assessed whether temporal changes in allele frequency were consistent with balancing selection maintaining genetic variation at *RXFP2*.

The *RXFP2* genotype was associated with annual reproductive success in adult males, as Ho^PHo^P males had significantly lower reproductive success than both Ho^+Ho^+ and Ho^+Ho^P males (Markov chain Monte Carlo (MCMC) generalized linear mixed model: $P_{MCMC} = 0.004$; Fig. 2a). The *RXFP2* genotype was also associated with male annual survival, such that Ho^+Ho^+ individuals had lower survival than both Ho^+Ho^P and Ho^PHo^P individuals (MCMC generalized linear mixed model: $P_{MCMC} = 0.006$ and 0.010, respectively; Fig. 2b). When reproductive success and survival were combined into an overall fitness metric, Ho^+Ho^P males had higher fitness than both Ho^+Ho^+ and Ho^PHo^P individuals (MCMC generalized linear mixed model: $P_{MCMC} = 0.042$ and 0.036, respectively; Fig. 2c). In females, there was no relationship between *RXFP2* genotype and survival, reproductive success or overall fitness (MCMC generalized linear mixed models: $P_{MCMC} > 0.05$). Full results for all models are given in Supplementary Tables 1 and 2.

Analyses of overall fitness of all sheep indicated that variation at *RXFP2* is maintained by overdominance, resulting in an equilibrium frequency of 0.529 for the Ho^P allele (bootstrap 95% confidence interval: 0.284–0.793). This is driven by each homozygous genotype being associated with reduced fitness (that is, Ho^+Ho^+ with reduced survival and Ho^PHo^P with reduced reproductive success), which has resulted in a net effect of highest overall fitness in Ho^+Ho^P individuals (Supplementary Table 3). The products of effective population size and selection coefficients were considerably greater than one for annual reproductive success, survival and overall fitness, confirming that selection is probably more important than genetic drift in maintaining genetic variation at *RXFP2* within this population[20,21] (Supplementary Table 3).

Although the frequency of the Ho^P allele increased over the study period (linear regression: $b = 0.426\%$ per year, adjusted $R^2 = 0.653$, $P = 1.74 \times 10^{-5}$; Supplementary Fig. 2), simulations of the expected change in frequency given the pedigree indicated that this increase is not greater than would be expected by chance (gene-drop simulation, one-tailed $P = 0.109$), further showing that the change in frequency is unlikely to be driven by directional selection. The observations that selection coefficients are strong enough to exceed the effects of drift, yet temporal changes in allele frequency are no greater than expected by drift, may seem contradictory. However, because selection coefficients on reproduction and survival are in opposite directions (from the context of a homozygous genotype) with a net effect of overdominance, marked temporal shifts in allele frequencies are not expected to occur. In fact, allele frequencies seem to have converged on and possibly stabilized at something close to the equilibrium frequency (Supplementary Fig. 2). Furthermore, analysis of *RXFP2* haplotype sharing between Soay sheep and other breeds of *Ovis aries* indicates that the polymorphism may have been present in Soay sheep throughout much of their long history on St Kilda (Supplementary Note 2). Therefore, as a result of the large contribution of *RXFP2* to horn growth, genetic variation in horn morphology is maintained by a trade-off between reproductive success (favouring the large-horn allele Ho^+) and survival (favouring the small-horn allele Ho^P) in male Soay sheep.

The analyses presented here have allowed an empirical test of the key hypotheses proposed to explain genetic variation in sexually selected traits. As a single locus explains nearly all of the genetic variation in horn type and size[6] (see Supplementary Note 1), the genic capture hypothesis can be discounted. More plausible explanations involving major loci, particularly intra-locus sexually antagonistic selection, were also ruled out as *RXFP2* is associated with fitness components in males, but not in females. Instead, variation at *RXFP2* seems to be maintained by a simple trade-off in males, where only heterozygous males (Ho^+Ho^P) are uncompromised by poor survival or low reproductive success; different horn types in females are merely a genetic consequence of the trade-off in males. Investigation of selection coefficients supports the idea that there is a net effect of overdominance, or heterozygote advantage, at this locus in Soay sheep. Despite being one of the most intuitive explanations for the maintenance of genetic variation, convincing examples of overdominance remain rare[22,23]. Therefore, selection on *RXFP2* in male Soay sheep may be an additional entry to what remains a short list of compelling cases of heterozygote advantage[24].

The exact mechanism by which the *RXFP2* genotype influences variation in male survival and reproductive success in Soay sheep is unknown. However, it does seem to be mediated by the effect of *RXFP2* on horns, as scurred Ho^PHo^P males have lower reproductive success than normal-horned Ho^PHo^P males (Supplementary Note 3). Fitness variation at *RXFP2* may be due to differences in energy expenditure by the three genotypes in relation to their horn size in males, particularly during the rut. Larger, normal-horned males can spend 50% of their time holding individual female consorts, but spend less time feeding and must defend their consort against harassment by juveniles and competition from other dominant males[25,26]. Subordinate males, including juveniles and smaller horned males, can spend just 5% of their time in consorts and instead mate opportunistically, actively seeking undefended females. Therefore, as *RXFP2* genotype directly affects horn phenotype, this may determine whether males are likely to engage in a mating strategy with high annual reproductive success but low survival, or a strategy with low annual reproductive success but high survival.

Ultimately, the findings in this study advance our understanding of the maintenance of genetic variance in a trait under natural and sexual

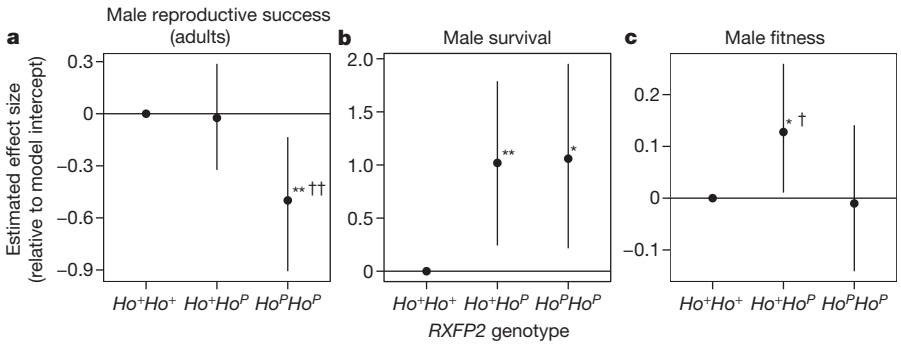


Figure 2 | Annual fitness variation and *RXFP2* genotype. **a**, Reproductive success in adult males ($n = 640$). **b**, Survival in all males ($n = 1{,}243$). **c**, Overall fitness in all males ($n = 1{,}204$). Effect sizes were estimated from the posterior mode of a MCMC generalized linear mixed model and are given relative to the model intercept at Ho^+Ho^+. Vertical bars indicate the 95% credible interval. The single asterisk and double asterisk indicate a significant difference from the intercept at Ho^+Ho^+ at $P_{MCMC} = 0.05$ and 0.01, respectively; single and double daggers indicate the same for the model intercept at Ho^+Ho^P.

selection in a wild population. The ability to identify the gene responsible for sexually selected variation has made it possible to distinguish between competing hypotheses for the persistence of horn variation and investigate an evolutionary response (or stasis). Further studies that dissect the genetic architecture of sexually selected traits in other systems will make it possible to establish which mechanisms are of most general relevance to explaining the evolution of sexually selected traits.

METHODS SUMMARY

Study population. The Soay sheep of the St Kilda archipelago (57° 49′ N, 8° 34′ W) are a feral population of Neolithic domestic sheep[27] that have been studied on an individual basis since 1985.

Estimation of reproductive success. A total of 5,880 sheep sampled between 1980 and 2012 were genotyped at 51,135 single nucleotide polymorphisms (SNPs) on an Ovine SNP50 BeadChip[28]. Of the 38,404 polymorphic loci that passed quality control, 315 informative, unlinked SNPs from the chip were used to construct a pedigree to determine individual reproductive success.

***RXFP2* genotyping.** A total of 1,750 sheep (796 males and 954 females) sampled between 1990 and 2008 with information on annual reproductive success and annual survival up to 2011 were genotyped at a diagnostic SNP in strong linkage disequilibrium with the putative *RXFP2* alleles, Ho^{+} and Ho^{P} (ref. 6).

Detecting fitness differences between *RXFP2* genotypes. The relationship between *RXFP2* genotype and annual reproductive success, survival and overall fitness was modelled in males and females using a generalized linear mixed model framework with a Markov chain Monte Carlo (MCMC) method[29].

Determining selection coefficients at *RXFP2*. Mean values of survival, reproductive success and overall fitness for each *RXFP2* genotype were calculated from the raw data[3]. Relative fitness values were determined by dividing the mean value for all three genotypes by the maximum value (that is, the genotype with the highest fitness has a relative fitness equal to 1) and selection coefficients and equilibrium frequencies were determined using dominance or partial dominance models where appropriate[3].

Temporal trend in *RXFP2* allele frequencies. Gene-drop simulations ($n = 1{,}000$ iterations) were used to model the expected change in frequency of the allele Ho^{P} due to genetic drift only (that is, absence of directional selection) between 1990 and 2008 in pedigreed individuals[30].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 January; accepted 18 July 2013.

Published online 21 August 2013.

1. Pomiankowski, A. & Møller, A. P. A resolution of the lek paradox. *Proc. R. Soc. Lond. B* **260,** 21–29 (1995).
2. Kingsolver, J. G. *et al.* The strength of phenotypic selection in natural populations. *Am. Nat.* **157,** 245–261 (2001).
3. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* (Longman, 1996).
4. Promislow, D. E. L. Costs of sexual selection in natural populations of mammals. *Proc. R. Soc. Lond. B* **247,** 203–210 (1992).
5. Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends Ecol. Evol.* **24,** 280–288 (2009).
6. Johnston, S. E. *et al.* Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol. Ecol.* **20,** 2555–2566 (2011).
7. Rowe, L. & Houle, D. The lek paradox and the capture of genetic variance by condition dependent traits. *Proc. R. Soc. Lond. B* **263,** 1415–1421 (1996).
8. Tomkins, J. L., Radwan, J., Kotiaho, J. S. & Tregenza, T. Genic capture and resolving the lek paradox. *Trends Ecol. Evol.* **19,** 323–328 (2004).
9. Chippindale, A. K., Gibson, J. R. & Rice, W. R. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in *Drosophila. Proc. Natl Acad. Sci. USA* **98,** 1671–1675 (2001).
10. Andersson, M. *Sexual Selection* (Princeton Univ. Press, 1994).
11. Kruuk, L. E. B., Slate, J. & Wilson, A. J. New answers for old questions: the evolutionary quantitative genetics of wild animal populations. *Annu. Rev. Ecol. Evol. Syst.* **39,** 525–548 (2008).
12. Chenoweth, S. F. & McGuigan, K. The genetic basis of sexually selected variation. *Annu. Rev. Ecol. Evol. Syst.* **41,** 81–101 (2010).
13. Stapley, J. *et al.* Adaptation genomics: the next generation. *Trends Ecol. Evol.* **25,** 705–712 (2010).
14. Slate, J. *et al.* Genome mapping in intensively studied wild vertebrate populations. *Trends Genet.* **26,** 275–284 (2010).
15. Ellegren, H. & Sheldon, B. C. Genetic basis of fitness differences in natural populations. *Nature* **452,** 169–175 (2008).
16. Robinson, M. R., Pilkington, J. G., Clutton-Brock, T. H., Pemberton, J. M. & Kruuk, L. E. B. Live fast, die young: trade-offs between fitness components and sexually antagonistic selection on weaponry in Soay sheep. *Evolution* **60,** 2168–2181 (2006).
17. Preston, B. T., Stevenson, I. R., Pemberton, J. M., Coltman, D. W. & Wilson, K. Overt and covert competition in a promiscuous mammal: the importance of weaponry and testes size to male reproductive success. *Proc. R. Soc. Lond. B* **270,** 633–640 (2003).
18. Dominik, S., Henshall, J. M. & Hayes, B. J. A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. *Anim. Genet.* **43,** 468–470 (2011).
19. Kijas, J. W. *et al.* Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* **10,** e1001258 (2012).
20. Hedrick, P. *Genetics of Populations* (Jones and Bartlett, 2005).
21. Connallon, T. & Clark, A. G. A general population genetic framework for antagonistic selection that accounts for demography and recurrent mutation. *Genetics* **190,** 1477–1489 (2012).
22. Allison, A. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* **1,** 290–294 (1954).
23. Greaves, J. H., Redfern, R., Ayres, P. B. & Gill, J. E. Warfarin resistance: a balanced polymorphism in the Norway rat. *Genet. Res.* **30,** 257–263 (1977).
24. Gemmell, N. J. & Slate, J. Heterozygote advantage for fecundity. *PLoS ONE* **1,** e125 (2006).
25. Grubb, P. *Island Survivors: the Ecology of the Soay Sheep of St Kilda*, Ch. 8, 195–223 (Athlone Press, 1974).
26. Stevenson, I. R., Marrow, B., Preston, B. T., Pemberton, J. M. & Wilson, K. *Soay Sheep: Dynamics and Selection in an Island Population*, Ch. 9 243–275 (Cambridge Univ. Press, 2004).
27. Chessa, B. *et al.* Revealing the history of sheep domestication using retrovirus integrations. *Science* **324,** 532–536 (2009).
28. Kijas, J. W. *et al.* A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS ONE* **4,** e4668 (2009).
29. Hadfield, J. MCMC methods for multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. *J. Stat. Softw.* **33,** 1–22 (2010).
30. Gratten, J. *et al.* Selection and microevolution of coat pattern are cryptic in a wild population of sheep. *Mol. Ecol.* **21,** 2977–2990 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the numerous Soay sheep project members and volunteers for collection of data and samples; M. Robinson, J. Hadfield, D. Childs and D. Nussey for statistical advice and discussions; J. McEwan, N. Pickering and J. Kijas for SNP information; D. Beraldi, E. Brown and P. Ellis for laboratory assistance; L. Evenden, J. Gibson and L. Murphy at the Wellcome Trust Clinical Research Facility Genetics Core for genome-wide SNP genotypes; I. Stevenson for database development; G. Prior and A. Ozgul for images; National Trust for Scotland and Scottish Natural Heritage for permission to work on St Kilda; and QinetiQ and Eurest for logistical support. The Soay sheep project is funded by the Natural Environment Research Council (NERC). SNP genotyping was funded by NERC and the European Research Council (ERC). S.E.J. was funded by a Biotechnology and Biological Sciences Research Council CASE studentship.

Author Contributions J.G.P., T.H.C.-B. and J.M.P. organized the long-term collection of phenotypic data and DNA samples. S.E.J. and J.S. designed the study. S.E.J., C.B. and J.G. performed laboratory work and C.B. constructed the pedigree. S.E.J. and J.G. analysed the data. S.E.J. and J.S. wrote the paper and all authors contributed to revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.E.J. (Susan.Johnston@ed.ac.uk) or J.S. (j.slate@shef.ac.uk).

METHODS

Study population. The Soay sheep of the St Kilda archipelago (57° 49′ N, 8° 34′ W) are a feral population of Neolithic domestic sheep[27] that have been studied on an individual basis since 1985. Animals were handled in strict accordance with UK Home Office ethical regulations and all work was licensed under the UK Animals (Scientific Procedures) Act 1986.

Pedigree construction. To determine individual reproductive success measures, a pedigree was constructed using molecular data for both maternity and paternity. Sheep with available blood and/or tissue samples ($n = 5{,}880$) were typed at 51,135 SNPs on the Ovine SNP50 BeadChip[28] using an Illumina Bead Array genotyping platform. Highly informative and unlinked SNPs were selected for parentage analysis by carrying out linkage-disequilibrium-based SNP pruning in the software PLINK v1.07 (ref. 31), with the following parameters: SNPs with minor allele frequency >0.4 retained, variance inflation factor = 1.01 and sliding windows of 50 SNPs with the window shifted 5 SNPs at each step. 315 SNP loci with a pairwise R^2 of <0.01 passed these criteria and were used for pedigree construction. Maternity and paternity of 5,626 sheep born between 1980 and 2012 were inferred simultaneously in the R package MasterBayes[32] implemented in R v2.14.0, with 20,000 iterations, a burn-in of 5,000 iterations and a thinning interval of 10 iterations. Sheep were included in the list of candidate parents if still alive during the rut preceding the spring the lamb was born in, and candidate parents were discarded if showing more than 8 mismatches with the progeny. Some males were not genotyped on the Ovine SNP50 BeadChip, but had previously been included in pedigrees constructed using 14-18 microsatellite loci[33]; these fathers were retained in the current pedigree when they had been assigned with >95% confidence, and there was no more than one mismatch between father and offspring or between the mother–father–offspring trio over all microsatellite loci.

Phenotypic data. Annual life history data were collected for Soay sheep during the period 1985 to 2011. Survival and reproductive success was recorded from November of a given year to the November of the following year for all individuals where this information was available; survival was scored as a binary trait and reproductive success was scored as the number of offspring surviving to at least the first November after birth (≥6 months old). Reproductive success was calculated for males only when they had at least one paternity or had been observed in the study area during the rut. A measure of annual 'overall fitness' was also calculated, where the contribution of an individual to the population count in the following year was calculated as the sum of individual survival and half of the number of offspring; this measure was transformed into integer values by multiplying all values by two. Models of fitness also included the covariate body weight (kg), which was measured during the August of the year the fitness measure was made. In total, there were 1,310 records on 796 males and 2,782 records on 954 females where all measures were available.

***RXFP2* genotyping.** A total of 1,750 sheep (796 males and 954 females) sampled between 1990 and 2008 with known phenotypic information were genotyped at a single diagnostic SNP in the 3′ untranslated region of *RXFP2* on ovine chromosome 10 (SNP ID: G100364-AS001072, developed by AgResearch Ltd, Invermay Agricultural Centre). The alleles at this diagnostic SNP are in strong linkage disequilibrium with putative alleles Ho^+ and Ho^P described in previous gene mapping papers; therefore, we refer to the diagnostic SNP genotypes in terms of their predicted genotype at *RXFP2* (ref. 6). We confirmed (1) the association of *RXFP2* with horn phenotype and (2) the suitability of use of this diagnostic SNP in this study: methods and results are presented in Supplementary Note 1. The diagnostic SNP was genotyped using a multiplex SNP-SCALE protocol[6,34], with additional animals typed on an Illumina BeadXpress Veracode GoldenGate SNP typing platform.

Detecting fitness differences between *RXFP2* genotypes using a mixed-model framework. The relationships between *RXFP2* genotype and annual reproductive success, annual survival and annual overall fitness were modelled using a generalized linear mixed model framework (GLMM) with an MCMC method in the R package MCMCglmm[29] in R v2.15.3. Models were fitted in males and females separately owing to differences in the distribution of fitness measures between the sexes. In males, reproductive success was modelled in all males and separately in lambs (age 1) and in adults (age 2+), as lamb males are much less likely to sire offspring. Survival was modelled with a binomial error structure, and reproductive success and overall fitness were modelled with Poisson error structures. Fixed effects included the age of measurement as a linear function, body weight and *RXFP2* genotype (fitted as a three level factor corresponding to genotypes Ho^+Ho^+, Ho^+Ho^P and Ho^PHo^P). Random effects were the animal identity (to account for repeated measures on individual sheep), year of birth and the year of fitness measurement, where the two latter effects accounted for variation attributed to specific environmental effects associated with these years. Effect sizes were estimated with posterior specification of previous distributions for random effect structures in the model. All models were run for N iterations and sampled 1,000 times after burn-in (see Supplementary Table 1 for specific values of N, burn-in and thinning interval for each model). Models were accepted if the independence of the samples in the posterior distribution (that is, the autocorrelation) was <0.1. Effect sizes and credible intervals for fixed and random effects were estimated from the posterior mode of the sampled iterations.

Determining selection coefficients at *RXFP2*. To determine selection coefficients and the expected frequency of *RXFP2* at equilibrium, it was necessary to calculate the relative fitness values of the genotypes, rather than their absolute values[3]. Therefore, we calculated the mean value of annual survival, annual reproductive success and annual overall fitness for each *RXFP2* genotype from the raw data in all sheep of both sexes ($n = 2{,}523$). Relative fitness values were determined by dividing the mean value for all three genotypes by the maximum value (that is, the genotype with the highest fitness will always have relative fitness equal to 1). In dominance or partial dominance models (that is, the two homozygotes are the most and least fit of the three genotypes), the relative fitnesses of the heterozygote and the less fit homozygote are denoted as $1 - hs$ and $1 - s$, respectively, where hs and s are their respective selection coefficients. In cases where heterozygotes have the highest fitness (known as overdominance or heterozygote advantage), the relative fitnesses of each homozygote are denoted as $1 - s_1$ and $1 - s_2$, respectively, where s_1 and s_2 are the selection coefficients for their respective homozygotes. If overdominance exists, then an equilibrium frequency can be calculated[3] as $q = s_2/(s_1 + s_2)$. If the product of selection coefficients and effective population size (N_e) is much greater than 1, then selection is likely to be more important in determining changes in allele frequency than drift; N_e was previously estimated as 194 individuals using data from a subset of 190 Soay sheep typed on the Ovine SNP50 Beadchip[28]. The 95% confidence interval for each selection coefficient was estimated using a bootstrap method resampling the data 1,000 times, where the fitness values were calculated relative to a value of 1 for the heterozygote, and s_1 and s_2 calculated as for the overdominance model.

Temporal trend in *RXFP2* allele frequencies. Gene-drop simulations ($n = 1{,}000$ iterations) were used to model the expected change in frequency of the allele Ho^P due to genetic drift only (that is, absence of directional selection) between 1990 and 2008, in all pedigreed individuals during this period ($n = 4{,}738$), given an identical pedigree to the true one[30]. For each parent, a randomly chosen allele was transmitted to their offspring; individuals with one or two unknown parents were assigned the unknown allele according to the population allele frequencies in the year they were born. For each simulated pedigree, the change in allele frequency over the study period was modelled using linear regression in animals that had been genotyped at *RXFP2*, generating a distribution of regression slopes that are expected due to drift but no selection. The probability of obtaining the observed change in frequency due to drift alone was determined by comparing the slope of the linear regression in the observed data with the distribution of slopes from the gene-drop simulations. All simulations and regressions were performed in R v2.15.3.

31. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
32. Hadfield, J. D., Richardson, D. S. & Burke, T. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol. Ecol.* **15,** 3715–3730 (2006).
33. Hayward, A. D. *et al.* Natural selection on a measure of parasite resistance varies across ages and environmental conditions in a wild mammal. *J. Evol. Biol.* **24,** 1664–1676 (2011).
34. Kenta, T. *et al.* Multiplex SNP-SCALE: a cost-effective medium-throughput single nucleotide polymorphism genotyping method. *Mol. Ecol. Resour.* **8,** 1230–1238 (2008).

doi:10.1038/nature12503

Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens

Katharine M. Ng[1]*, Jessica A. Ferreyra[1]*, Steven K. Higginbottom[1], Jonathan B. Lynch[1], Purna C. Kashyap[1]†, Smita Gopinath[1], Natasha Naidu[2], Biswa Choudhury[2], Bart C. Weimer[3], Denise M. Monack[1] & Justin L. Sonnenburg[1]

The human intestine, colonized by a dense community of resident microbes, is a frequent target of bacterial pathogens. Undisturbed, this intestinal microbiota provides protection from bacterial infections. Conversely, disruption of the microbiota with oral antibiotics often precedes the emergence of several enteric pathogens[1–4]. How pathogens capitalize upon the failure of microbiota-afforded protection is largely unknown. Here we show that two antibiotic-associated pathogens, *Salmonella enterica* serovar Typhimurium (*S. typhimurium*) and *Clostridium difficile*, use a common strategy of catabolizing microbiota-liberated mucosal carbohydrates during their expansion within the gut. *S. typhimurium* accesses fucose and sialic acid within the lumen of the gut in a microbiota-dependent manner, and genetic ablation of the respective catabolic pathways reduces its competitiveness *in vivo*. Similarly, *C. difficile* expansion is aided by microbiota-induced elevation of sialic acid levels *in vivo*. Colonization of gnotobiotic mice with a sialidase-deficient mutant of *Bacteroides thetaiotaomicron*, a model gut symbiont, reduces free sialic acid levels resulting in *C. difficile* downregulating its sialic acid catabolic pathway and exhibiting impaired expansion. These effects are reversed by exogenous dietary administration of free sialic acid. Furthermore, antibiotic treatment of conventional mice induces a spike in free sialic acid and mutants of both *Salmonella* and *C. difficile* that are unable to catabolize sialic acid exhibit impaired expansion. These data show that antibiotic-induced disruption of the resident microbiota and subsequent alteration in mucosal carbohydrate availability are exploited by these two distantly related enteric pathogens in a similar manner. This insight suggests new therapeutic approaches for preventing diseases caused by antibiotic-associated pathogens.

The intestinal microbiota is composed of trillions of microbial cells that together form a complex, dynamic and highly competitive ecosystem[5,6]. Limited nutrients and high microbial densities are likely to have a key role in protecting the host against invading microbes[7]. Carbohydrates derived from diet or host play a well-established part in sustaining the resident members of the microbiota[8–10], and more recently have been shown to have important roles in gut microbiota–pathogen dynamics[11–14]. Oral antibiotic use is one of the leading risk factors for disease associated with *Salmonella* spp. and *Clostridium difficile*, consistent with increased enteric vulnerability upon disruption of the resident microbiota[1–4,15]. In addition, mouse models of *S. typhimurium* or *C. difficile* infection commonly require disruption of the intestinal microbiota with antibiotics to promote pathogen expansion within the lumen of the gut and to initiate disease[16–19]. Deciphering the numerous mechanisms by which the microbiota prevents bacterial pathogen expansion and how microbiota disruption enables pathogens to circumvent these mechanisms remains an important task.

We used transcriptional profiling of *Salmonella typhimurium* from orally infected gnotobiotic mice to gain insight into the pathogen's biology while inhabiting the gastrointestinal tract. Our goal was to reveal adaptations of the pathogen within a 'low-complexity' gnotobiotic microbiota that might be relevant to antibiotic-induced microbiota disruption. Mice that were monoassociated with the model gut symbiont *Bacteroides thetatiotaomicron* (*Bt*) were used as a simplified model of a microbiota that is susceptible to pathogen emergence within the gut. Five days after *S. typhimurium* infection of the *Bt*-monoassociated or germ-free mice (Fig. 1a), caecal contents were collected and subjected to transcriptional profiling using a custom *S. typhimurium* Gene Chip. In the presence of *Bt*, all 59 *S. typhimurium* genes that displayed significantly altered expression relative to infection of germ-free mice were upregulated (Supplementary Table 1). Functional classification of these genes revealed enriched cluster of orthologous groups (COG) categories: 'carbohydrate metabolism and transport' and 'secondary metabolites biosynthesis, transport, and catabolism' (Supplementary Fig. 2). Genes encoding host mucin carbohydrate metabolism pathways are prominently represented in this gene set, including three operons encoding catabolic pathways for sialic acid, fucose and the fucose catabolite propanediol (*nan*, *fuc* and *pdu*, respectively) (Fig. 1b). We surveyed expression of genes within the *nan* and *fuc* operons 1 day after *S. typhimurium* infection in germ-free or *Bt*-monoassociated mice, to determine whether these operons identified by expression profiling on day 5 post infection also display high expression earlier in the infection. *S. typhimurium nanE* and *fucI* are significantly upregulated 1 day after infection of *Bt*-monoassociated mice relative to infection of germ-free mice (*nanE*, 6.0-fold, $P = 1.47 \times 10^{-5}$; *fucI*, 3.5-fold, $P = 0.0028$) (Fig. 1c) when *S. typhimurium* densities and host pathology are similar between colonization states (Supplementary Figs 3 and 4). These data are consistent with *S. typhimurium* catabolizing sialic acid and fucose in the lumen of the gut in a *Bt*-dependent manner soon after infection.

We next constructed mutant strains of *S. typhimurium* to quantitatively assess the requirement of sialic acid and fucose during expansion *in vivo*. Deletion of *nanA* and *fucI*, the first committed steps in the sialic acid and fucose utilization pathways, abolished growth of the strains on the respective sugars (Supplementary Fig. 5). In competition experiments, *Bt*-monoassociated mice co-infected with wild-type *S. typhimurium* and a *nanA*/*fucI* double mutant strain (*St*-Δ*nanA*Δ*fucI*) revealed that the mutant had a significant disadvantage on days 1 and 2 after infection (day 1, competitive index (CI) = 1.87, $P = 0.028$; day 2, CI = 1.45, $P = 0.016$; Fig. 1d). This mutant, however, displayed no competitive disadvantage when competing with wild-type *S. typhimurium* within germ-free mice, consistent with *S. typhimurium*'s sialic acid and fucose use being microbiota-dependent (day 1, $P = 0.26$). The competitive index was not significantly different between the two colonization conditions (Supplementary Fig. 5), however this is probably because of the small amount of free sialic acid present in the germ-free mouse gut (see Fig. 2a).

[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Glycobiology Research and Training Center, University of California, San Diego, California 92093, USA. [3]Department of Population Health and Reproduction, University of California, Davis, California 95616, USA. †Present address: Department of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota 55905, USA.
*These authors contributed equally to this work.

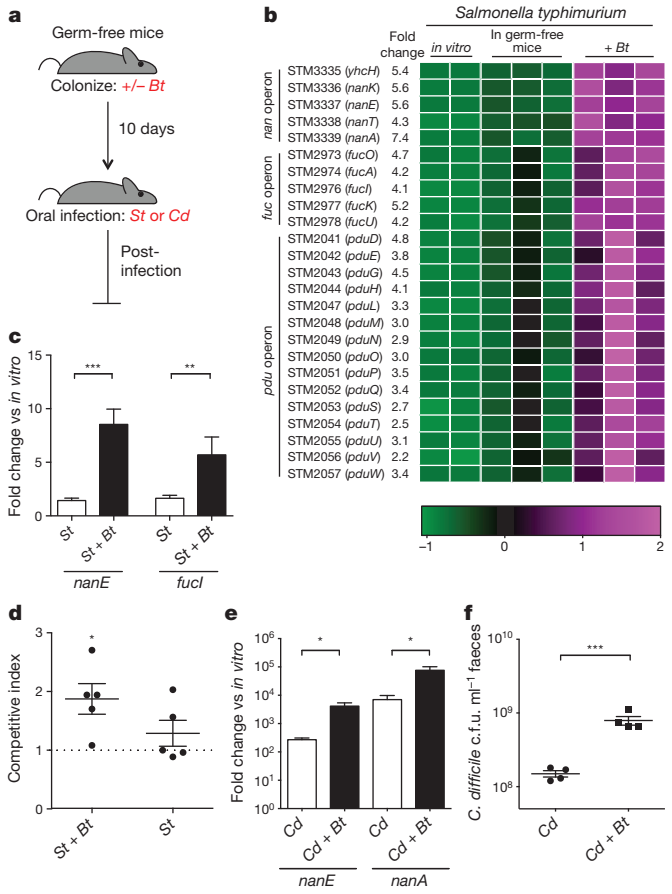


Figure 1 | *B. thetaiotaomicron* facilitates *S. typhimurium* and *C. difficile* carbohydrate utilization during emergence. **a**, Schematic of mouse infection experiments. *Bt, B. thetaiotaomicron*; *Cd, C. difficile*; *St, S. typhimurium*. **b**, *S. typhimurium* operons displaying significant differences in gene expression levels *in vivo* in the presence and absence of *B. thetaiotaomicron,* 5 days post infection. Colours indicate the deviation of each gene's signal above (purple) and below (green) its mean expression value across all six *in vivo* samples and duplicate *in vitro* growths conducted in minimal medium. **c**, Induction of *S. typhimurium nanE* and *fucI* in caecal contents 1 day post infection relative to growth in LB broth ($n = 9$ and 4 for *St* and *St* + *Bt*, respectively). **d**, Competitive index (CI) of wild-type *S. typhimurium*/*St*-Δ*nanA*Δ*fucI* in *Bt*-monoassociated (*St* + *Bt*) and ex-germ-free (*St*) mice 1 day post infection. Horizontal bars indicate the geometric means of CI values and individual CI values are represented with dots ($n = 5$ per group). **e**, Induction of *C. difficile nan* genes in caecal contents 3 days post infection relative to growth in minimal medium containing 0.5% glucose ($n = 4$ per group). **f**, *C. difficile* density in faeces 1 day post infection ($n = 4$ per group). Error bars indicate s.e.m. $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$.

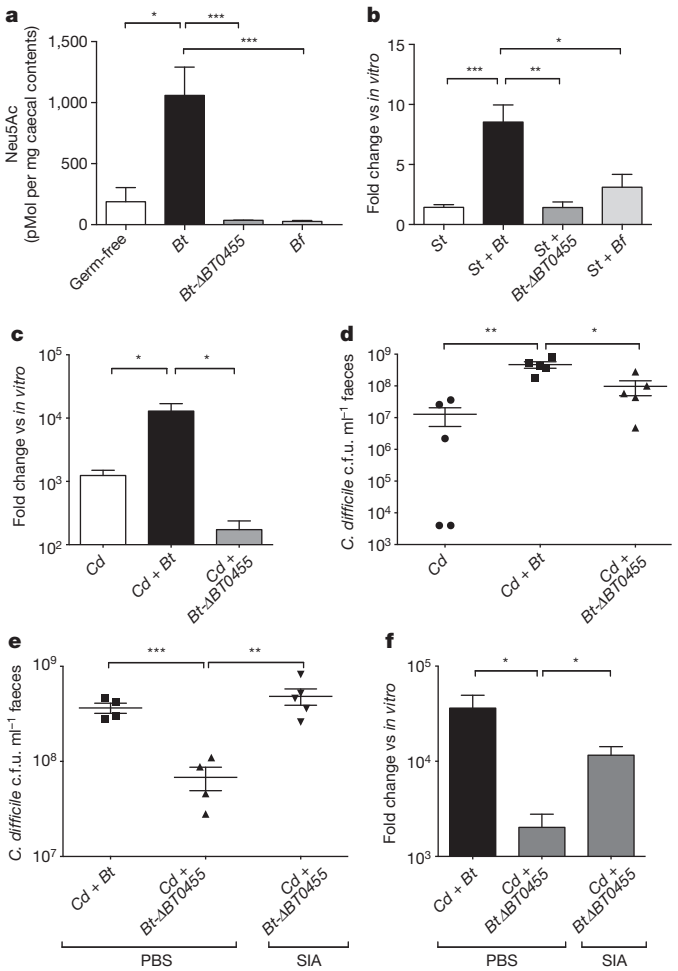


Figure 2 | *B. thetaiotaomicron* liberated sialic acid promotes emergence of *S. typhimurium* and *C. difficile*. **a**, Levels of free sialic acid in caecal contents in germ-free and gnotobiotic mice monoassociated for 10 days ($n = 3$, 3, 5 and 5, respectively). **b**, Fold change of expression of *S. typhimurium nanE* in caecal contents 1 day post infection relative to growth *in vitro* ($n = 9$, 4, 5 and 5, respectively). **c**, Induction of *C. difficile nanE* expression in caecal contents 3 days post infection relative to growth in minimal medium containing 0.5% glucose ($n = 4$ per group). **d**, *C. difficile* density in faeces 1 day post infection ($n = 5$ per group). **e**, *C. difficile* density 1 day post infection in faeces of PBS buffer or exogenous free sialic acid (SIA) treated mice. ($n = 4$–5 per group). **f**, Induction of *C. difficile nanE* gene expression 1 day post infection in faeces of PBS buffer or exogenous free sialic acid (SIA) treated mice relative to growth in minimal medium containing 0.5% glucose ($n = 5$ per group). Error bars indicate s.e.m. $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$.

C. difficile possesses a sialic acid catabolic operon, like *S. typhimurium*, but encodes no apparent genes for fucose consumption (Supplementary Fig. 6). To identify whether *C. difficile* also expresses sialic acid catabolism genes during its expansion within the gut, we quantified the expression of two genes within the *nan* operon, *nanE* and *nanA,* by quantitative PCR with reverse transcription (qRT–PCR) of RNA extracted from gnotobiotic mouse caecal contents. *C. difficile nanE* and *nanA* displayed elevated expression in *Bt*-monoassociated mice relative to expression levels observed when *C. difficile* colonized germ-free mice alone (*nanE*, 15-fold higher expression, $P = 0.02$; *nanA* 11-fold higher expression, $P = 0.039$; Fig. 1e). The presence of *Bt* in the gut of gnotobiotic mice resulted in an increased density of *C. difficile* one day post infection compared to infection of germ-free mice (1.5×10^8 versus 7.9×10^8 colony-forming units (c.f.u.) ml^{-1}; $P = 0.0009$; Fig. 1f).

Many commensal and pathogenic bacteria can utilize sialic acids from their hosts as a source of energy, carbon and nitrogen[20]. However, some bacteria, such as *B. thetatiotaomicron* encode the sialidase required to cleave and release this terminal sugar from the mucosal glycoconjugates, but lack the catabolic pathway (that is, the *nan* operon) required to consume the liberated monosaccharide. Presumably, the release of sialic acids allows *B. thetatiotaomicron* to access highly coveted underlying carbohydrates in the mucus[10,21]. Conversely, *S. typhimurium* and *C. difficile* encode the *nan* operon but each lacks the sialidase required for sialic acid liberation[22,23].

We quantified levels of free sialic acids in the caeca of *Bt*-monoassociated and germ-free mice. *Bt*-monoassociated mice exhibited a significantly higher concentration of the common sialic acid *N*-acetylneuraminic acid (Neu5Ac) versus germ-free mice, consistent with the ability of *Bt* to liberate but not consume the monosaccharide (1059 pmoles mg^{-1}, *Bt*-associated; 188 pmoles mg^{-1}, germ-free; $P = 0.029$; Fig. 2a). Colonization of mice with *Bt*-Δ*BT0455* (a mutant strain of *Bt* lacking a predicted cell surface sialidase that achieves the same density as wild type *in vivo*; Supplementary Fig. 7) did not result in increased free sialic acid,

nor did colonization with *Bacteroides fragilis* (*Bf*), which encodes both a sialidase and the *nan* operon and is therefore able to catabolize Neu5Ac (Fig. 2a). Expression of *S. typhimurium*'s *nan* operon was reduced upon infection of gnotobiotic mice colonized with *Bt*-Δ*BT0455* or *B. fragilis*, consistent with *S. typhimurium*'s dependence upon elevated levels of microbiota liberated sialic acid (Fig. 2b).

Loss of *Bt*-liberated sialic acid affects *C. difficile* in a manner similar to that observed with *S. typhimurium*. The *nan* gene expression in *C. difficile* was lower in mice colonized with the sialidase-deficient mutant *Bt*-Δ*BT0455* relative to expression in the presence of *Bt*-colonized mice (*nanE*, 75-fold higher expression, $P = 0.0187$; Fig. 2c). Furthermore, *C. difficile* density decreased in infected mice colonized with *Bt*-Δ*BT0455* mutant relative to densities in mice colonized with wild-type *Bt* (9.7×10^7 versus 4.6×10^8 c.f.u. ml^{-1}; $P = 0.0143$), illustrating the importance of *Bt*-liberated sialic acid in *C. difficile* expansion *in vivo* (Fig. 2d; Supplementary Fig. 8). Free sialic acid was orally administered to *Bt*-Δ*BT0455* and *C. difficile* co-colonized mice to determine if exogenous administration of the monosaccharide could reverse the decrease in *C. difficile* density by complementing the sialidase deficiency in this model. *C. difficile* densities increased 1 day post infection in *Bt*-Δ*BT0455* monoassociated mice fed free sialic acid compared to unsupplemented controls (4.8×10^8 versus 6.8×10^7 c.f.u. ml^{-1}; $P = 0.0066$) reaching densities similar to those observed in the presence of wild type (Fig. 2e). Furthermore, expression of *C. difficile nanE* increases in the sialic acid fed *Bt*-Δ*BT0455*-associated mice, further demonstrating that sialic acid use by *C. difficile* occurs concomitant with its increased densities *in vivo* (*nanE*, 58-fold higher expression over PBS buffer treated controls, $P = 0.019$; Fig. 2f). Notably, free sialic acid administration to germ-free mice infected with *C. difficile* resulted in higher densities of the pathogen in the lumen of the gut, confirming the important role of this monosaccharide *in vivo* (Supplementary Fig. 9). These data demonstrate that sialic acid catabolism by *C. difficile* promotes higher densities of the pathogen and depends upon the availability of the liberated monosaccharide within the lumen of the gut.

To determine whether sialic acid use is relevant to pathogen proliferation in an antibiotic-treated complex microbiota, we quantified free sialic acids in the caeca of conventional mice before and after antibiotic treatment. Levels of free Neu5Ac were very low within untreated conventional mice, consistent with efficient partitioning of Neu5Ac between members of an undisturbed complex microbiota (Fig. 3a). However, antibiotic-treated mice exhibited elevated levels of free sialic acid 1 day after treatment (725 pmoles mg^{-1}, 1 day post streptomycin, compared to 17 pmoles mg^{-1} in untreated mice; $P = 0.0019$), a time frame that coincides with pathogen expansion and acute microbiota disturbance (Supplementary Fig. 10)[24]. The pool of free sialic acids decreased by day 3 post treatment, consistent with recovery of the microbiota after antibiotic treatment[24] (Fig. 3a). *St*-Δ*nanA* and *St*-Δ*nanA*Δ*fucI* mutants both showed a competitive defect relative to wild-type *S. typhimurium* 1 day after infection in antibiotic-treated conventional mice (*St*-Δ*nanA*, CI = 1.83 $P = 0.0095$; *St*-Δ*nanA*Δ*fucI*, CI = 2.77, $P = 0.036$), consistent with sialic acid and fucose utilization providing an advantage to *S. typhimurium* during emergence (Fig. 3b). The lack of statistical significance of the phenotype in the *fucI* single mutant suggests mechanisms that are compensatory for fucose catabolism in this experimental model (Supplementary Fig. 11). To test whether *C. difficile* relies upon sialic acid catabolism in post-antibiotic expansion, we quantified the expression of the *nan* operon in antibiotic-treated conventional mice 1 day post infection. Coincident with expansion of *C. difficile*, the *nan* operon was highly induced compared to basal expression *in vitro* (*nanA*, 230-fold, $P = 0.0358$; *nanT*, 112-fold, $P = 0.0217$) confirming that *C. difficile* expresses this operon at high levels during its post-antibiotic expansion within a complex microbiota (Fig. 3c). As a test of sialic acid catabolism importance in *C. difficile* proliferation, we constructed a *nanT* mutant strain of *C. difficile* (*Cd-nanT*$^-$) that is deficient in sialic acid consumption (Supplementary Fig. 4). *Cd-nanT*$^-$ was significantly compromised in post-antibiotic

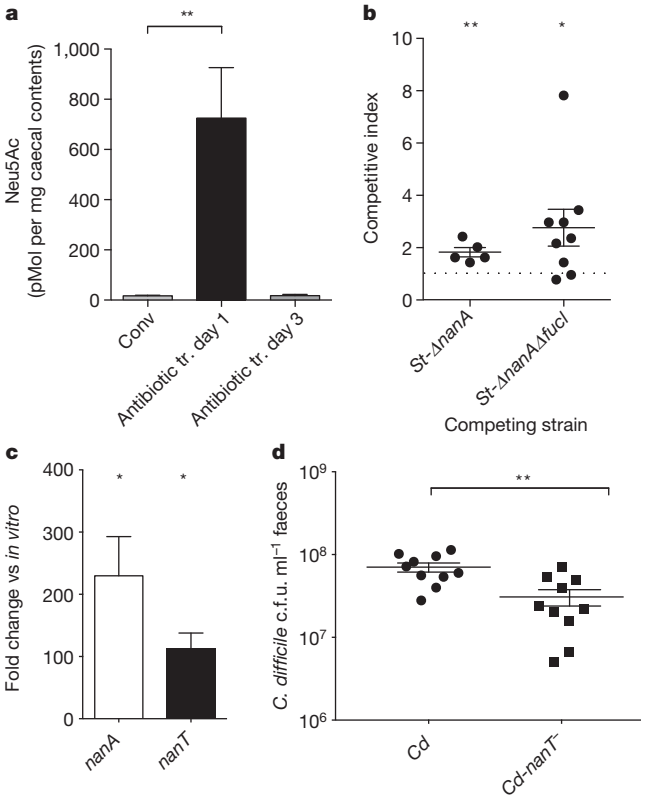


Figure 3 | *S. typhimurium* and *C. difficile* use mucin-derived monosaccharides resulting from antibiotic treatment of conventional mice. **a**, Levels of free sialic acid in caecal contents of conventional mice (Conv), antibiotic-treated mice 1 day and 3 days post treatment ($n = 8$, 9 and 3, respectively). **b**, Competitive index of wt *S. typhimurium* versus *S. typhimurium* mutants in caecal contents (*St-ΔnanA*) or faeces (*St-ΔnanAΔfucI*) of antibiotic-treated conventional mice. Horizontal bars indicate the geometric means of CI values and individual CI values are represented with dots ($n = 5$ and 9, respectively). **c**, Induction of *C. difficile nanA* and *nanT* expression in faecal samples 1 day post infection of antibiotic-treated conventional mice relative to growth in minimal medium containing 0.5% glucose ($n = 4$ per group). **d**, Density of wild-type *C. difficile* or a mutant deficient in sialic acid consumption (*Cd-nanT*$^-$) 3 days post-infection in faeces of antibiotic-treated conventional mice. ($n = 10$ per group). Error bars indicate s.e.m. *$P < 0.05$ and **$P < 0.01$.

expansion of conventional mice relative to wild-type *C. difficile* (3.1×10^7 versus 7.0×10^7 c.f.u. ml^{-1}; $P = 0.0023$) demonstrating the importance of sialic acid catabolism to *C. difficile* in attaining high densities in the context of an antibiotic-disrupted complex microbiota (Fig. 3d).

Recent studies have illustrated that enteric bacterial pathogens can subvert aspects of host inflammation to hold potential competitors within the microbiota at bay and enable pathogen proliferation[7,25–27]. Our results indicate that the antibiotic-associated pathogens *S. typhimurium* and *C. difficile* exploit increases in mucosal carbohydrate availability that occur upon disruption of the competitive ecosystem in which nutrients are typically efficiently consumed by endogenous community members. The transient post-antibiotic increase in monosaccharides liberated by the resident microbiota from host mucus provides a window of opportunity for these pathogens to expand to densities sufficient to induce self-promoting host inflammation (Supplementary Fig. 1). Implicit in these findings are new potential therapeutic strategies to combat post-antibiotic pathogen expansion.

METHODS SUMMARY

Bacterial strains and culture conditions. Strains were grown as follows: *B. thetaiotaomicron* (VPI-5482), TYG; *S. typhimurium* (SL1344), LB; *C. difficile* (630), RCM. For strains and primers see Supplementary Table 2.

Mice. Germ-free Swiss-Webster mice were maintained in gnotobiotic isolators, in accordance with A-PLAC, the Stanford IACUC. Conventional Swiss-Webster mice (SWRF, Taconic) were used for antibiotic-treated experiments.

Expression analyses. *S. typhimurium* transcriptomics in caecal contents were conducted using custom-made GeneChips. Robust multi-chip average-multi-species (RMA-MS) normalized signals were analysed for significant differences using significance analysis of microarrays (SAM). For primers see Supplementary Table 3.

Quantification of sialic acids. Free sialic acids were isolated from caecal content supernatants using a 1K MWCO filter and subjected to 1,2-diamino-4, 5-methylenedioxybenzene (DMB) derivatization before HPLC analysis.

Statistical analyses. The Student's *t*-test was used for statistical calculations. $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$. Error bars indicate s.e.m. n indicates the number of mice used per condition.

Full Methods and any associated references are available in the online version of the paper.

Received 17 May 2012; accepted 24 July 2013.
Published online 1 September 2013.

1. Doorduyn, Y., Van Den Brandhof, W. E., Van Duynhoven, Y. T., Wannet, W. J. & Van Pelt, W. Risk factors for *Salmonella* Enteritidis and Typhimurium (DT104 and non-DT104) infections in The Netherlands: predominant roles for raw eggs in Enteritidis and sandboxes in Typhimurium infections. *Epidemiol. Infect.* **134,** 617–626 (2006).
2. Pavia, A. T. *et al.* Epidemiologic evidence that prior antimicrobial exposure decreases resistance to infection by antimicrobial-sensitive *Salmonella*. *J. Infect. Dis.* **161,** 255–260 (1990).
3. Pépin, J. *et al.* Emergence of fluoroquinolones as the predominant risk factor for *Clostridium difficile*-associated diarrhea: a cohort study during an epidemic in Quebec. *Clin. Infect. Dis.* **41,** 1254–1260 (2005).
4. Kelly, C. P., Pothoulakis, C. & LaMont, J. T. *Clostridium difficile* colitis. *N. Engl. J. Med.* **330,** 257–262 (1994).
5. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–1920 (2005).
6. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
7. Stecher, B. *et al.* Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathog.* **6,** e1000711 (2010).
8. Chang, D. E. *et al.* Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proc. Natl Acad. Sci. USA* **101,** 7427–7432 (2004).
9. Sonnenburg, J. L. *et al.* Glycan foraging *in vivo* by an intestine-adapted bacterial symbiont. *Science* **307,** 1955–1959 (2005).
10. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4,** 447–457 (2008).
11. Fabich, A. J. *et al.* Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. *Infect. Immun.* **76,** 1143–1152 (2008).
12. Kamada, N. *et al.* Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. *Science* **336,** 1325–1329 (2012).
13. Pacheco, A. R. *et al.* Fucose sensing regulates bacterial intestinal colonization. *Nature* **492,** 113–117 (2012).
14. Maltby, R., Leatham-Jensen, M. P., Gibson, T., Cohen, P. S. & Conway, T. Nutritional basis for colonization resistance by human commensal *Escherichia coli* strains HS and Nissle 1917 against *E. coli* O157:H7 in the mouse intestine. *PLoS ONE* **8,** e53957 (2013).
15. Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* **6,** e280 (2008).
16. Hapfelmeier, S. & Hardt, W. D. A mouse model for *S. typhimurium*-induced enterocolitis. *Trends Microbiol.* **13,** 497–503 (2005).
17. Lawley, T. D. *et al.* Host transmission of *Salmonella enterica* serovar Typhimurium is controlled by virulence factors and indigenous intestinal microbiota. *Infect. Immun.* **76,** 403–416 (2008).
18. Lawley, T. D. *et al.* Antibiotic treatment of *Clostridium difficile* carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. *Infect. Immun.* **77,** 3661–3669 (2009).
19. Chen, X. *et al.* A mouse model of *Clostridium difficile*-associated disease. *Gastroenterology* **135,** 1984–1992 (2008).
20. Vimr, E. R., Kalivoda, K. A., Deszo, E. L. & Steenbergen, S. M. Diversity of microbial sialic acid metabolism. *Microbiol. Mol. Biol. Rev.* **68,** 132–153 (2004).
21. Marcobal, A. *et al.* Consumption of human milk oligosaccharides by gut-related microbes. *J. Agric. Food Chem.* **58,** 5334–5340 (2010).
22. Hoyer, L. L., Hamilton, A. C., Steenbergen, S. M. & Vimr, E. R. Cloning, sequencing and distribution of the *Salmonella typhimurium* LT2 sialidase gene, *nanH*, provides evidence for interspecies gene transfer. *Mol. Microbiol.* **6,** 873–884 (1992).
23. Sebaihia, M. *et al.* The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nature Genet.* **38,** 779–786 (2006).
24. Stecher, B. *et al.* *Salmonella enterica* serovar typhimurium exploits inflammation to compete with the intestinal microbiota. *PLoS Biol.* **5,** e244 (2007).
25. Winter, S. E. *et al.* Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* **467,** 426–429 (2010).
26. Lupp, C. *et al.* Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host Microbe* **2,** 119–129 (2007).
27. Barman, M. *et al.* Enteric salmonellosis disrupts the microbial ecology of the murine gastrointestinal tract. *Infect. Immun.* **76,** 907–915 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Sonnenburg for comments on the manuscript; M. St. Onge for technical assistance; and A. Shen, N. Minton and R. Knight for valuable help and reagents. This research was supported by R01-DK085025 (to J.L.S.), NSF graduate fellowships (to K.M.N. and J.A.F) and a Stanford Graduate Fellowship (to K.M.N.). J.L.S. holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund.

Author Contributions K.M.N., J.A.F., D.M.M. and J.L.S. designed experiments. K.M.N., J.A.F., S.K.H., J.B.L., P.C.K., N.N., B.C. and J.L.S. performed experiments. K.M.N., J.A.F., P.C.K., S.G., N.N., D.M.M. and J.L.S. analysed data. D.M.M. and B.C.W. contributed reagents. K.M.N., J.A.F. and J.L.S. wrote the paper.

Author Information Microbiota enumeration (16S rRNA) datasets have been deposited in the EMBL European Nucleotide Archive (ENA) under accession number ERP003629 and can also be found in the QIIME Database under the study ID 1958 (http://www.microbio.me/qiime/). Gene Chip datasets are available in the GEO database under accession number GSE49076. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L.S. (jsonnenburg@stanford.edu).

METHODS

Bacterial strains and culture conditions. *B. thetaiotaomicron* (ATCC 29148, also known as VPI-5482), was grown anaerobically (6% H_2, 20% CO_2, 74% N_2) overnight in TYG medium (1% tryptone, 0.5% yeast extract, 0.2% glucose, w/v) supplemented with 100 mM potassium phosphate buffer, pH 7.2, 4.1 mM cysteine, 200 μM histidine, 6.8 μM $CaCl_2$, 140 nM $FeSO_4$, 81 μM $MgSO_4$, 4.8 mM $NaHCO_3$, 1.4 mM NaCl, 1.9 μM haematin, plus 5.8 μM vitamin K_3.

All strains of *S. typhimurium* were derived from wild-type strain SL1344, which is naturally streptomycin-resistant. Using the methods of Datsenko and Wanner[28], mutant strains were first constructed in strain LT2, verified by PCR and then transduced into SL1344 using P22 phage transduction. Mutant strains and primers used in their generation are listed in Supplementary Table 2. Growth defects were not observed on glucose for either mutant and the presence of sialic acid did not pose a toxicity issue with the *nanA* mutant as has been previously reported for *E. coli*[29], consistent with its polarity that compromises *nanT* expression (Supplementary Fig. 5e). For colonization experiments, *S. typhimurium* strains were grown in Luria-Bertani (LB) broth at 37 °C with aeration or on LB agar plates, with the appropriate antibiotics (200 μg ml^{-1} streptomycin, 30 μg ml^{-1} kanamycin). Minimal medium used for transcriptional profiling consisted of 100 mM KH_2PO_4, pH 7.2, 15 mM NaCl, 8.5 mM $(NH_4)_2SO_4$, 4 mM L-cysteine, 1.9 mM haematin plus 200 mM L-histidine, 100 mM $MgCl_2$, 1.4 mM $FeSO_4$, 50 mM $CaCl_2$, 1 mg ml^{-1} vitamin K_3, 5 ng ml^{-1} vitamin B_{12} and 0.5% glucose (w/v). For evaluation of growth on various monosaccharides, strains were grown in M9 minimal media supplemented with 0.02% w/v histidine. Faecal densities (c.f.u.) of *S. typhimurium* were quantified by duplicate sampling with 1 μl loops, and subsequent dilution and spot plating on plain LB agar for gnotobiotic experiments and LB agar with streptomycin for conventional experiments.

C. difficile strain 630 was used in all *C. difficile* experiments and was cultured in reinforced clostridial medium (RCM) plus cysteine (Becton Dickinson) anaerobically (6% H_2, 20% CO_2, 74% N_2). *C. difficile* growth curves were generated using minimal medium composed of ammonium sulphate, sodium carbonate, calcium chloride, magnesium chloride, manganese chloride, cobalt chloride, histidine hematin, vitamin B_{12}, vitamin K_1, $FeSO_4$ and 1% Bacto Tryptone diluted 1:1 with 1% or 0.5% carbon source. $D_{600\,nm}$ (OD_{600}) was monitored using a BioTek PowerWave 340 plate reader (BioTek, Winooski, VT) every 30 min, at 37°C anaerobically (6% H_2, 20% CO_2, 74% N_2). Faecal densities (c.f.u.) of *C. difficile* were quantified by duplicate sampling with 1 μl loops and subsequent dilution and spot plating on brain heart infusion agar (Becton Dickinson) with 10% v/v of defibrinated horse blood (Lampire Biological Laboratories) supplemented with 25 mg l^{-1} erythromycin. For quantification of *C. difficile* c.f.u. in conventional mice, 1 μl of faeces was serially diluted in PBS and plated onto CDMN plates, composed of *C. difficile* agar base (Oxoid) with 7% v/v of defibrinated horse blood (Lampire Biological Laboratories), supplemented with 32 mg l^{-1} moxalactam (Santa Cruz Biotechnology) and 12 mg l^{-1} norfloxacin (Sigma-Aldrich). Plates were incubated overnight at 37 °C in an anaerobic chamber (Coy). Colonies identified as *C. difficile* were validated by colony PCR.

To construct the *nanT* null mutant (*Cd-nanT*$^-$), the ClosTron method for targeted gene disruption in *C. difficile* and detailed protocol were used[30,31]. SOEing PCRs with primers IBS, EBS1d, EBS2 and EBS (see Supplementary Table 2) were used to assemble and amplify the product for intron targeting, as outlined in the TargeTron users' manual (Sigma-Aldrich). The retargeting sequence was digested with BsrGI/HindIII and cloned into pMTL007C-E2. The resulting plasmid was transformed into HB101/pRK24 for conjugation into JIR8094[32] (a generous gift from A. Shen) to generate *Cd-nanT*$^-$.

Reagents and mice. Germ-free Swiss-Webster mice were maintained in gnotobiotic isolators and fed an autoclaved standard diet (Purina LabDiet 5K67) or a polysaccharide-deficient diet[33], in accordance with A-PLAC, the Stanford IACUC. All animals were 6–12-weeks of age and both genders were used. For all experiments involving *C. difficile* colonization of germ-free mice, the diet was switched to polysaccharide-deficient chow one day before inoculation with *C. difficile*. Conventional Swiss-Webster mice (RFSW, Taconic) were used for *S. typhimurium* and *C. difficile* antibiotic-treated experiments. The number of animals per group was chosen as the minimum likely required for conclusions of biological significance, established from prior experience. Randomization was not possible in the gnotobiotic setting and blinding was not applicable.

Conventional mice were orally gavaged with 20 mg streptomycin dissolved in water 24 h before infection and starved 18 h before infection. Mice were infected via oral gavage of 14 h overnight cultures of *S. typhimurium* resuspended in PBS. For single infections of gnotobiotic mice, 10^8 c.f.u. of *S. typhimurium* were gavaged. For *S. typhimurium* competitive index experiments, pure cultures of wild-type and mutant bacteria were diluted to equal densities, mixed in a 1:1 ratio and serial diluted in PBS to a total of 10^3 c.f.u. per 200 μl. Each mouse was orally gavaged with 200 μl of this dilution. Throughout the experiment, faecal samples were taken and dilutions were plated on LB agar plates containing streptomycin, which allows for growth of both the wild-type and mutant strains. Colonies from these plates were then patched onto LB agar plus kanamycin plates to determine the proportion of kanamycin-resistant (mutant) cells. With each sample, the ratio of kanamycin-sensitive (wild-type) bacteria to kanamycin-resistant (mutant) bacteria was divided by the kan^s/kan^r ratio determined from the original inoculum to produce the competitive index. Significance was evaluated using one-sample *t*-tests with a theoretical mean of 1. All competitive indices were determined for faecal samples with the exception of *St-ΔnanA,* which was surveyed in caecal contents.

For *C. difficile* experiments involving conventional mice, antibiotics were administered in the water for 3 days, starting 6 days before inoculation including: kanamycin (0.4 mg ml^{-1}), gentamycin (0.035 mg ml^{-1}), colistin (850 U ml^{-1}), metronidazole (0.215 mg ml^{-1}) and vancomycin (0.045 mg ml^{-1})[19]. Mice were then switched to regular water for 2 days and administered 1 mg of clindamycin by oral gavage 1 day before inoculation with *C. difficile*. Inoculations were by oral gavage at a density of 10^8 c.f.u. from overnight cultures.

For sialic acid administration experiments, *N*-acetylneuraminic acid (Calbiochem or Santa Cruz Biotechnology) was administered in the water at a 1% concentration. Additionally, mice were orally gavaged 1 mg of sialic acid twice a day. The amount of sialic acid in the caecal contents were calculated to equal approximately 700 pmoles mg^{-1} of caecal contents, which mirrors the average concentration of free sialic acids we quantified post antibiotic treatment (725 pmoles mg^{-1}).

Expression analysis. Genome-wide transcriptional profiling of *S. typhimurium* was conducted using custom-made Gene Chips, which contain probes for all annotated coding sequences for *S. typhimurium* LT2. RNA was purified from caecal contents and *in vitro* culture and complementary DNA (cDNA) was prepared, fragmented and labelled as described[9].

Gene Chip data were RMA-MS normalized as described[34] and log2 transformed. Statistical significance for differential gene expression was determined using significance analysis of microarrays (SAM)[35]. The delta parameter was adjusted to achieve a false discovery rate (FDR) nearest to 10% and this delta value was used to select significantly-regulated genes.

qRT–PCR analysis was performed on RNA extracted from caecal or faecal contents by phenol–chloroform extraction and bead beating. Superscript II (Invitrogen) was used to convert RNA to cDNA and SYBR Green (ABgene) in a MX3000P thermocycler (Stratagene) was used. Fold changes were normalized to *in vitro* growths in minimal medium containing 0.5% glucose for *C. difficile* and LB for *S. typhimurium*.

Quantification of sialic acids. All steps were carried out at 4 °C to minimize enzymatic hydrolysis. Approximately 200 mg of flash-frozen caecal contents were weighed out and resuspended in 400 μl of dH_2O. Samples were vortexed for 30 min at maximum speed and centrifuged for 15 min at 14,000*g* in a tabletop centrifuge. The supernatant was stored and the pellet resuspended in an additional 400 μl of dH_2O. The tubes were vortexed individually until the pellet was dispersed and then all samples were vortexed for 30 min, centrifuged, and supernatants were pooled. This process was repeated once more for a total volume of approximately 1 ml. Then 700 μl of each sample was filtered through a Pall 1K MWCO filter for 9 h at 7,000*g*. Samples were derivatized with 1,2-diamino-4,5-methylenedioxybenzene (DMB) as described previously[36]. The resulting product was analysed by reverse-phase HPLC using a C18 column (Dionex) at a flow rate of 0.9 ml min^{-1}, using a gradient of 5% to 11% acetonitrile in 7% methanol. The excitation and emission were 373 and 448 nm, respectively. The DMB-derivatized sialic acids were identified and quantified by comparing elution times and peak areas to known standards.

16S rRNA microbial community composition analysis. Faecal DNA was isolated and amplicons generated of the 16S rRNA V4 region (515F, 806R). Samples were sequenced at the Medical Genome Facility, Mayo Clinic using the MiSeq (Illumina) platform[37]. Data analysis was done using QIIME[38]. Single-end reads were analysed to determine operational taxonomic units (OTUs) at 97% sequence similarity using uclust (http://www.drive5.com/usearch/index.html). Taxonomy was assigned using RDP classifier against the GreenGenes database and a phylogenetic tree was built using FastTree. The OTU table was rarified to a sequencing depth of 900 for each set of samples. Beta diversity was determined using unweighted and weighted UniFrac[39].

Statistical analyses. The Student's *t*-test was used for statistical calculations; $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$. Error bars indicate s.e.m. *n* indicates the number of mice used per condition. Normal distribution was assumed for all data and no deviations were noted. Grubbs' test was used to identify and eliminate statistical outliers.

28. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97,** 6640–6645 (2000).

29. Vimr, E. R. & Troy, F. A. Identification of an inducible catabolic system for sialic acids (*nan*) in *Escherichia coli*. *J. Bacteriol.* **164,** 845–853 (1985).
30. Heap, J. T. *et al.* The ClosTron: mutagenesis in *Clostridium* refined and streamlined. *J. Microbiol. Methods* **80,** 49–55 (2010).
31. Adams, C. M. *et al.* Structural and functional studies of the CspB protease required for *Clostridium* spore germination. *PLoS Pathog.* **9,** e1003165 (2013).
32. O'Connor, J. R. *et al.* Construction and analysis of chromosomal *Clostridium difficile* mutants. *Mol. Microbiol.* **61,** 1335–1351 (2006).
33. Sonnenburg, E. D. *et al.* Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141,** 1241–1252 (2010).
34. Stevens, J. R. *et al. Statistical issues in the normalization of multi-species microarray data.* Appl. Stat. Agric. Proc. Conf. Appl. Stat. Agric. Kansas State Univ. 47–62 (National Agricultural Library, 2008).
35. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98,** 5116–5121 (2001).
36. Manzi, A. E., Diaz, S. & Varki, A. High-pressure liquid chromatography of sialic acids on a pellicular resin anion-exchange column with pulsed amperometric detection: a comparison with six other systems. *Anal. Biochem.* **188,** 20–32 (1990).
37. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
38. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
39. Lozupone, C., Hamady, M. & Knight, R. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).

doi:10.1038/nature12519

Immune clearance of highly pathogenic SIV infection

Scott G. Hansen[1]*, Michael Piatak Jr[2]*, Abigail B. Ventura[1], Colette M. Hughes[1], Roxanne M. Gilbride[1], Julia C. Ford[1], Kelli Oswald[2], Rebecca Shoemaker[2], Yuan Li[2], Matthew S. Lewis[1], Awbrey N. Gilliam[1], Guangwu Xu[1], Nathan Whizin[1], Benjamin J. Burwitz[1], Shannon L. Planer[1], John M. Turner[1], Alfred W. Legasse[1], Michael K. Axthelm[1], Jay A. Nelson[1], Klaus Früh[1], Jonah B. Sacha[1], Jacob D. Estes[2], Brandon F. Keele[2], Paul T. Edlefsen[3], Jeffrey D. Lifson[2] & Louis J. Picker[1]

Established infections with the human and simian immunodeficiency viruses (HIV and SIV, respectively) are thought to be permanent with even the most effective immune responses and antiretroviral therapies only able to control, but not clear, these infections[1–4]. Whether the residual virus that maintains these infections is vulnerable to clearance is a question of central importance to the future management of millions of HIV-infected individuals. We recently reported that approximately 50% of rhesus macaques (RM; *Macaca mulatta*) vaccinated with SIV protein-expressing rhesus cytomegalovirus (RhCMV/SIV) vectors manifest durable, aviraemic control of infection with the highly pathogenic strain SIVmac239 (ref. 5). Here we show that regardless of the route of challenge, RhCMV/SIV vector-elicited immune responses control SIVmac239 after demonstrable lymphatic and haematogenous viral dissemination, and that replication-competent SIV persists in several sites for weeks to months. Over time, however, protected RM lost signs of SIV infection, showing a consistent lack of measurable plasma- or tissue-associated virus using ultrasensitive assays, and a loss of T-cell reactivity to SIV determinants not in the vaccine. Extensive ultrasensitive quantitative PCR and quantitative PCR with reverse transcription analyses of tissues from RhCMV/SIV vector-protected RM necropsied 69–172 weeks after challenge did not detect SIV RNA or DNA sequences above background levels, and replication-competent SIV was not detected in these RM by extensive co-culture analysis of tissues or by adoptive transfer of 60 million haematolymphoid cells to naive RM. These data provide compelling evidence for progressive clearance of a pathogenic lentiviral infection, and suggest that some lentiviral reservoirs may be susceptible to the continuous effector memory T-cell-mediated immune surveillance elicited and maintained by cytomegalovirus vectors.

Clinical and experimental observations have suggested that HIV and SIV infections might be vulnerable to immune control or pharmacological clearance in the first few hours to days of infection, before both the viral amplification needed for efficient mutational escape and the establishment of the highly resilient viral reservoir that sustains the infection[4,6–8]. Cytomegalovirus (CMV) vectors were designed to exploit this putative window of vulnerability, based on their ability to elicit and indefinitely maintain high frequency, effector-differentiated, and broadly targeted virus-specific T cells in potential sites of early viral replication[5,9,10]. Indeed, the pattern of protection observed in approximately 50% of RhCMV/SIV vector-vaccinated RM after intrarectal SIVmac239 challenge was consistent with early immunological interception of the nascent SIV infection at the portal of viral entry and immune control before irreversible systemic spread[5]. Protected RM manifested a very transient viraemia at the onset of infection, followed by control of plasma SIV levels to below the threshold levels of quantification, except for occasional plasma viral 'blips' that waned over time, and after one year, demonstrated only trace levels of tissue-associated SIV RNA and DNA at necropsy using ultrasensitive assays. The occurrence of plasma viral blips and the recurrence of 'breakthrough' progressive SIV infection in 1 of the 13 RhCMV/SIV vector-protected RM at day 77 after infection indicated that SIV was not immediately cleared from these protected RM, but the failure to find more than trace levels of SIV nucleic acid in systemic lymphoid tissues was consistent with the productive infection being largely contained at the portal of entry with the possibility of eventual clearance. Given the crucial importance of understanding the degree to which a highly pathogenic lentivirus can be contained or even cleared by adaptive immunity, we sought to define more precisely the spread and dynamics of SIV infection in RM that controlled the infection as a consequence of RhCMV/SIV vector vaccination, and in particular, the extent to which residual SIV was eventually cleared from these animals.

To establish the extent of SIV spread early after the onset of RhCMV/SIV vector-mediated control, we studied a group of five RM vaccinated with RhCMV vectors containing SIV Gag, Rev/Tat/Nef, Env and Pol (but not Vif) inserts that were taken to necropsy within 24 days of controlling plasma viraemia after intrarectal inoculation with SIVmac239. All of these RM had measureable SIV RNA in plasma for one or two weekly time points after challenge, followed by at least three consecutive weekly samples with plasma SIV RNA below 30 copy equivalents (equiv.) per ml, and at the time of necropsy, below 5 copy equiv. ml^{-1}, as measured by an ultrasensitive assay (Fig. 1a). Infection was confirmed by the *de novo* development of T-cell responses against SIV Vif (not included in the vaccine) in all RM (Fig. 1b and Supplementary Fig. 1a). As previously described[5], protection occurred without anamnestic boosting of vaccine-elicited SIV-specific $CD8^+$ T-cell responses in blood (Fig. 1b), and at necropsy, robust $CD4^+$ and $CD8^+$ T-cell responses to the SIV proteins included in the RhCMV/SIV vaccine vectors were identified (Supplementary Fig. 1b). We then used ultrasensitive, nested quantitative PCR (qPCR) and quantitative PCR with reverse transcription (qRT–PCR) assays to quantify SIV DNA and RNA, respectively, in the tissues of these protected RM, in comparison with tissues from three unchallenged, RhCMV/SIV vector-vaccinated RM (SIV^- controls), two unvaccinated RM with productive SIV infection (one progressor and one elite controller) and three RM with SIV infection suppressed with antiretroviral drug treatment (ART) (Fig. 1c, Supplementary Figs 2–4 and Supplementary Table 1). Two of the five RhCMV/SIV vector-protected RM showed levels of SIV DNA and RNA approaching the very low level background signal observed for SIV^- control RM. However, the other three showed readily measurable SIV RNA, not only in rectal/colonic mucosa (portal of entry), but also in lymph nodes draining the portal of entry (iliosacral and mesenteric lymph node groups), as well as in sites of presumed haematogenous spread: bone marrow, spleen and liver. The level of SIV RNA in the tissues of these three RM was less than that seen in progressive infection, but comparable to that in the elite SIV controller and in ART-suppressed SIV infection. Notably, however, levels of tissue-associated SIV DNA in the RhCMV/SIV vector-protected RM were all substantially lower than in the RM with elite control and ART suppression, probably reflecting virological control before, rather than after, peak viral replication

[1]Vaccine and Gene Therapy Institute and Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon 97006, USA. [2]AIDS and Cancer Virus Program, SAIC Frederick, Inc., Frederick National Laboratory, Frederick, Maryland 21702, USA. [3]Statistical Center for HIV/AIDS Research and Prevention, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.
*These authors contributed equally to this work.

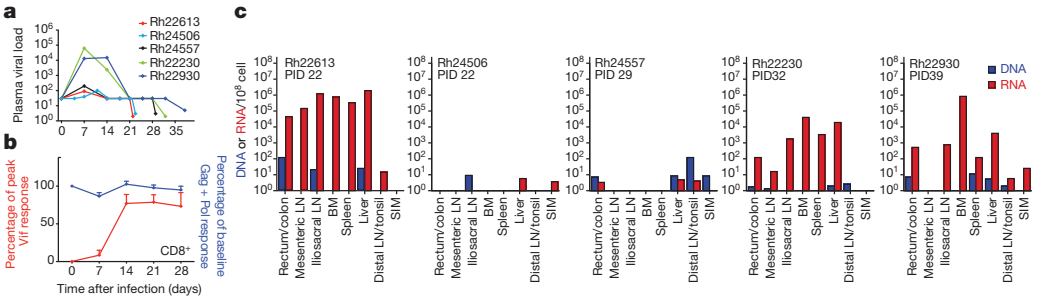


Figure 1 | Virological analysis of early RhCMV/SIV vector-mediated protection. a, Plasma viral load (measured as log(copy equiv. per ml)) profiles of five RhCMV/SIV vector-vaccinated RM with complete control of viraemia after intrarectal SIVmac239 challenge. All five RM controlled viraemia to below the 30 copy equiv. ml^{-1} limit of quantification for the standard plasma viral load assay used for all pre-necropsy samples, and to below the 1–5 copy equiv. ml^{-1} limit of detection for the ultrasensitive plasma viral load assay used on necropsy samples (individual detection limits for each terminal sample shown). **b**, Frequencies of peripheral blood memory $CD8^+$ T cells specific for SIV proteins that were (Gag plus Pol) or were not (Vif) included in the RhCMV/SIV vectors, shown before and after the onset of the controlled SIV infection. The response frequencies (mean ± s.e.m.) were normalized to the response frequencies immediately before SIV infection for the vaccine-elicited SIV Gag- and Pol-specific responses, and to the peak frequencies after SIV infection for the *de novo* SIV Vif-specific responses. **c**, Analysis of tissue-associated SIV DNA and RNA (copy equiv. per 10^8 cell equiv.) in the five RhCMV/SIV vector-protected RM at necropsy using ultrasensitive quantitative PCR and RT–PCR. BM, bone marrow; LN, lymph nodes; PID, post-infection day; SIM, small intestine mucosa.

in the RhCMV/SIV vector-protected RM, and the limited time for SIV DNA^+ cells to accumulate in these RM before necropsy. Although these data suggest a much smaller SIV reservoir in the RhCMV/SIV vector-protected RM than in the SIV^+ controls, including the RM with ART-suppressed SIV infection, we were able to recover replication-competent SIV from iliosacral lymph nodes and spleen in all five of the RhCMV/SIV-protected RM taken to early necropsy (and from bone marrow and mesenteric lymph nodes in three of these five RM), including the two RM with near background levels of SIV RNA by nested qRT–PCR (Table 1). This replication-competent SIV was found in tissues manifesting only minimal interferon-stimulated gene expression, significantly less than found in either progressive or ART-suppressed SIV infection (Supplementary Fig. 5). Taken together, these data demonstrate that in RhCMV/SIV vector-protected RM, SIV can escape the portal of entry and establish infection in draining lymph nodes, as well as bone marrow, spleen and liver, before stringent control.

After intrarectal inoculation, SIV infection has been reported to spread to draining lymph nodes within 4 h (ref. 11), a rate of dissemination that may preclude SIV-specific effector memory T cells from containing the infection within the mucosa. By contrast, the development of SIV infection after intravaginal inoculation has been reported to require local amplification, with distal spread only after 4–5 days[6]. To determine whether RhCMV/SIV vector-elicited T-cell responses might locally control and perhaps clear an intravaginal SIV challenge, we compared the outcome of repeated, limiting dose intravaginal SIVmac239 challenge in cycling female RM vaccinated twice (weeks 0 and 14) with RhCMV/SIV vectors (group A) versus similar RM vaccinated twice with RhCMV vectors encoding non-SIV inserts (group B) or left unvaccinated (group C), with challenge 78 weeks after initial vaccination (Supplementary Fig. 6). The immunogenicity of RhCMV/SIV vectors in these female RM was similar to that described for male RM with robust, effector memory-biased SIV-specific $CD4^+$ and $CD8^+$ T-cell responses to all SIV inserts (Supplementary Figs 7 and 8), but little to no SIV Env-specific antibody responses (Supplementary Fig. 9). As previously described for intrarectal SIV challenge of male RM[5], RhCMV/SIV vector vaccination did not significantly affect the number of intravaginal SIV challenges required to achieve infection relative to control-vaccinated and unvaccinated RM (Supplementary Fig. 10), but did markedly alter the course of SIV infection with 9 out of 16 RhCMV/SIV vector-vaccinated female RM manifesting stringent (MHC class I allele-independent) control of plasma viraemia compared with none of 18 infected female control RM (Fig. 2a and Supplementary Table 2).

Table 1 | Replication-competent SIV by inductive co-culture at necropsy

	Progressor	Elite controller	ART-suppressed	RhCMV/SIV vector-protected										
				Early term					Medium and long term					
Animal	Rh23657	Rh21582	Rh25708	Rh22613	Rh24506	Rh24557	Rh22230	Rh22930	Rh24514	Rh26467	Rh24552	Rh24272	Rh24399	Rh24250
Virus	SIVmac239	SIVmac239	SIVmac251	SIVmac239	SIVmac239	SIVmac239	SIVmac239	SIVmac239	SIVmac239	SIVmac251	SIVmac239	SIVmac239	SIVmac239	SIVmac239
Plasma SIV (copy equiv. ml^{-1})	110,000	510	11	<2	<3	<3	<2	<5	<1	<1	<1	<1	<1	<1
Duration of infection at necropsy (weeks)	74	28	63*	3	3	4	5	5	69	76	166	167	167	172
SIV^+ cultures/total cultures														
Distal LN	80/80	32/80	4/49	0/80	0/80	0/80	0/60	1/100	0/80	0/80	0/80	0/60	0/80	0/58
Iliosacral LN	40/40	9/40	0/40	2/40	1/40	2/40	3/40	1/40	0/40	0/40	0/40	0/40	0/40	0/40
Mesenteric LN	80/80	2/80	3/80	1/80	0/80	0/80	2/80	1/60	0/80	0/80	0/80	0/60	0/60	0/65
Spleen	40/40	1/40	5/40	2/40	1/40	2/40	1/40	1/40	0/40	0/40	0/40	0/40	0/40	0/40
Liver	ND	0/20	2/19	1/40	0/20	2/20	ND	0/20	0/20	0/20	0/20	0/20	0/20	0/20
Bone marrow	ND	0/20	0/20	ND	1/20	2/20	1/20	0/20	0/16	0/13	ND	0/20	0/20	0/20
Total positive	240/240	44/280	14/248	6/280	3/280	8/280	7/240	4/280	0/276	0/273	0/260	0/240	0/260	0/243
				28/1,360					0/1,549					

The frequency of SIV^+ cultures (250,000 cells per culture) is shown. The frequencies of SIV^+ cultures using cells derived from the tissues of early-term RhCMV/SIV vector-protected RM is significantly higher than the frequencies of SIV^+ cultures using cells derived from tissues of medium- and long-term protected RM ($P < 0.0001$, Fisher's exact test). LN, lymph nodes; ND, no data.
* 24 weeks after ART initiation.

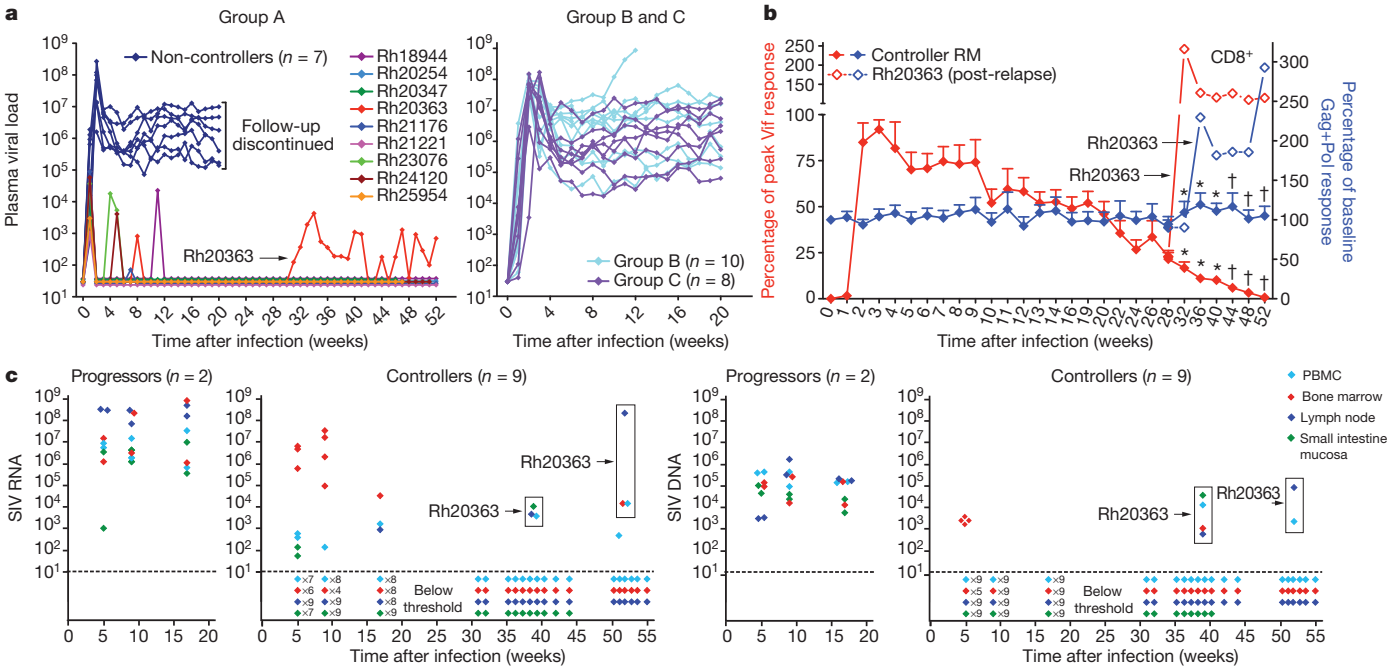


Figure 2 | Longitudinal analysis of RhCMV/SIV vector-mediated protection after intravaginal challenge. **a**, Plasma viral load (measured as log(copy equiv. per ml)) profiles of groups A (RhCMV/SIV vector-vaccinated), B (control RhCMV vector-vaccinated) and C (unvaccinated) RM after infection by repeated, limiting dose, intravaginal SIVmac239 challenge, with the day of infection defined as the challenge before the first above-threshold plasma viral load. The fraction of infected RM that met controller criteria (see Methods) in group A (9 out of 16) versus groups B and C (0 out of 18) was significantly different ($P = 0.0002$) by two-sided Fisher's exact test. Note that Rh20363 initially manifested aviraemic protection, but then relapsed with productive, albeit controlled, infection at week 31 after infection. **b**, Mean (and s.e.m.) frequencies of peripheral blood memory $CD8^+$ T cells specific for SIV proteins that were (Gag plus Pol) or were not (Vif) included in the RhCMV/SIV vectors, measured before and after the onset of SIV infection in the nine group A RM with initial aviraemic control (response frequencies normalized as described in Fig. 1b). Asterisks indicate $n = 8$ (minus Rh20363 post-relapse); daggers indicate $n = 7$ (minus Rh20363 and Rh20347, the latter used in the $CD8^+$ cell depletion study described in Supplementary Fig. 14). **c**, Quantification of tissue-associated SIV RNA (left) and DNA (right) (copies per 10^8 cell equiv.) in the designated longitudinal samples of the nine group A controllers versus two representative viraemic progressors. All sample types were analysed at weeks 5, 9 and 17 in all RM. All sample types were analysed a fourth time in all controller RM between post-infection weeks 30 and 40, and PBMCs, bone marrow and lymph node samples were analysed a fifth time in eight out of nine controller RM between post-infection weeks 42 and 55. Each symbol represents a single determination from the designated tissue, except when a multiplication factor is shown (for example, ×7 indicates a total of seven samples from different RM with below threshold measurements for that time point).

Five of these nine protected female RM manifested a second episode of transient plasma viraemia within the first 12 weeks after initial control, but overall, the fraction of protected female RM (followed for at least 30 weeks) with such plasma viral blips (56% versus 100%; $P = 0.02$ by Fisher's exact test) and the number of blips per RM (0.7 versus 6.0; $P < 0.0001$ by two-sided Wilcoxon rank sum test) were less than that observed in RhCMV/SIV vector-vaccinated male RM protected after intrarectal challenge[5]. Other characteristics of protection in these intravaginally challenged, RhCMV/SIV vector-vaccinated female RM were identical to those previously reported for RhCMV/SIV vector-mediated protection of male RM against intrarectal challenge[5], including development of *de novo* SIV Vif-specific $CD4^+$ and $CD8^+$ T-cell responses, lack of an anamnestic boost of the vaccine-elicited SIV-specific $CD4^+$ or $CD8^+$ T cells, lack of SIV Env seroconversion, and lack of $CD4^+$ T-cell depletion at mucosal effector sites (Fig. 2b and Supplementary Figs 9, 11 and 12).

To determine whether SIV infection spread from the cervical/vaginal mucosa in the nine RhCMV/SIV vector-protected female RM, we biopsied bone marrow, peripheral lymph nodes (axillary/inguinal) and small intestinal mucosa for nested qRT–PCR and qPCR analysis of SIV RNA and DNA, respectively, at 5, 9, 17 and >30 weeks after infection. Notably, in the first 9 weeks of infection, five of these nine RM manifested levels of SIV RNA in bone marrow comparable to levels seen in uncontrolled SIV infection, but, whereas in uncontrolled infection SIV RNA levels were similarly high in peripheral blood mononuclear cells (PBMCs), lymph nodes and intestinal mucosa, SIV RNA was either not detected or detected only at very low levels in these sites in the RhCMV/SIV vector-protected RM (Fig. 2c). Moreover, in contrast to uncontrolled infection, SIV DNA was inconsistently detected in the samples from the RhCMV/SIV vector-protected RM, and by 40 weeks after infection, all nine of the RhCMV/SIV vector-protected RM had at least one sample set in which both SIV RNA and DNA were below the level of detection in all sites. In eight of these RM (excluding Rh20363, see below), all samples obtained subsequent to 30 weeks after infection showed SIV RNA and DNA levels below the level of detection, with the exception of one PBMC sample with low-level SIV RNA (454 copy equiv. per 10^8 cells). The differences in the frequency of SIV detection in samples obtained at 5, 9 and 17 weeks versus >30 weeks after infection from these eight RM were highly significant ($P = 0.002$ for all samples, $P = 0.0006$ for bone marrow by two-sided Wilcoxon rank sum tests).

The ability to detect tissue-associated SIV early, but not late, after infection in these eight stably protected female RM, particularly in bone marrow, is consistent with initial spread and subsequent control and progressive clearance of SIV. In accordance with this, the frequencies of circulating SIV Vif-specific T cells, which are elicited and maintained by antigen derived from SIV infection (rather than the vaccine), progressively declined in these RM until these responses were no longer detectable (Fig. 2b and Supplementary Fig. 11). However, despite having no detectable SIV RNA or DNA in PBMCs and tissue samples at week 17, and declining SIV Vif-specific T-cell responses, one animal (Rh20363) showed the emergence of low-level productive SIV infection at week 31 after infection (Fig. 2a). The boosting of SIV-specific $CD4^+$ and $CD8^+$ T-cell responses (Fig. 2b and Supplementary Fig. 11), including *de novo* $CD8^+$ T-cell responses to canonical *Mamu-A*01*-restricted SIV epitopes (Supplementary Fig. 13), the appearance of cell-associated RNA and

DNA in subsequent PBMCs, lymph node and intestinal samples (Fig. 2c), and the induction of increased plasma and PBMC-associated SIV loads with experimental *in vivo* $CD8^+$ lymphocyte depletion (Supplementary Fig. 14) indicates that this RM spontaneously converted from a unique state of stringent viral containment with little or no continuing viral replication to a different state characterized by continuing, but low-level SIV replication (consistent with conventional 'elite' immunological control). In keeping with this, sequence analysis of the breakthrough virus 3 weeks after initial viral rebound showed little evolution from the initial SIVmac239 sequence except, notably, a putative escape mutation in the Tat-SL8 epitope sequence, consistent with early escape from the Tat-SL8-specific T-cell responses that developed after viral rebound at week 31 (Supplementary Figs 13 and 15). Given the enormous breadth of RhCMV/SIV vector-elicited $CD8^+$ T-cell responses[10], this limited sequence evolution suggests that the loss of aviraemic control in Rh20363 was more likely due to inadequate immune surveillance of residual infection than mutational escape. Experimental $CD8^+$ lymphocyte depletion was also performed on three RhCMV/SIV vector-protected female RM that retained aviraemic control, and in keeping with previous analysis of $CD8^+$ lymphocyte depletion of RhCMV/SIV vector-vaccinated male RM protected after intrarectal challenge[5,9], this treatment did not induce detectable plasma viraemia (Supplementary Fig. 14). However, one of these RM (Rh21176) transiently manifested unequivocal detection of SIV RNA (10 out of 10 replicates positive) and replication-competent SIV (7 out of 20 co-cultures positive) in lymph nodes at day 10 after $CD8^+$ lymphocyte depletion, demonstrating the presence of at least local, very low level residual SIV infection in this RM after 52 weeks of stringent control. In contrast to Rh20363, Rh21176 maintained aviraemic control, indicating that this RM's immune system either controlled or eliminated residual foci of SIV replication.

The finding that RhCMV/SIV vector-protected RM are able to control haematogenous SIV dissemination after both intrarectal and intravaginal challenge suggested that the immune responses elicited by these vectors might provide protection even when mucosal surfaces are bypassed. To assess this possibility, we challenged six RhCMV/SIV-vaccinated RM with low dose, intravenous SIVmac239, and found that two of these six RM manifested the same pattern of control observed after mucosal challenge—a transient, low-level viraemia associated with the development of an SIV Vif-specific T-cell response, and detection of SIV RNA in bone marrow (high level) and/or PBMCs (low level) early, but not late, after infection (Supplementary Fig. 16). Taken together, these data indicate that (1) RhCMV/SIV vector-elicited immune responses can mediate protection regardless of the route of SIV challenge, (2) viral control is both local and systemic, and (3) replication-competent SIV can persist in several sites for weeks to months in protected RM (even when aviraemic), but seems to decline over time.

To determine the ultimate fate of residual SIV in RhCMV/SIV vector-protected RM, we followed a total of ten protected RM for 69–180 weeks after infection (Fig. 3a, b). In all of these RM, plasma viral blips became increasingly infrequent over time, with no blips observed after 70 weeks. The frequency of the SIV infection-dependent, SIV Vif-specific $CD8^+$ T cells in blood also progressively declined in all RM until these responses were no longer detectable (Supplementary Fig. 17). In contrast to the SIV Vif-specific $CD8^+$ T-cell responses, the SIV-specific $CD8^+$ T-cell responses elicited by the RhCMV/SIV vectors remained stable, including high frequencies of $CD8^+$ T cells capable of recognizing autologous SIV-infected $CD4^+$ T cells (Supplementary Fig. 18). Analysis of six of these medium- to long-term protected RM at necropsy, including one RM that was $CD8^+$ lymphocyte-depleted 10 days before necropsy (Supplementary Fig. 19), confirmed the systemic loss of SIV Vif-specific T cells, and the maintenance of RhCMV vector-elicited, SIV-specific T cells (Supplementary Fig. 20). Most importantly, ultrasensitive, nested qRT–PCR and qPCR analysis of ≥54 tissues per animal (ten replicates per tissue, including extensive sampling of all tissues shown to contain SIV in the short-term RhCMV/SIV vector-protected RM) revealed extremely low to absent levels of SIV RNA and DNA that were indistinguishable from measurements in unchallenged RhCMV/SIV-vaccinated (SIV^-) controls (Fig. 3c, d, Supplementary Figs 2 and 21 and Supplementary Tables 1 and 3). Moreover, despite extensive sampling (>240 cultures per animal), no replication-competent SIV was isolated by co-culture analysis from the lymphoid tissues of these RM (Table 1). Finally, we asked whether the adoptive transfer of a total of 6×10^7

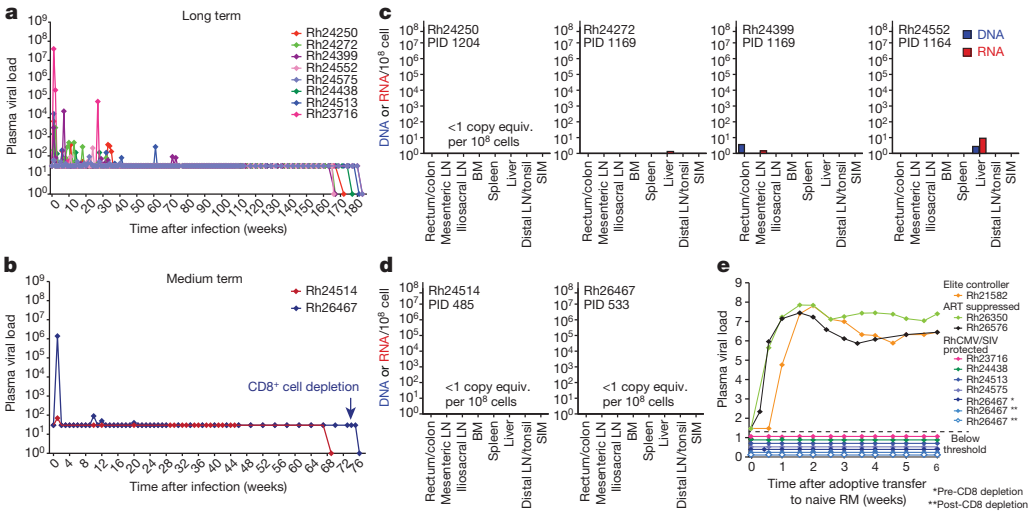


Figure 3 | Virological analysis of medium- to long-term RhCMV/SIV vector-mediated protection. **a**, **b**, Plasma viral load (measured as log(copy equiv. per ml)) profiles of ten RhCMV/SIV vector-vaccinated RM that controlled SIV infection after intrarectal challenge (eight long term (**a**) and two medium term (**b**)). The limit of detection for all pre-terminal plasma viral load assays is 30 copy equiv. ml^{-1}; the limit of detection for the ultrasensitive assay used on the terminal sample of the study was ≤1 copy equiv. ml^{-1}. Note that one of the RM with medium-term protection (Rh26467) was $CD8^+$ lymphocyte-depleted 10 days before the terminal sample. **c**, **d**, Quantification of tissue-associated SIV DNA and RNA in four long-term and two medium-term protected RhCMV/SIV-vaccinated RM studied at necropsy, including the $CD8^+$ cell-depleted RM (Rh26467). **e**, Assessment of residual replication-competent, cell-associated SIV in medium- and long-term protected RM by adoptive transfer of 6×10^7 haematolymphoid cells (3×10^7 blood leukocytes and 3×10^7 lymph node cells or, in one transfer from Rh26467, represented by the open symbol, 3×10^7 bone marrow leukocytes and 3×10^7 spleen cells) to SIV-naive RM with SIV infection in the recipient RM delineated by plasma viral load. Cell transfers from RM with conventional elite SIV control and ART-suppressed SIV infection resulted in rapid onset of SIV infection in the recipient RM, but no SIV infection was observed in RM receiving cells from medium- to long-term RhCMV/SIV vector-protected RM (including Rh26467, analysed both before and after $CD8^+$ cell depletion).

haematolymphoid cells (3×10^7 each of peripheral blood leukocytes and lymph node cells, or 3×10^7 each of bone marrow leukocytes and spleen cells) from three SIV^+ control RM (two with ART-suppressed infection and one elite controller), and five medium- or long-term RhCMV/SIV vector-protected RM (including one RM tested before and after $CD8^+$ cell depletion) would initiate infection in SIV-naive RM. Remarkably, although cells from the SIV^+ controls, including ART-suppressed RM, rapidly initiated SIV infection in the SIV-naive recipients (manifested by the onset of SIV replication and induction of SIV Vif-specific T-cell responses), no evidence of SIV infection was observed in the SIV-naive recipients receiving cells from the medium- and long-term RhCMV/SIV vector-protected RM (Fig. 3e and Supplementary Fig. 22). Taken together, these data provide strong evidence that after being unequivocally infected with SIV, these RhCMV/SIV vector-vaccinated RM cleared detectable infection, such that by all measured criteria (lack of plasma viral blips, absence of Vif-specific T-cell responses, extensive ultrasensitive qRT–PCR and qPCR and co-culture analysis, and adoptive transfer) these RM were indistinguishable from RhCMV/SIV vector-vaccinated controls that had never been exposed to SIV. Although we cannot rule out residual virus below our level of detectability, or in tissues not examined, these data strongly support progressive immune-mediated clearance of an established lentivirus infection, leading to a situation meeting criteria for a functional cure[12] and consistent with possible viral eradication.

In the past 5 years, the HIV/AIDS vaccine field has concluded that a prophylactic HIV/AIDS vaccine must prevent or eliminate HIV infection, as it is thought that any residual infection runs a high risk of eventual progression[13]. Our demonstration here that the virus-specific, effector memory T cells maintained by a persistent vector can shut down productive SIV infection, and by maintaining immune surveillance over time, functionally cure and possibly eradicate this infection, indicates that an effector memory T-cell-targeted HIV/AIDS vaccine could (by itself, or combined with antibody-targeted approaches) provide meaningful long-term efficacy. Our results also suggest that an effector memory T-cell-targeted vaccine might contribute to HIV cure strategies. Although the SIV reservoirs that initially develop in RhCMV/SIV vector-vaccinated controllers are smaller in size, and possibly different in character from HIV/SIV reservoirs in the setting of ART administration initiated in chronic infection, it is conceivable that the indefinitely persistent, unconventionally targeted[10], viral-specific T cells elicited and maintained by CMV vectors—alone or in combination with agents designed to activate HIV gene expression[1,2,12]—might exert potent immune pressure on cells with any HIV protein expression (including expression of viral antigen by stochastically activated, latently infected cells) and thereby facilitate depletion of residual HIV reservoirs in patients on suppressive ART. It is also possible that these responses might stringently control recrudescent 'rebound' infection after ART withdrawal in a manner analogous to their control of primary SIV infection in this study. In summary, the ability of CMV vectors to implement continuous, long-term, and potent antipathogen immune surveillance makes them promising candidates for vaccine strategies intended to prevent and cure HIV/AIDS, as well as other chronic infections.

METHODS SUMMARY

RhCMV/SIV vectors expressing SIV Gag, Rev/Tat/Nef, Pol and Env or irrelevant control inserts were used as described[5,9]. RM were challenged with SIVmac239 by intrarectal, intravaginal or intravenous inoculation using a repeated (weekly), limiting dose protocol[5,9]. RM were considered infected after detection of plasma SIV RNA ≥ 30 copy equiv. ml^{-1} and the *de novo* development of T-cell responses to SIV Vif, an SIV antigen not included in the RhCMV/SIV vectors. Selected RM were depleted of $CD8^+$ lymphocytes, as described[14]. SIV-specific T-cell responses were measured in mononuclear cells from blood and tissues by flow cytometric intracellular cytokine analysis[5,9]. Levels of SIV RNA and DNA in tissue/cell preparations were quantified, respectively, by an ultrasensitive, nested qRT–PCR and qPCR approach[5]. Replication-competent SIV was detected in mononuclear cells by inductive co-cultivation assays, as described[15]. To ascertain the presence of occult replication-competent SIV below the level of detection of the co-cultivation assay, 60 million cells from blood, lymph nodes, bone marrow or spleen from RhCMV vector-protected RM or RM with spontaneously controlled or ART-suppressed SIV infection were adoptively transferred by intravenous infusion to SIV-naive RM, with infection in the recipient RM determined as described above.

Full Methods and any associated references are available in the online version of the paper.

Received 1 May; accepted 1 August 2013.
Published online 11 September 2013.

1. Chun, T. W. & Fauci, A. S. HIV reservoirs: pathogenesis and obstacles to viral eradication and cure. *AIDS* **26,** 1261–1268 (2012).
2. Lewin, S. R., Evans, V. A., Elliott, J. H., Spire, B. & Chomont, N. Finding a cure for HIV: will it ever be achievable? *J. Int. AIDS Soc.* **14,** 4 (2011).
3. Deeks, S. G. & Walker, B. D. Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. *Immunity* **27,** 406–416 (2007).
4. Picker, L. J., Hansen, S. G. & Lifson, J. D. New paradigms for HIV/AIDS vaccine development. *Annu. Rev. Med.* **63,** 95–111 (2012).
5. Hansen, S. G. *et al.* Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* **473,** 523–527 (2011).
6. Haase, A. T. Early events in sexual transmission of HIV and SIV and opportunities for interventions. *Annu. Rev. Med.* **62,** 127–139 (2011).
7. Lifson, J. D. *et al.* Containment of simian immunodeficiency virus infection: cellular immune responses and protection from rechallenge following transient postinoculation antiretroviral treatment. *J. Virol.* **74,** 2584–2593 (2000).
8. Sáez-Cirión, A. *et al.* Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of Early Initiated Antiretroviral Therapy ANRS VISCONTI Study. *PLoS Pathog.* **9,** e1003211 (2013).
9. Hansen, S. G. *et al.* Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nature Med.* **15,** 293–299 (2009).
10. Hansen, S. G. *et al.* Cytomegalovirus vectors violate $CD8^+$ T cell epitope recognition paradigms. *Science* **340,** 1237874 (2013).
11. Ribeiro dos Santos, P. *et al.* Rapid dissemination of SIV follows multisite entry after rectal inoculation. *PLoS ONE* **6,** e19493 (2011).
12. Deeks, S. G. *et al.* Towards an HIV cure: a global scientific strategy. *Nature Rev. Immunol.* **12,** 607–614 (2012).
13. Burton, D. R. *et al.* A blueprint for HIV vaccine discovery. *Cell Host Microbe* **12,** 396–407 (2012).
14. Okoye, A. *et al.* Profound $CD4^+/CCR5^+$ T cell expansion is induced by $CD8^+$ lymphocyte depletion but does not account for accelerated SIV pathogenesis. *J. Exp. Med.* **206,** 1575–1588 (2009).
15. Shen, A. *et al.* Novel pathway for induction of latent virus from resting $CD4^+$ T cells in the simian immunodeficiency virus/macaque model of human immunodeficiency virus type 1 latency. *J. Virol.* **81,** 1660–1670 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the AIDS Vaccine Research in Nonhuman Primates Consortium of the National Institute of Allergy and Infectious Diseases (NIAID; U19 AI095985), as well as other NIAID grants (RO1 AI060392, R37 AI054292, P01 AI094417 and U19 AI096109); the Bill & Melinda Gates Foundation-supported Collaboration for AIDS Vaccine Discovery (CAVD); the International AIDS Vaccine Initiative (IAVI) and its donors, particularly the US Agency for International Development (USAID); the National Center for Research Resources (P51 OD011092); and the National Cancer Institute (contract HHSN261200800001E). We thank A. Sylwester, A. Okoye, C. Kahl, S. Hagen, R. Lum, Y. Fukazawa, S. Shiigi and L. Boshears for technical or administrative assistance; C. Miller, N. Miller and T. Friedrich for provision of SIV stocks; K. Reimann for provision of the CD8-depleting monoclonal antibody; D. Watkins for MHC typing; D. Montefiori for neutralizing antibody assays; and A. Townsend for figure preparation.

Author Contributions S.G.H. planned and performed animal experiments, and analysed immunological and virological data, assisted by A.B.V., C.M.H., R.M.G., J.C.F., M.S.L., A.N.G., G.X. and N.W. M.P. and J.D.L. planned and performed SIV quantification, assisted by K.O., R.S. and Y.L. B.J.B. and J.B.S. performed infected cell recognition assays. B.F.K. performed sequencing analysis. J.D.E. performed immunohistological studies. S.L.P., J.M.T., A.W.L. and M.K.A. managed the animal protocols. J.A.N. and K.F. supervised CMV vector design and development. J.D.L. planned and supervised SIV quantification and immunohistological experiments. P.T.E. performed all statistical analyses. L.J.P. conceived the RhCMV vector strategy, supervised all experiments, analysed data, and wrote the paper, assisted by S.G.H., J.D.L. and J.B.S.

Author Information New SIVmac239 sequences reported in this manuscript are accessible in GenBank under accessions KF439057, KF439058 and KF439059. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D.L. (lifsonj@mail.nih.gov) or L.J.P. (pickerl@ohsu.edu).

METHODS

Rhesus macaques. Ninety-nine purpose-bred male and female RM (*M. mulatta*) of Indian genetic background were used with the approval of the Oregon National Primate Research Center Animal Care and Use Committee, under the standards of the US National Institutes of Health Guide for the Care and Use of Laboratory Animals. These animals were specific pathogen-free as defined by being free of cercopithicine herpesvirus 1, D-type simian retrovirus, simian T-lymphotrophic virus type 1, rhesus rhadinovirus, and *Mycobacterium tuberculosis.* MHC-1 genotyping for common *Mamu* alleles such as *Mamu-A*01*/*-A*02* and *Mamu-B*08*/*-B*17* was performed by sequence-specific priming PCR, as described[16]. The 99 RM include 15 RhCMV/SIV vector-vaccinated RM (7–10 years of age) with aviraemic control of SIV infection (14 with SIVmac239, 1 with SIVmac251) after intrarectal challenge (5 with short-term follow up; 10 with medium- to long-term follow up); 52 RM (4–19 years of age) vaccinated with RhCMV/SIV or control RhCMV vectors or left unvaccinated before SIVmac239 challenge (42 intravaginal, 10 intravenous); 12 SIV^+ RM (5–12 years of age; 8 with progressive SIVmac239 infection, 1 with spontaneously controlled SIVmac239 infection, 3 with ART-suppressed SIVmac251 infection); and 20 SIV-naive RM (4–14 years of age) used as negative controls (these include 4 RM vaccinated with RhCMV/SIV vectors) or as SIV- and RhCMV vector-naive recipients in the adoptive transfer experiments. Early (<1 year) follow-up of eight of the RhCMV/SIV vector-vaccinated RM with long-term, aviraemic control of SIV infection was previously reported[5], with this study extending that follow-up from 1 to >3 years. RhCMV/Gag, Rev/Tat/Nef, Env and Pol-1 and Pol-2 vectors were administered subcutaneously at a dose of 5×10^6 plaque forming units per vector. The control antigen-expressing RhCMV vector was used at a total dose of 2.5×10^6 plaque forming units to match the total dose of the RhCMV/SIV vectors. RM were vaccinated twice with RhCMV vectors, 14 weeks apart. All SIV challenges (intrarectal, intravaginal and intravenous) used a repeated limiting dose protocol with dosing designed to require >1 challenge for infection of >60% of challenged RM, and to infect all or nearly all challenged RM with ≤10 (weekly) challenges for intrarectal or intravaginal inoculation (300 focus-forming units) and ≤3 (every third week) challenges for intravenous inoculation (0.2 focus-forming units). RM were considered SIV-infected (and challenge discontinued) with the onset of plasma viral load ≥ 30 copy equiv. ml^{-1} and the *de novo* development of $CD4^+$ and $CD8^+$ T-cell responses to SIV Vif, an SIV antigen not included in the RhCMV/SIV vectors. RM were considered controllers if plasma viral load became undetectable (<30 copy equiv. ml^{-1}) within 2 weeks of the initial positive plasma viral load and was then maintained below threshold for three consecutive determinations. RM with progressive SIV infection were followed for 20 weeks after infection, or if progression was rapid, until the onset of AIDS. ARTs consisted of two reverse transcriptase inhibitors (20 mg day^{-1} tenofovir, 50 mg day^{-1} emtricitabine), an integrase inhibitor (240 mg day^{-1} raltegravir) and a protease inhibitor (600 mg twice daily darunavir boosted with 100 mg twice daily ritonavir). Selected RM were depleted of $CD8^+$ lymphocytes by administration of 10, 5, 5 and 5 mg per kg body weight of the CD8α monoclonal antibody M-T807R1, a modified version of the cM-T807 humanized anti-CD8 monoclonal antibody with rhesus constant and variable framework regions (http://nhpreagents.bidmc.harvard.edu), administered on days 0, 3, 7 and 10, respectively[14]. Tissues obtained by biopsy or at necropsy were processed for mononuclear cell preparation, virological and/or immunohistological analysis as previously described[5,17]. For adoptive transfer experiments, freshly obtained peripheral blood and bone marrow buffy coats were prepared by centrifugation (400*g* for 20 min). These buffy coats and/or freshly obtained whole lymph node cell or splenocyte preparations were washed three times in saline before intravenous infusion, with each RM receiving 3×10^7 peripheral blood leukocytes plus 3×10^7 lymph node cells (or in 1 RM, 3×10^7 bone marrow leukocytes plus 3×10^7 splenocytes) over 1 h.

Vectors and viruses. The construction and characterization of the strain 68-1-derived RhCMV/SIV vectors, including RhCMV(Gag), RhCMV(Rev/Tat/Nef), RhCMV(Env) and RhCMV(Pol-1) and RhCMV(Pol-2), have been previously described[5,9]. A control RhCMV vector expressing an *M. tuberculosis* Ag85B-ESAT6 fusion protein under the control of the EF1-α promoter was constructed with the same E/T recombination approach and RhCMV (68-1) bacterial artificial chromosome used for RhCMV/SIV construction[9]. RhCMV vector stocks were titred using primary rhesus fibroblasts in a tissue culture infectious dose 50 ($TCID_{50}$) assay. The pathogenic SIV challenge stocks used in these experiments were generated by expanding the SIVmac239 clone (or SIVmac251 swarm) in RM PBMCs, and were titred using the CMMT-CD4-LTR-β-Gal (sMAGI) cell assay (NIH AIDS Reagent Program).

Viral detection assays. Plasma viral loads were determined by quantitative RT–PCR as previously described[18,19]. Ultrasensitive determinations of plasma viral loads at necropsy were achieved by concentrating virus from the larger volumes of material available by ultracentrifugation (6041 10-ml tubes and 4018 crown assembly, Seton Scientific; T1270 rotor, Thermo-Sorvall Scientific) at 170,000*g* for 30 min before processing RNA. Reactions were also run in triplicate and followed the analysis recommendations described previously[20] permitting per reaction determinations of 1 copy (2 of 3 positive amplifications) and threshold sensitivities correspondingly lower, dependent on the amount of plasma input. Plasma viral sequencing of Rh20363 post-rebound was performed by synthesis of complementary DNA with SIV gene-specific primers followed by sequencing using the single genome amplification strategy[21] (GenBank accessions KF439057, KF439058 and KF439059). Quantitative assessment of SIV DNA and RNA in isolated cells and tissues was determined by the quantitative hybrid real-time/digital RT–PCR and PCR assays, essentially as previously described[5], but with modifications to allow more efficient processing of samples larger than ~100 mg and to increase sample throughput. With respect to the former, the large tissue samples were directly disrupted in TriReagent (Molecular Research Center) using two 11-mm stainless steel balls over 10–15 stainless steel hex nuts (5.5-mm wide) as grinding media, rather than first attempting to cryogenically pulverize the tissue. With respect to the latter, the RT–PCR and PCR assay conditions were also modified to reduce reaction volumes to allow use of 384-well plates. The cDNA reactions were reduced to 15 μl, comprising 10 μl sample plus 5 μl concentrated reaction cocktail and contained 2 mM SIVnestR01 primer, 10 U RNAsin, and 50 U MoMLV reverse transcriptase (Promega). The cDNA synthesis stage of the thermal profile was optimal for MoMLV reverse transcription at 37 °C for 60 min, as opposed to 42 °C for 40 min. The RT–PCR preamplification reactions were 25 μl in volume with 1.25 U PlatinumTaq polymerase (Life Technologies) and 2.5 μl of this reaction was transferred to 20 μl of real-time PCR reaction mix with 1 U PlatinumTaq polymerase. For DNA determinations, the preamplification reactions were 20 μl in volume, comprising 10 μl sample and 10 μl reaction cocktail; 2 μl of this 'nested' reaction was transferred to 20 μl of real-time PCR reaction mix. As previously described, for both RNA and DNA determinations, 12 replicate reactions were tested per sample including a spike of RNA or DNA internal control sequence standard in two of the twelve reactions to assess overall amplification efficiency and assess potential inhibition of the PCR or RT–PCR reactions. The amount of DNA or RNA standard added to replicate reactions to monitor inhibition and PCR performance was typically 10–100 copies, depending on the anticipated level of SIV sequences present. Samples showing greater than a five cycle shift in amplification of the spiked standard, compared to amplification in the absence of specimen nucleic acid, corresponding to less than 74% overall amplification efficiency, were diluted and re-assayed. Quantitative determinations for samples showing amplification in all replicates were derived directly with reference to a standard curve. Quantitative determinations for samples showing fewer than 10 positive amplifications in replicates were derived from the frequency of positive amplifications, corresponding to the presence of at least one target copy in a reaction, according to a Poisson distribution of a given median copy number per reaction. To avoid false positives in biopsy material, in which the specimen size and total number of specimens is limited, we required a minimum of two positive reactions out of ten for a sample to be considered positive. The presence of inducible, replication-competent SIV in mononuclear cell preparations derived from different tissue sites at necropsy was detected by co-cultivation of 2.5×10^5 unfractionated cells from each tissue with 2×10^5 CEMx174 cells (×20 replicates per tissue, cell numbers permitting; CEMx174 cells obtained from NIH AIDS Research & Reference Reagent Program)[15]. After 18 days, each culture was stained for CD3, CD4 and intracellular SIVgag-p27 (monoclonal antibody 55-2F12), with positive cultures based on ≥0.5% CEMx174 cells with intracellular SIV Gag expression over background by flow cytometry.

Immunological assays. SIV-specific $CD4^+$ and $CD8^+$ T-cell responses were measured in blood and tissues by flow cytometric intracellular cytokine analysis, as previously described in detail[5,9]. To determine T-cell responses to SIV peptide mixes or individual peptides, mononuclear cells were incubated with mixes of overlapping 15-amino-acid peptides comprising SIV proteins or individual epitopic 8- to 10-amino-acid peptides (with every individual peptide always at 2 μg ml^{-1}) and the co-stimulatory molecules CD28 and CD49d (BD Biosciences) for 1 h, followed by addition of brefeldin A (Sigma-Aldrich) for an additional 8 h. Co-stimulation without antigen served as a background control. To determine responses to autologous SIV-infected cells, SIV^+ and SIV^- target cells were produced by spinoculating (or not spinoculating) activated $CD4^+$ T cells with sucrose-purified SIVmac239, followed by culturing the cells for 4 days and then purifying the $CD4^+$ cells with CD4 microbeads and LS columns (Miltenyi Biotec), as described[22]. These cell preparations were >95% $CD4^+$ T cells and the SIV-infected preparations were >50% SIV^+ following enrichment. SIV^+ versus SIV^- T cells were then incubated with microbead-purified $CD8^+$ T cells at an effector:target ratio of 40:1 under the same conditions used for peptide-specific flow cytometric intracellular cytokine analysis. After incubation, stimulated cells were stored at 4 °C until staining with combinations of fluorochrome-conjugated monoclonal antibodies including: SP34-2 (CD3; Pacific Blue, PerCP-Cy5.5), L200 (CD4; AmCyan), SK-1 (CD8α; APC, PerCP-Cy5.5), CD28.2 (CD28; PE, PE-TexasRed), DX2 (CD95; APC, PE), 15053 (CCR7;

Pacific Blue), B56 (Ki-67; FITC), MAB11 (TNF-α; APC, FITC, PE), B27 (IFN-γ; APC, FITC) and FN50 (CD69; PE, PE-TexasRed). Data was collected on an LSR-II (BD Biosciences). Analysis was performed using FlowJo software (Tree Star). In all analyses, gating on the lymphocyte population was followed by the separation of the $CD3^+$ T-cell subset and progressive gating on $CD4^+$ and $CD8^+$ T-cell subsets. Antigen-responding cells in both $CD4^+$ and $CD8^+$ T-cell populations were determined by their intracellular expression of CD69 and either or both of the IFN-γ and TNF cytokines. After subtracting background, the raw response frequencies were memory corrected, as previously described[5,9]. In selected experiments, cells responding to SIV peptides by production of either or both of IFN-γ and TNF were directly phenotyped with respect to the memory markers CD28 and CCR7 (refs 5, 9). Titres of SIV Env-specific antibodies were determined by neutralization of tissue culture-adapted SIVmac251 using a luciferase reporter gene assay[23].

Immunohistology. Immunohistochemistry was performed using a biotin-free polymer approach (Golden Bridge International) on 5-μm tissue sections mounted on glass slides, which were dewaxed and rehydrated with double-distilled H_2O. Heat-induced epitope retrieval was performed by heating sections in 0.01% citraconic anhydride containing 0.05% Tween-20 in a pressure cooker set at 122–125 °C for 30 s. Slides were incubated with blocking buffer (TBS with 0.05% Tween-20 and 0.5% casein) for 10 min. For APOBEC3G, slides were incubated with rabbit anti-APOBEC3G (1:100; Sigma HPA001812) diluted in blocking buffer overnight at 4 °C. Slides were washed in 1× TBS with 0.05% Tween-20, endogenous peroxidases blocked using 1.5% (v/v) H_2O_2 in TBS, pH 7.4, for 10 min, incubated with rabbit polink-2 horseradish peroxidase (HRP) and developed with impact DAB (3,3′-diaminobenzidine; Vector Laboratories). For ISG15 staining, after the heat-induced epitope retrieval step, the slides were loaded on an IntelliPATH autostainer (Biocare Medical) and stained with optimal conditions determined empirically that consisted of a blocking step using blocking buffer (TBS with 0.05% Tween-20 and 0.5% casein) for 10 min and an endogenous peroxidase block using 1.5% (v/v) H_2O_2 in TBS, pH 7.4, for 10 min. Rabbit anti-ISG (1:250; Sigma HPA004627) was diluted in blocking buffer and incubated for 1 h at room temperature. Tissue sections were washed and developed using the Rabbit Polink-1 HRP staining system (Golden Bridge International) according to manufacturer's recommendations. Sections were developed with impact DAB (Vector Laboratories). All slides were washed in H_2O, counterstained with haematoxylin, mounted in Permount (Fisher Scientific), and scanned at high magnification (×200) using the ScanScope CS System (Aperio Technologies), yielding high-resolution data from the entire tissue section. Representative regions of interest (500 mm^2) were identified and high-resolution images extracted from these whole-tissue scans. The percentage area of the T-cell zone that stained for APOBEC3G and ISG15 was quantified using Photoshop CS5 and Fovea tools.

Statistical analysis. The RM used in the vaccine efficacy analysis were randomly assigned to vaccine groups with randomization stratified to balance groups for expression of protective MHC alleles. All reported experiments were conducted once and are reported fully. The criteria for categorizing post-challenge RM into 'protected' and 'unprotected' groups were established previously[5]. Experimenters were not explicitly blinded to the treatment assignments of the RM, nor were the analyses conducted by blinded investigators. All statistical analyses were conducted using non-parametric and model-independent analysis procedures either for the main analysis or as a sensitivity analysis, and in every sensitivity analysis the result was qualitatively consistent with the main analysis. No tests depended on an assumption of equal variance across compared groups. The only exceptions were the time series analyses, which were conducted with two model-based (regression) approaches; the residuals of these analyses were evaluated and found to be consistent with homoschedasticity and normality requirements, and the results were consistent across approaches. For comparisons of continuous-valued data from independent samples, we applied bivariate Mann–Whitney *U* tests[24], also known as Wilcoxon rank sum tests. For comparisons of dichotomous values across groups, we applied Fisher's exact tests[25]. We estimated confidence bounds for binomial proportions using the Wilson score method, as described in Agresti and Coull[26]. We compared group means of positivity frequencies for which we have repeated binary measures on individual RM using mixed effects logistic regression (with individual RM mean deviations from group means modelled as a normally-distributed random effect). We compared confidence intervals for RM groups to confidence intervals for individual RM (from other groups) by directly determining overlap of the intervals (and we used the estimated random effect variance and estimated group means to conduct z-tests in sensitivity analyses, which yielded consistent results). We compared Kaplan–Meier curves using the log-rank test[27]. We conducted time series analyses using standard linear regression with time as the primary predictor, and we used Gaussian first-order autoregressive process models in sensitivity analyses, which yielded consistent results. All tests were conducted as two-tailed tests with a type-I error rate of 5%. We used the R statistical computing language[28] for all statistical analyses.

16. Loffredo, J. T. *et al. Mamu*-B*08-positive macaques control simian immunodeficiency virus replication. *J. Virol.* **81,** 8827–8832 (2007).
17. Tabb, B. *et al.* Reduced inflammation and lymphoid tissue immunopathology in rhesus macaques receiving anti-tumor necrosis factor treatment during primary simian immunodeficiency virus infection. *J. Infect. Dis.* **207,** 880–892 (2013).
18. Cline, A. N., Bess, J. W., Piatak, M. Jr & Lifson, J. D. Highly sensitive SIV plasma viral load assay: practical considerations, realistic performance expectations, and application to reverse engineering of vaccines for AIDS. *J. Med. Primatol.* **34,** 303–312 (2005).
19. Venneti, S. *et al.* Longitudinal *in vivo* positron emission tomography imaging of infected and activated brain macrophages in a macaque model of human immunodeficiency virus encephalitis correlates with central and peripheral markers of encephalitis and areas of synaptic degeneration. *Am. J. Pathol.* **172,** 1603–1616 (2008).
20. Palmer, S. *et al.* New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. *J. Clin. Microbiol.* **41,** 4531–4536 (2003).
21. Keele, B. F. *et al.* Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. *J. Exp. Med.* **206,** 1117–1134 (2009).
22. Sacha, J. B. & Watkins, D. I. Synchronous infection of SIV and HIV *in vitro* for virology, immunology and vaccine-related studies. *Nature Protocols* **5,** 239–246 (2010).
23. Montefiori, D. C. Evaluating neutralizing antibodies against HIV, SIV, and SHIV in luciferase reporter gene assays. *Curr. Protoc. Immunol.* **Chapter 12,** Unit 12.11 (2005).
24. Wolfe, D. A. & Hollander, M. *Nonparametric Statistical Methods* (Wiley, 1973).
25. Fisher, R. A. The logic of inductive inference. *J. Roy. Stat. Soc. A* **98,** 39–54 (1935).
26. Agresti, A. & Coull, B. A. Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.* **52,** 119–126 (1998).
27. Harrington, D. P. & Fleming, T. R. A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566 (1982).
28. R Development Core Team. *R, A Language and Environment for Statistical Computing*; http://www.R-project.org (2011).

LETTER

doi:10.1038/nature12501

Podoplanin maintains high endothelial venule integrity by interacting with platelet CLEC-2

Brett H. Herzog[1,2]*, Jianxin Fu[1,3,4]*, Stephen J. Wilson[5], Paul R. Hess[6], Aslihan Sen[6], J. Michael McDaniel[1], Yanfang Pan[1,2], Minjia Sheng[1], Tadayuki Yago[1], Robert Silasi-Mansat[1], Samuel McGee[1], Frauke May[7], Bernhard Nieswandt[7], Andrew J. Morris[8], Florea Lupu[1], Shaun R. Coughlin[5], Rodger P. McEver[1,2], Hong Chen[1,2], Mark L. Kahn[6] & Lijun Xia[1,2,3,4]

Circulating lymphocytes continuously enter lymph nodes for immune surveillance through specialized blood vessels named high endothelial venules[1–5], a process that increases markedly during immune responses. How high endothelial venules (HEVs) permit lymphocyte transmigration while maintaining vascular integrity is unknown. Here we report a role for the transmembrane *O*-glycoprotein podoplanin (PDPN, also known as gp38 and T1α)[6–8] in maintaining HEV barrier function. Mice with postnatal deletion of *Pdpn* lost HEV integrity and exhibited spontaneous bleeding in mucosal lymph nodes, and bleeding in the draining peripheral lymph nodes after immunization. Blocking lymphocyte homing rescued bleeding, indicating that PDPN is required to protect the barrier function of HEVs during lymphocyte trafficking. Further analyses demonstrated that PDPN expressed on fibroblastic reticular cells[7], which surround HEVs, functions as an activating ligand for platelet C-type lectin-like receptor 2 (CLEC-2, also known as CLEC1B)[9,10]. Mice lacking fibroblastic reticular cell PDPN or platelet CLEC-2 exhibited significantly reduced levels of VE-cadherin (also known as CDH5), which is essential for overall vascular integrity[11,12], on HEVs. Infusion of wild-type platelets restored HEV integrity in *Clec-2*-deficient mice. Activation of CLEC-2 induced release of sphingosine-1-phosphate[13,14] from platelets, which promoted expression of VE-cadherin on HEVs *ex vivo*. Furthermore, draining peripheral lymph nodes of immunized mice lacking sphingosine-1-phosphate had impaired HEV integrity similar to *Pdpn*- and *Clec-2*-deficient mice. These data demonstrate that local sphingosine-1-phosphate release after PDPN–CLEC-2-mediated platelet activation is critical for HEV integrity during immune responses.

Lymph nodes (LNs) are essential sites for immune responses. They are organized into lobules, which are surrounded by lymphatic sinuses that deliver antigens from afferent lymphatic vessels to LNs for identification by naive lymphocytes that continually home through HEVs (Supplementary Fig. 1). Lymphocyte trafficking is particularly prominent in mucosal LNs, as most foreign antigens enter the body through mucosal epithelium, and in draining peripheral LNs during immune responses[1,15]. How HEVs accommodate a high rate of lymphocyte trafficking while maintaining their integrity remains unknown.

Platelets support vascular integrity in inflamed tissues by still undefined mechanisms[16]. Whether, and if so how, platelets protect HEV integrity in the LN is unexplored. PDPN, a ligand for the platelet activating receptor CLEC-2, is highly expressed in LNs. We developed mice with tamoxifen-inducible global deletion of PDPN (*Pdpn*$^{f/f}$;*CagCre*, Supplementary Fig. 2a–d), and focused first on the mucosal LNs of mice around the weaning-age because development of adaptive immunity occurs early and primarily through mucosal LNs[17]. Tamoxifen administration from postnatal day 1 (P1) to P5 resulted in ~90% reduction of PDPN at P15 and complete loss at 1 month in both mucosal and peripheral LNs (Supplementary Figs 2e and 3a, b, d). Beginning at P15 and progressively worsening, *Pdpn*$^{f/f}$;*CagCre* pups exhibited massive bleeding primarily in mucosal LNs including mesenteric LNs (MLNs) and cervical LNs (CLNs), but rarely in peripheral (inguinal and popliteal) LNs (Fig. 1a, Supplementary Fig. 3). Histology and confocal imaging of MLNs revealed large numbers of extravasated red blood cells (RBCs) around HEVs but not non-HEV vessels of *Pdpn*$^{f/f}$;*CagCre* mice (Fig. 1b–d). *Pdpn* deletion starting at 3–4 weeks of age resulted in a similar mucosal LN bleeding phenotype, suggesting that PDPN is also important for LN vascular integrity in adults (Supplementary Fig. 3c, d).

In LNs, PDPN is expressed by endothelial cells of lymphatic vessels but not by blood vessels including HEVs (Supplementary Fig. 4a). However, PDPN is also highly expressed on fibroblastic reticular cells (FRCs), which surround HEVs and express ER-TR7, αSMA and PDGFRβ[7,18](Supplementary Fig. 4b–d). To address whether PDPN on FRCs is essential for LN vascular integrity, we generated *Pdpn*$^{f/f}$;*PdgfrbCre* mice, which lack PDPN in FRCs but otherwise exhibit normal FRC organization (Supplementary Fig. 5b, e). Similar to *Pdpn*$^{f/f}$;*CagCre* mice, *Pdpn*$^{f/f}$;*PdgfrbCre* mice developed bleeding in mucosal LNs (Fig. 1e). *Pdpn*$^{f/f}$;*PdgfrbCre* mice also had reduced levels of PDPN on lymphatic endothelial cells (LECs) in LNs (Supplementary Fig. 5b–d). To rule out the contribution of endothelial PDPN to LN bleeding, we developed *Pdpn*$^{f/f}$;*Tie2Cre* mice, which lack PDPN specifically in LECs but not in FRCs (Supplementary Fig. 6a–d). Consistent with the previously described role of PDPN in the separation of blood and lymphatic vessels during embryonic development[8], *Pdpn*$^{f/f}$;*Tie2Cre* mice exhibited blood–lymphatic vessel mixing phenotype (Supplementary Fig. 6e, data not shown). However, *Pdpn*$^{f/f}$;*Tie2Cre* mice did not exhibit bleeding around HEVs in the LN (Supplementary Fig. 6e, f). These results indicate that PDPN on FRCs rather than LECs prevents bleeding in LNs.

CLEC-2 is the only known receptor for PDPN[9,10,19]. To determine its importance for LN vascular integrity, we depleted CLEC-2 in wild-type neonates using a CLEC-2-specific monoclonal antibody, INU1 (ref. 19). Administration of INU1 resulted in bleeding in mucosal LNs at P15 similar to *Pdpn*$^{f/f}$;*CagCre* mice (Fig. 1f, Supplementary Fig. 7a). To determine the role of CLEC-2 in adult LNs, we made *Clec-2*$^{-/-}$ bone marrow chimaeras (*Clec-2*$^{-/-}$ BM chimaera, Supplementary Fig. 7b, c), which, consistent with deleting *Pdpn* in adult mice, developed bleeding in mucosal LNs (Supplementary Fig. 7d). CLEC-2 is primarily expressed on platelets[10,19]; however, it is also expressed on myeloid and dendritic cells. To verify whether CLEC-2 on platelets is required to protect LN vascular integrity, we developed mice lacking CLEC-2

[1]Cardiovascular Biology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma 73104, USA. [2]Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104, USA. [3]Jiangsu Institute of Hematology, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215006, China. [4]Key Laboratory of Thrombosis and Hemostasis of Ministry of Health, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215006, China. [5]Cardiovascular Research Institute, University of California, San Francisco, California 94158, USA. [6]Department of Medicine and Division of Cardiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [7]University Hospital Wurzburg and Rudolf Virchow Center, DFG Research Center for Experimental Biomedicine, 97080 Wurzburg, Germany. [8]Division of Cardiovascular Medicine, University of Kentucky and Lexington Veterans Affairs Medical Center, Lexington, Kentucky 40502, USA.
*These authors contributed equally to this work.

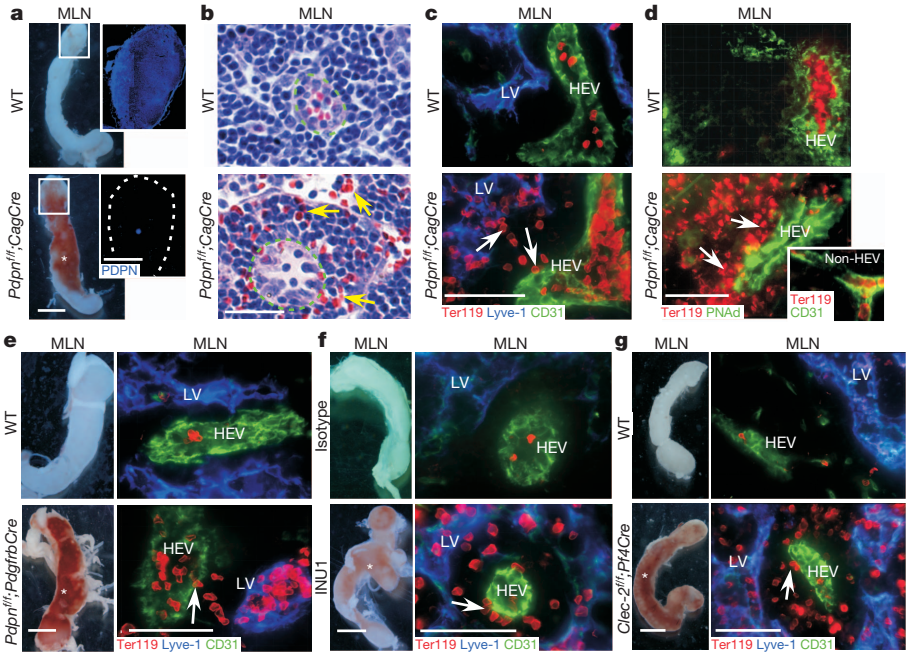


Figure 1 | Loss of fibroblastic reticular cell PDPN or platelet CLEC-2 leads to spontaneous mucosal LN bleeding. **a**, Gross morphology of MLNs. Insets contain montages of confocal images of MLN cryosections showing PDPN expression in wild-type (WT) and *Pdpn*$^{f/f}$;*CagCre* mice. **b**, Haematoxylin-and-eosin-stained MLN sections. Yellow arrows indicate extravasated RBCs outside HEVs (dashed line) of *Pdpn*$^{f/f}$;*CagCre* mice. **c**, Confocal images of MLNs from wild-type and *Pdpn*$^{f/f}$;*CagCre* mice reveal extravasated RBCs (arrows) outside HEVs, some of which are present in lymphatic vessels (LVs). Ter119 staining indicates RBCs. CD31 marks endothelial cells. Lyve-1 marks LVs. **d**, Immunostaining of MLN cryosections using HEV-specific marker PNAd. Inset shows that no bleeding occurred around CD31$^+$/PNAd$^-$ non-HEV vessels in MLNs. **e**, Gross morphology and confocal images of MLN cryosections from wild-type and *Pdpn*$^{f/f}$;*PdgfrbCre*. **f**, Gross morphology and confocal images of MLNs from P15 wild-type mice treated with isotype control (rat IgG1κ) or the CLEC-2 depleting antibody, INU1. **g**, Gross morphology and confocal images of MLN cryosections from wild-type and *Clec-2*$^{f/f}$;*Pf4Cre* mice. Data are representatives of ≥ 12 mice per group. Scale bars, 2 mm (light microscopy images), 50 μm (**b** and confocal images). Asterisk indicates bleeding in the LN. Arrows indicate extravasated RBCs. Tissues were from 1-month-old mice unless otherwise specified.

specifically on platelets (*Clec-2*$^{f/f}$;*Pf4Cre,* Supplementary Fig. 8a–c). *Clec-2*$^{f/f}$;*Pf4Cre* mice developed spontaneous bleeding in mucosal LNs reminiscent of *Pdpn*-deficient mice (Fig. 1g, Supplementary Fig. 8d). Collectively, these results indicate that FRC PDPN and platelet CLEC-2 are required to protect LN vascular integrity.

To explore where platelet CLEC-2 interacts with PDPN in the LN, we stained cryosections of LNs and found frequent associations of platelets with PDPN$^+$ FRCs at the abluminal sides of HEVs of wild-type but not *Pdpn*-deficient mice (Fig. 2a, Supplementary Fig. 9). Furthermore, intravenously transfused, fluorescently labelled CLEC-2-expressing wild-type platelets were detected on the abluminal sides of HEVs in close association with PDPN$^+$ FRCs of *Clec-2*$^{-/-}$ BM chimaeras (Fig. 2b), supporting the hypothesis that platelets interact with FRCs around HEVs in a PDPN–CLEC-2-dependent manner.

Mucosal LNs but not peripheral LNs of *Pdpn*- or *Clec-2*-deficient mice exhibit spontaneous bleeding, suggesting phenotypic differences between these two types of LNs. We found that mucosal LNs were significantly more permeable to Evans blue dye than peripheral LNs in the steady-state (Fig. 2c). We noted that mucosal but not peripheral HEVs express MAdCAM-1 (Supplementary Fig. 10a, b). However, after an immune challenge (ovalbumin/complete Freund's adjuvant, OVA/CFA), HEVs in peripheral LNs of wild-type mice had increased MAdCAM-1 expression and permeability, resembling mucosal HEVs (Supplementary Fig. 10b, data not shown). Importantly, *Pdpn*$^{f/f}$;*CagCre* or *Clec-2*$^{f/f}$;*Pf4Cre* but not wild-type mice developed bleeding in draining peripheral LNs after OVA/CFA challenge (Fig. 2d). These results suggest a crucial role for PDPN–CLEC-2 interactions in maintaining vascular integrity of steady-state mucosal and immunized peripheral LNs, which share a similar 'reactive' functional and molecular phenotype and are specifically vulnerable to bleeding. Consistent with this, bleeding was observed in mucosal LNs of 3-week-old *Pdpn*$^{f/f}$;*PdgfrbCre* mice, but not *Pdpn*$^{f/f}$;*PdgfrbCre* mice bred into a *Rag1*$^{-/-}$ background that lacks lymphocytes (Fig. 2e). In addition, depleting CLEC-2 with INU1 resulted in mucosal LN bleeding in 2-week-old wild-type but not *Rag1*$^{-/-}$ mice, highlighting the importance of lymphocytes in promoting mucosal LN bleeding (Supplementary Fig. 11a). We next administered a monoclonal antibody to L-selectin (Mel-14) that blocks lymphocyte trafficking through HEVs (Supplementary Fig. 11b)[20]. Mel-14 treatment considerably reduced the bleeding observed in the draining peripheral LN of immunized *Pdpn*$^{f/f}$;*CagCre* mice compared with an isotype control (Fig. 2f, Supplementary Fig. 11c). A previous study shows that activated platelets facilitate lymphocyte adhesion to HEVs[21]. However, whether platelets transmigrate with or following lymphocytes to the abluminal space of HEVs to fulfil additional function is unknown. We found that draining peripheral LNs from immunized wild-type mice had a significant increase in the number of platelets at the abluminal sides of HEVs, which was abolished after Mel-14 treatment, demonstrating that platelets migrate across HEVs in a lymphocyte transmigration-dependent manner to interact with FRCs (Supplementary Fig. 11d). These data support the model that interactions between platelets and FRCs are critical for preventing bleeding in 'reactive' LNs during immune responses when lymphocyte trafficking is increased.

To determine whether impaired HEV barrier function is responsible for the observed LN bleeding when PDPN is absent, we intravenously injected fluorescein isothiocyanate (FITC)-conjugated dextran immediately before euthanasia. FITC–dextran was contained within HEVs of wild-type MLNs. However, FITC–dextran leaked from HEVs, but not from non-HEV blood vessels in LNs and other organs, of *Pdpn*$^{f/f}$;*CagCre* mice (Fig. 3a, Supplementary Fig. 12a and data not shown). Furthermore, intravenously injected, fluorescently labelled RBCs were detected outside of HEVs in *Pdpn*$^{f/f}$;*CagCre* but not in wild-type mice (Fig. 3b). Ultrastructural analyses indicated that abnormal gaps between endothelial cell membranes were specific to HEVs of *Pdpn*$^{f/f}$;*CagCre* mice, and RBCs were frequently observed between or outside of high endothelial cells in *Pdpn*$^{f/f}$;*CagCre* mice (Fig. 3c, Supplementary Fig. 12b). Together, these data indicate that, in the absence of PDPN, impaired vascular barrier function and defective junctions are restricted to HEVs.

VE-cadherin (VE-cad) is an integral component of endothelial adherens junctions and barrier function[22]. We found that levels of VE-cad were reduced in HEVs of *Pdpn*$^{f/f}$;*CagCre* MLNs starting at P8, before the onset of bleeding (Fig. 3d, Supplementary Fig. 12c, d), consistent with the idea that loss of HEV junctional integrity contributes to the onset of bleeding. Furthermore, following the loss of PDPN, β-catenin, another component of endothelial adherens junctions[22], was decreased

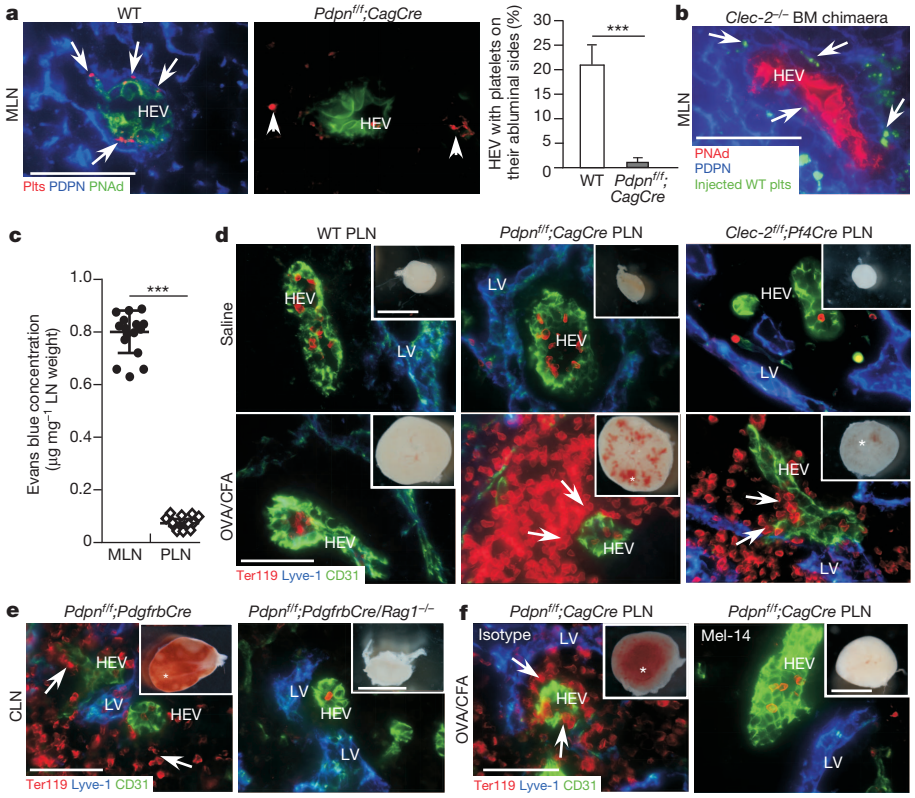


Figure 2 | FRC PDPN and platelet CLEC-2 protect LN vascular integrity during immune responses. **a**, Confocal images of MLN cryosections stained with antibodies to PNAd, platelets (plts) and PDPN. Arrows indicate platelets on the abluminal side of HEVs. Arrowheads indicate platelets that are not associated with HEVs. Bar graph on the right represents percentage of HEVs with platelets on their abluminal side (mean ± s.d., 250 HEVs per mouse, $n = 3$). **b**, Confocal images of PNAd$^+$ HEVs, PDPN, and transfused fluorescently labelled wild-type platelets in MLNs. Arrows indicate platelets on the abluminal side of HEVs. **c**, Comparison of Evans blue permeability between MLNs and popliteal LNs (PLNs) of 1-month-old wild-type mice, 5 min after intravenous injections (mean ± s.d., $n = 15$ per group). **d**, Gross morphology (insets) and confocal images of PLNs after OVA/CFA challenge. **e**, Gross morphology (insets) and confocal images of CLNs from 3-week-old mice with (*Pdpn*$^{f/f}$;*PdgfrbCre*) or without (*Pdpn*$^{f/f}$;*PdgfrbCre*/*Rag1*$^{-/-}$) lymphocytes. **f**, Gross morphology (insets) and confocal images of PLNs from 1-month-old mice, 1 week after OVA/CFA challenge and injections with a monoclonal antibody (Mel-14) that blocks L-selectin-dependent lymphocyte homing or with an isotype control rat IgG. Data represent at least 8 LNs per group from at least three experiments. Scale bars, 2 mm (light microscopy), 50 μm (confocal). Asterisk indicates bleeding and arrows mark extravasated RBCs around HEVs (**d**–**f**). ***$P < 0.001$.

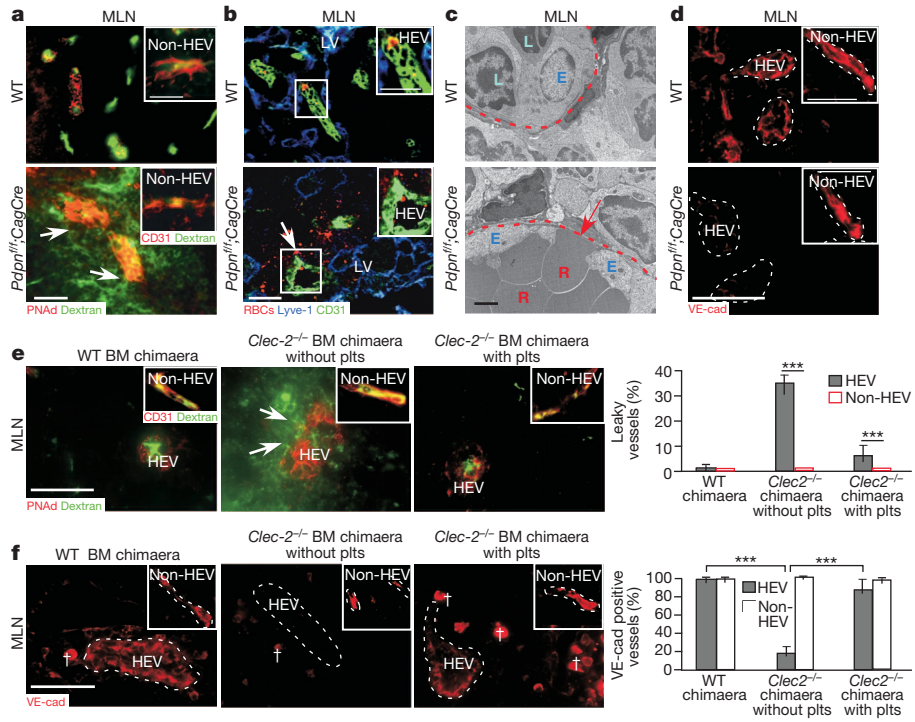


Figure 3 | Interactions between FRC PDPN and platelet CLEC-2 are critical for HEV junctional integrity. **a**, Confocal images of intravenously injected FITC–dextran (2,000 kDa) in P15 MLNs. Arrows indicate vascular leak of FITC–dextran. Insets show non-HEV blood vessels. **b**, Confocal images of intravenously injected RBCs (fluorescently labelled, red) in MLNs from 2-month-old mice. Arrow shows labelled RBCs outside of HEVs (CD31$^+$). **c**, Transmission electron micrographs of HEVs in MLNs. Arrow indicates gaps and RBCs between high endothelial cells. **d**, Confocal images of VE-cad in P8 MLNs. Insets show VE-cad staining in non-HEV blood vessels. **e**, Confocal analysis of intravenously injected FITC–dextran in MLNs from wild-type or *Clec2*$^{-/-}$ BM chimaeras with or without previous transfusions of wild-type platelets. Arrows indicate vascular leak of FITC–dextran. Inset depicts non-HEV blood vessels. Graphs on the right quantify number of leaking vessels (mean ± s.d., 300 vessels per group, $n = 3$). **f**, Anti-VE-cad staining of HEVs and non-HEV blood vessels in wild-type BM chimaeras, and *Clec2*$^{-/-}$ BM chimaeras without or with previous transfusions with wild-type platelets. Dagger indicates nonspecific staining as also observed in isotype controls. Graphs on the right quantify VE-cad staining (mean ± s.d., 300 vessels per group, $n = 3$). Tissues were from 1-month-old (**a**–**c**) or 12-week post-BM transplantation (**e**, **f**) mice unless otherwise specified. Dashed lines mark HEVs. Data represent at least three individual experiments. Scale bars, 2 μm (transmission electron microscopy), 50 μm (confocal images (inset **a**, **b** and **d**, 25 μm)). ***$P < 0.001$.

(Supplementary Fig. 12e) and expression of N-cadherin, which is essential for endothelial–stromal cell junctions[22], was reduced around HEVs (Supplementary Fig. 12f). Further analyses indicate that VE-cad was expressed at higher levels in peripheral HEVs than in mucosal HEVs of wild-type mice (Supplementary Fig. 13a). ZO-1 (also known as TJP1)[22], another junction molecule, was detected on peripheral but not mucosal HEVs. However, HEVs in draining peripheral LNs of immunized wild-type mice exhibited reduced VE-cad and ZO-1 expression, resembling MLNs of wild-type mice (Supplementary Fig. 13a, b). Consistent with the bleeding in draining peripheral LNs of immunized *Pdpn*- or *Clec-2*-deficient mice (Fig. 2d), these LNs exhibited a further reduction of VE-cad on their HEVs (Supplementary Fig. 13a), which was normalized by blocking lymphocyte trafficking with Mel-14 (Supplementary Fig. 13c). Furthermore, blocking VE-cad increased permeability in draining peripheral LNs of immunized wild-type mice (Supplementary Fig. 14). These results support the idea that interactions of PDPN and CLEC-2 preserve HEV barrier function primarily by promoting VE-cad expression on HEVs.

HEVs of *Clec-2*$^{-/-}$BM chimaeras exhibited increased permeability to intravenously injected FITC–dextran (Fig. 3e). To test whether this phenotype could be rescued by wild-type platelets, *Clec-2*$^{-/-}$ BM chimaeras were transfused daily with wild-type platelets for 4 days before euthanasia. HEVs of *Clec-2*$^{-/-}$ BM chimaeras that received wild-type platelets exhibited significantly decreased permeability to FITC–dextran compared to controls (Fig. 3e). Notably, wild-type and *Clec-2*$^{-/-}$BM chimaeras transfused with wild-type platelets expressed comparable levels of VE-cad on HEVs (Fig. 3f, Supplementary Fig. 15). Together, these results demonstrate that PDPN and platelet CLEC-2 are essential for preserving adherens junctions of HEVs.

Platelet aggregation is essential for haemostasis[16,23]. However, blocking aggregation with a monoclonal antibody to integrin αIIbβ3 did not increase permeability in draining peripheral LNs of immunized wild-type mice (Supplementary Fig. 14). This suggests that a platelet function mediated by PDPN–CLEC-2 apart from aggregation is required for HEV integrity. Sphingosine-1-phosphate (S1P)[13,14], a bioactive lipid, is a strong candidate for mediating this platelet function as it is known to regulate vascular integrity through interactions with its G-protein-coupled receptors on endothelial cells[13].

Although platelets generate and store S1P[24], whether PDPN–CLEC-2-mediated platelet activation causes S1P release is unknown. To test this, we stimulated wild-type and *Clec-2*$^{-/-}$ platelets with the monoclonal antibody INU1, which activates CLEC-2 signalling *in vitro*[19], and observed a CLEC-2-dependent S1P release from platelets (Fig. 4a). Furthermore, PDPN^{+} but not PDPN^{-} melanoma cells induced the release of S1P from wild-type but not *Clec-2*$^{-/-}$ platelets (Supplementary Fig. 16a, b), indicating that interactions between PDPN and CLEC-2 induce S1P release from platelets. Next, we found that MLN slices cultured *ex vivo* with fetal bovine serum (FBS), but not lipid-depleted FBS, retained VE-cad expression on their HEVs (Fig. 4b, Supplementary Fig. 16c), supporting the idea that lipids, such as S1P, maintain HEV adherens junctions. The addition of wild-type but not *Clec-2*$^{-/-}$ platelets rescued VE-cad levels on HEVs of MLN slices cultured in lipid-depleted FBS-containing media. Wild-type platelets were unable to restore HEV VE-cad expression on MLN slices from *Pdpn*-deficient mice (Fig. 4b, Supplementary Fig. 16c), demonstrating that FRC PDPN and platelet CLEC-2 are required for normal VE-cad expression on HEVs. The increase of VE-cad on HEVs in the presence of wild-type platelets is due, at least in part, to the activation of the S1P

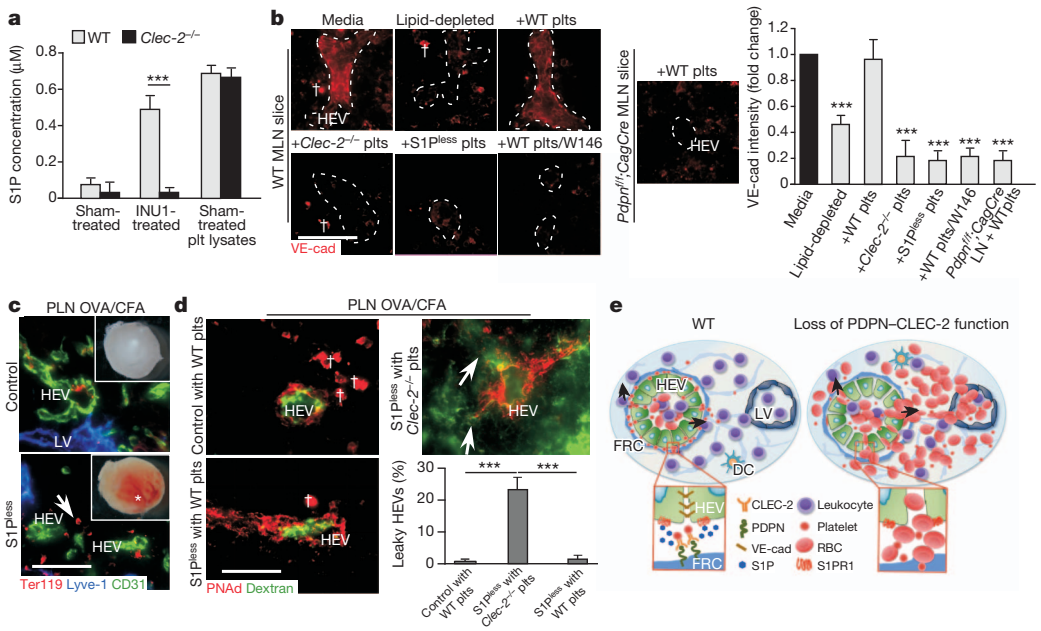


Figure 4 | S1P release from platelets after PDPN–CLEC-2-dependent activation contributes to HEV barrier function. **a**, S1P concentrations in supernatants of wild-type and *Clec-2*$^{-/-}$platelets after incubation with CLEC-2 activating antibody, INU1 or isotype control (sham-treated). S1P in platelet lysates was used as the positive control (mean ± s.d., $n = 4$ mice per group representing two individual experiments). **b**, Representative images of VE-cad staining of HEVs from wild-type LN slices incubated for 1.5 h with DMEM and normal FBS (media), DMEM and lipid-depleted FBS, lipid-depleted and wild-type platelets, lipid-depleted and *Clec-2*$^{-/-}$ platelets, lipid-depleted and S1P^{less} platelets, or lipid-depleted and wild-type platelets plus S1PR1 antagonist W146. *Pdpn*$^{f/f}$*;CagCre* LN slices incubated with lipid-depleted and wild-type platelets were controls. One hundred HEVs were analysed per condition. Dashed lines mark HEVs. Dagger marks nonspecific staining as also observed in isotype controls. Graphs represent ratios of VE-cad intensities on HEVs relative to that of wild-type lymph node slices cultured with media (mean ± s.d., $n = 20$ HEVs per group). **c**, Gross morphology (insets) and confocal images of draining PLNs after immunization. Asterisk indicates bleeding. Arrow marks bleeding (Ter119^{+}) around an HEV (CD31^{+}). **d**, Confocal images of PLN HEVs from S1P^{less} mice transfused with wild-type or *Clec-2*$^{-/-}$ platelets for 4 days after intravenous FITC–dextran injection. Arrows indicate vascular leak of FITC–dextran. Graphs on the right quantify leaking HEVs (mean ± s.d., 50 HEVs per group, $n = 3$). **e**, Model depicting how PDPN maintains HEV integrity during lymphocyte trafficking. FRC PDPN engages CLEC-2 on extravasated platelets in the perivenular space of HEVs and induces local release of S1P, which promotes VE-cad expression on the wild-type HEV (left). In contrast, loss of the interaction results in impaired HEV integrity and subsequent bleeding (right). **b–d**, Data are from three individual experiments. Scale bars, 50 μm. ***$P < 0.001$.

receptor 1 (S1PR1), because the S1PR1 antagonist, W146 (ref. 14), blocked the wild-type platelet-induced VE-cad expression on HEVs (Fig. 4b, Supplementary Fig. 16c). Furthermore, platelets from S1P-deficient (referred to here as $S1P^{less}$) mice[13], did not rescue VE-cad expression on HEVs of MLNs (Fig. 4b, Supplementary Fig. 16c), supporting the importance of platelet S1P. $S1P^{less}$ mice did not develop spontaneous mucosal LN bleeding, probably due to reduced lymphocyte transmigration through HEVs because of lower circulating lymphocytes[25] (Supplementary Fig. 17a). However, after OVA/CFA challenge of the hindlimb, the draining peripheral LNs of $S1P^{less}$ mice exhibited bleeding, increased permeability to FITC–dextran, and decreased VE-cad expression on HEVs, resembling mice lacking PDPN or CLEC-2 (Fig. 4c, d, Supplementary Fig. 17b–d). A daily transfusion for four consecutive days of wild-type but not *Clec-2*$^{-/-}$ platelets resulted in higher levels of VE-cad on HEVs and reduced vascular leak of injected FITC–dextran from HEVs of draining peripheral LNs of immunized $S1P^{less}$ mice (Fig. 4d, Supplementary Fig. 18). Taken together, these data indicate that PDPN–CLEC-2-dependent local release of S1P from platelets plays a critical role in maintaining HEV integrity during immune responses.

Unlike other venules, HEVs are not circumscribed by typical pericytes and collagen-containing matrix that would activate platelets. Instead, they are surrounded by a perivenular sleeve of FRCs that sequester collagen fibres[26,27]. Our findings reveal an important new role for FRC PDPN, which is well positioned to interact with CLEC-2 on extravasated platelets (Fig. 4e). S1P in the blood is known to regulate vascular integrity[13]. Our data demonstrate that PDPN–CLEC-2-dependent platelet activation causes release of S1P in the perivenular space that preserves VE-cad expression on HEVs (Fig. 4e). Thus, cross-talk between FRCs, platelets and HEVs is essential to maintain HEV integrity in situations of increased lymphocyte trafficking such as chronic inflammation. Recently, components of the CLEC-2 signalling pathway have been implicated in preventing inflammation-induced haemorrhage in the skin and lung[28]. Therefore, PDPN–CLEC-2-mediated local S1P release from platelets may protect vascular integrity in other inflamed tissues.

METHODS SUMMARY

Mice were maintained in specific-pathogen-free facilities and used under protocols approved by the IACUC of the Oklahoma Medical Research Foundation. For all experiments, a minimum of six mutants of each line and littermates were examined unless otherwise stated based on the highly penetrant phenotypes observed in our preliminary studies. Data are expressed as mean ± s.d. and represent at least three experiments unless otherwise specified. Statistical analysis was performed with Student's *t*-tests and differences were considered significant when $P < 0.05$. Generation of novel mouse lines, bone marrow transplantation, depletion of CLEC-2, platelet transfusion, immune challenge, L-selectin blocking, permeability assays, platelet and RBC labelling/transfusions, immunofluorescence imaging, flow cytometry, electron microscopy, S1P concentration analysis and *ex vivo* LN slice experiments are described in detail in the Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 5 October 2012; accepted 23 July 2013.
Published online 1 September 2013.

1. Butcher, E. C. & Picker, L. J. Lymphocyte homing and homeostasis. *Science* **272,** 60–67 (1996).
2. Drayton, D. L., Liao, S., Mounzer, R. H. & Ruddle, N. H. Lymphoid organ development: from ontogeny to neogenesis. *Nature Immunol.* **7,** 344–353 (2006).
3. Girard, J. P., Moussion, C. & Forster, R. HEVs, lymphatics and homeostatic immune cell trafficking in lymph nodes. *Nature Rev. Immunol.* **12,** 762–773 (2012).
4. Rosen, S. D. Ligands for L-selectin: homing, inflammation, and beyond. *Annu. Rev. Immunol.* **22,** 129–156 (2004).
5. von Andrian, U. H. & Mempel, T. R. Homing and cellular traffic in lymph nodes. *Nature Rev. Immunol.* **3,** 867–878 (2003).
6. Breiteneder-Geleff, S. *et al.* Podoplanin, novel 43-kd membrane protein of glomerular epithelial cells, is down-regulated in puromycin nephrosis. *Am. J. Pathol.* **151,** 1141–1152 (1997).
7. Farr, A. G. *et al.* Characterization and cloning of a novel glycoprotein expressed by stromal cells in T-dependent areas of peripheral lymphoid tissues. *J. Exp. Med.* **176,** 1477–1482 (1992).
8. Fu, J. *et al.* Endothelial cell *O*-glycan deficiency causes blood/lymphatic misconnections and consequent fatty liver disease in mice. *J. Clin. Invest.* **118,** 3725–3737 (2008).
9. Bertozzi, C. C. *et al.* Platelets regulate lymphatic vascular development through CLEC-2-SLP-76 signaling. *Blood* **116,** 661–670 (2010).
10. Watson, S. P., Herbert, J. M. & Pollitt, A. Y. GPVI and CLEC-2 in hemostasis and vascular integrity. *J. Thromb. Haemost.* **8,** 1456–1467 (2010).
11. Carmeliet, P. *et al.* Targeted deficiency or cytosolic truncation of the VE-cadherin gene in mice impairs VEGF-mediated endothelial survival and angiogenesis. *Cell* **98,** 147–157 (1999).
12. Grazia Lampugnani, M. *et al.* Contact inhibition of VEGF-induced proliferation requires vascular endothelial cadherin, β-catenin, and the phosphatase DEP-1/CD148. *J. Cell Biol.* **161,** 793–804 (2003).
13. Camerer, E. *et al.* Sphingosine-1-phosphate in the plasma compartment regulates basal and inflammation-induced vascular leak in mice. *J. Clin. Invest.* **119,** 1871–1879 (2009).
14. Gaengel, K. *et al.* The sphingosine-1-phosphate receptor S1PR1 restricts sprouting angiogenesis by regulating the interplay between VE-cadherin and VEGFR2. *Dev. Cell* **23,** 587–599 (2012).
15. Soderberg, K. A. *et al.* Innate control of adaptive immunity via remodeling of lymph node feed arteriole. *Proc. Natl Acad. Sci. USA* **102,** 16315–16320 (2005).
16. Goerge, T. *et al.* Inflammation induces hemorrhage in thrombocytopenia. *Blood* **111,** 4958–4964 (2008).
17. Renz, H., Brandtzaeg, P. & Hornef, M. The impact of perinatal immune development on mucosal homeostasis and chronic inflammation. *Nature Rev. Immunol.* **12,** 9–23 (2012).
18. Chyou, S. *et al.* Fibroblast-type reticular stromal cells regulate the lymph node vasculature. *J. Immunol.* **181,** 3887–3896 (2008).
19. May, F. *et al.* CLEC-2 is an essential platelet-activating receptor in hemostasis and thrombosis. *Blood* **114,** 3464–3472 (2009).
20. Gallatin, W. M., Weissman, I. L. & Butcher, E. C. A cell-surface molecule involved in organ-specific homing of lymphocytes. *Nature* **304,** 30–34 (1983).
21. Diacovo, T. G. *et al.* Platelet-mediated lymphocyte delivery to high endothelial venules. *Science* **273,** 252–255 (1996).
22. Dejana, E., Tournier-Lasserve, E. & Weinstein, B. M. The control of vascular integrity by endothelial cell junctions: molecular basis and pathological implications. *Dev. Cell* **16,** 209–221 (2009).
23. Hodivala-Dilke, K. M. *et al.* β3-integrin-deficient mice are a model for Glanzmann thrombasthenia showing placental defects and reduced survival. *J. Clin. Invest.* **103,** 229–238 (1999).
24. Ulrych, T. *et al.* Release of sphingosine-1-phosphate from human platelets is dependent on thromboxane formation. *J. Thromb. Haemost.* **9,** 790–798 (2011).
25. Pappu, R. *et al.* Promotion of lymphocyte egress into blood and lymph by distinct sources of sphingosine-1-phosphate. *Science* **316,** 295–298 (2007).
26. Gretz, J. E., Anderson, A. O. & Shaw, S. Cords, channels, corridors and conduits: critical architectural elements facilitating cell interactions in the lymph node cortex. *Immunol. Rev.* **156,** 11–24 (1997).
27. Katakai, T. *et al.* Lymph node fibroblastic reticular cells construct the stromal reticulum via contact with lymphocytes. *J. Exp. Med.* **200,** 783–795 (2004).
28. Boulaftali, Y. *et al.* Platelet ITAM signaling is critical for vascular integrity in inflammation. *J. Clin. Invest.* **123,** 908–916 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Kincade and L. Thompson for critical reading of the manuscript; R. Adams for providing *PdgfrbCre* mice; and M. Kinter and M.C. Marlin for technical assistance. Work was supported by grants from the National Institutes of Health (GM103441, GM097747, HL085607, HL093242, HL103432, HL065590, HL112788), VA Merit Award (BX001984), the American Heart Association (SDG7410022), the Deutsche Forschungsgemeinschaft (SFB688), National Natural Science Foundation of China (30928010), Jiangsu Provincial Special Program of Medical Science (BL2012005) and Jiangsu Province's Key Medical Center (ZX201102).

Author Contributions B.H.H. and J.F. designed and performed experiments, analysed results and drafted the manuscript. J.M.M., Y.P., M.S., T.Y., R.S.-M., S.M., A.J.M. and F.L. performed experiments. S.J.W., P.R.H., A.S., F.M., B.N. and S.R.C. supplied key reagents and mice. R.P.M., H.C. and M.L.K. helped analyse results and commented on the manuscript. L.X. designed and supervised research and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.X. (Lijun-Xia@omrf.org).

METHODS

Mice. To generate mice with *loxP*-flanked *Pdpn* alleles (*Pdpn*$^{f/f}$), a targeting vector was constructed in which exon 2, the major coding exon of *Pdpn* gene, was flanked by *loxP* sites (*Pdpn*$^{f/f}$), and the neomycin resistance selection cassette (Neo) was flanked by *Frt* sites (Supplementary Fig. 2a). The NotI-linearized construct was electroporated into C57BL/6-derived embryonic stem (ES) cells, and correctly targeted clones were identified by Southern blots. The *Frt*-flanked Neo was removed by transient expression of Flp recombinase in the positive ES clones to avoid potential undesirable effects of the Neo cassette. ES cells with normal karyotype bearing a floxed *Pdpn* allele were microinjected into B6/Tyr blastocytes, which were subsequently implanted into pseudopregnant foster mothers. Male chimaeras were bred with C57BL/6J females for germline transmission. Heterozygous mice were then crossed to generate *Pdpn*$^{f/f}$ mice (Supplementary Fig. 2a, b).

To generate mice with inducible deletion of PDPN (*Pdpn*$^{f/f}$;*CagCre*), *Pdpn*$^{f/f}$ mice were crossed with the CAG-Cre-ER^{T2} Tg mice (B6.Cg-Tg(CAG-cre/Esr1*)5Amc/J, Jackson Laboratories)[29]. To induce postnatal deletion of *Pdpn*, tamoxifen (MP Biomedical) was dissolved in ethanol/sunflower oil (1:9) and administered orally (20 μg per day) to pups from postnatal day (P) 1–5. Adult deletion was accomplished by administering tamoxifen orally (1 mg per day) for 5 consecutive days beginning at P21, then once a week thereafter. Wild-type littermates (*Pdpn*$^{f/w}$; *CagCre* or *Pdpn*$^{f/f}$) treated with the same regimen were used as controls. Mice deficient for *Pdpn* in pericytes/fibroblasts including FRCs (*Pdpn*$^{f/f}$;*PdgfrbCre*) or in endothelial cells (*Pdpn*$^{f/f}$;*Tie2Cre*) were generated by crossing *Pdpn*$^{f/f}$ mice with *PdgfrbCre* Tg mice (Tg(Pdgfrb-cre)9Rha)[30] or *Tie2Cre* Tg mice (Tg(Tek-cre)1Ywa)[31], respectively.

Conditional *Clec-2* knockout mice were generated in which exons 3 and 4 of the *Clec-2* allele are flanked by *loxP* sites (*Clec-2*$^{f/f}$). Deletion of exons 3 and 4 induces a premature stop codon that blocks the expression of the extracellular domain of CLEC-2 (Supplementary Fig. 8a). ES clones with correct homologous recombination were microinjected into B6 blastocytes, which were subsequently implanted into pseudopregnant foster mothers. Male chimaeras were bred with ACTB-FLP1 females (B6.Cg-Tg(ACTFLPe)9205Dym/J, Jackson Laboratory) for germline transmission and removal of the Neo cassette. Heterozygous mice were then crossed to generate *Clec-2*$^{f/f}$ mice. *Clec-2*$^{f/f}$ mice were crossed with *Pf4Cre* Tg mice (C57BL/6-Tg(Pf4-cre)Q3Rsko/J, Jackson Laboratory) to generate mice deficient for *Clec-2* on platelets (*Clec-2*$^{f/f}$;*Pf4Cre* mice).

Clec-2$^{-/-}$ (ref. 9), *Rag1*$^{-/-}$ (Jackson Laboratories)[32] and S1P^{less} mice[13] were previously described.

Mice were housed in specific pathogen-free barrier facilities. All mice were of mixed genetic background (129S and C57BL/6J) unless otherwise stated. Sex- and age-matched wild-type littermate controls were used for all experiments. Mice were included in studies if they exhibited deletion efficiency of more than 75% for the *Pdpn* deficiency and 95% for the *Clec-2* deficiency. Random assignment to treatment groups was used in antibody blocking and platelet transfusion studies *in vivo*. No blinding was used in this study. Animal studies were approved by the Institutional Animal Care and Use Committee of the Oklahoma Medical Research Foundation.

Microscopy. Organs were photographed at autopsy. For histology, samples were fixed in 4% paraformaldehyde (PFA) overnight at 4 °C, washed and embedded in paraffin. Sections (5 μm) were stained with haematoxylin and eosin. Confocal microscopy was performed as previously described[8]. Briefly, tissues were fixed in 4% PFA overnight at 4 °C, washed in PBS, cryoprotected in 20% sucrose in PBS at 4 °C overnight, embedded in 50% tissue freezing medium/50% OCT and cryosectioned (20–30-μm). Sections were blocked for 1 h at room temperature and incubated with primary antibodies overnight at 4 °C. Secondary antibodies were added for 1 h at room temperature. Primary antibodies were to murine thrombocytes (Accurate Chemical & Scientific Corporation), PDPN (clone 8.1.1, Developmental Studies Hybridoma Bank), CD41 (clone MWReg30, eBioscience), PDGFRβ (Santa Cruz), CD31 (clone 2H8, Abcam), CD31 (clone MEC13.3, BD Pharmingen), Ter119 (BD Pharmingen), PNAd (clone MECA-79, BioLegend), ER-TR7 (BioLegend), VE-cadherin (clone 11D4.1, BD Pharmingen), N-cadherin (Abcam), β-catenin (clone 14, BD Pharmingen), MAdCAM-1 (clone MECA-367, from E. Butcher, Department of Pathology, Stanford University School of Medicine), as well as biotinylated-goat anti-mouse Lyve-1 (R&D Systems). Corresponding secondary antibodies were conjugated with DyLight 488, AlexaFluor 555, or DyLight 649 (Jackson ImmunoResearch). Alternatively, Cy3-conjugated anti-αSMA (clone 1A4, Sigma-Aldrich) was also used. Images were collected using an Olympus IX81 with DSU spinning-disk confocal microscope and a Hamamatsu ORCA-R^2 camera. Images were analysed using Slidebook 5.0 (Intelligent Imaging Innovations).

Immunoblotting. As previously described[8], tissues were collected, homogenized and total protein concentration was calculated using a spectrophotometer (Eppendorf). Tissue lysates containing 20 μg total protein were resolved on a 10% SDS–PAGE resolving gel and transferred to an Immobilon-P membrane (EMD Millipore). Membranes were probed with a Syrian hamster anti-mouse PDPN (clone 8.1.1), rat anti-mouse VE-cad (clone BV13, Abcam), or mouse anti-GAPDH monoclonal antibodies. Membranes were washed and probed with horseradish peroxidase-conjugated secondary antibodies. Signal was developed using ECL reagents (Thermo Scientific).

Bone marrow chimaeras. Bone marrow cells from both femurs and tibias were collected from wild-type and a limited number of *Clec-2*$^{-/-}$ mice that survived after weaning in our facility by flushing the marrow cavity with 5 ml of HBSS containing 1 mM EDTA, as previously described[8]. The cell suspension was then run through a 100 μm mesh cell strainer to remove aggregates, centrifuged at 300*g* for 8 min and resuspended in sterile saline. Cells (5×10^6) in 200 μl saline were injected retro-orbitally into wild-type recipient mice that had been irradiated with 1100 rad. Engraftment efficiency was determined using flow cytometry analysis on a FACSCalibur (Becton Dickinson) with an antibody against CLEC-2 (clone 17D9, from C. Sousa, Immunobiology Laboratory, Cancer Research UK, London Research Institute).

Flow cytometric analysis of PDPN expression in LNs. PDPN expression on LN cells was analysed by flow cytometry. Briefly, LNs were digested in digestion buffer (RPMI-1640 (Invitrogen) containing 0.2 mg ml^{-1} collagenase P (Roche), 0.1 mg ml^{-1} DNase I (Invitrogen) and 0.8 mg ml^{-1} dispase (Roche)) at 37 °C for 20 min. Digested LNs were vigorously mixed to ensure disruption of capsule and release of leukocytes. Cell suspensions were placed in ice-cold PBS containing 2% FCS and 5 mM EDTA and centrifuged at 300*g* for 10 min at 4 °C. Pellets were resuspended in digestion buffer. Washed cells (5×10^6) were blocked for 15 min on ice (in HBSS containing Ca^{2+} and Mg^{2+}, 2% horse serum, 20 μg ml^{-1} anti-CD16/CD32 (clone 2.4G2, BD Pharmingen)). Isolated cells were incubated with antibodies to PDPN (clone 8.1.1) and biotinylated CD31 (clone MEC13.3, BD Pharmingen) for 30 min on ice and washed. Cells were then incubated with fluorescent dye-conjugated antibodies: AF488-conjugated goat anti-Syrian hamster IgG (Jackson ImmunoResearch), PerCP-conjugated streptavidin (BioLegend) and PE-conjugated rat anti-mouse CD45 (clone 30-F11, BD Pharmingen) for 20 min on ice. Flow cytometry analysis was performed on a FACSCalibur (Becton Dickinson).

Postnatal CLEC-2 depletion. Wild-type and *Rag1*$^{-/-}$ mice were intraperitoneally (i.p.) injected with 8 μg per gram of body weight of a monoclonal antibody to CLEC-2 (clone INU1)[19] at P1, P6 and P11. These mice were then killed on P15.

Labelling of erythrocytes or platelets. For erythrocyte labelling, whole blood from wild-type mice was collected into EDTA-coated Microvette 500 K3E tubes (Sarstedt), washed in HBSS (without Ca^{2+} and Mg^{2+}) and centrifuged at 1200*g* for 10 min. Cells were resuspended in HBSS (without Ca^{2+} and Mg^{2+}) at a 1:5 dilution of the original volume and incubated with 5 μM CM-DiI (Invitrogen). Washed, labelled RBCs were resuspended at a 75% haematocrit and 200 μl were intravenously injected into mice. For platelet labelling, platelet-rich plasma was obtained by collecting whole blood into an Eppendorf tube containing 20 IU heparin, adding 500 μl modified Tyrode's buffer (137 mM NaCl, 0.3 mm Na_2HPO_4, 2 mM KCl, 12 mM $NaHCO_3$, 5 mM HEPES, 5 mM glucose, pH 7.3) containing 0.35% BSA and centrifuged for 8 min at 100*g* at room temperature. Platelets were collected by centrifuging for 10 min at 1000*g*. Washed platelets were incubated with 5 μM CellTracker Green (Invitrogen) for 30 min at 37 °C. Labelled platelets were washed in HBSS (without Ca^{2+} and Mg^{2+}) and resuspended in HBSS (without Ca^{2+} and Mg^{2+}) before intravenous injection. Efficiency of labelling was confirmed using a FACSCalibur (Becton Dickinson) and analysed using CellQuest Pro (Becton Dickinson).

Vascular permeability analysis. To determine the vascular permeability in MLNs, 250 μg of lysine-fixable FITC-dextran (2,000 kDa, Invitrogen) in 40 μl total volume physiologic saline was injected immediately before killing animals. Frozen sections of MLNs were stained with anti-CD31 or PNAd antibodies and analysed for the presence of dextran outside the blood vessels. Alternatively, 500 μg of lysine-fixable FITC–dextran (2,000 kDa, Invitrogen) or fluorescently labelled RBCs resuspended at a 75% haematocrit were intravenously injected into mice 1 min or 10 min, respectively, before the mice were killed. Frozen sections of LN were stained with an anti-CD31 antibody and analysed for the presence of fluorescently labelled RBCs outside the blood vessels. Evans blue dye (2% in saline) was intravenously injected 5 min before killing the mice. Mice were perfused with heparanized saline and LNs were placed in formamide (Fisher Scientific) overnight at 55 °C. Tissues were removed from formamide and absorbance was measured at 620 nm in a spectrophotometer[33].

Transmission electron microscopy (TEM). LNs were removed and fixed overnight at 4 °C in 2% gluteraldehyde in 0.1 M cacodylate buffer (pH 7.3). Tissues were washed in 0.1 M cacodylate buffer and post-fixed in 1% osmium tetroxide (1.5 h) and 1% tannic acid (1 h). Tissues were dehydrated through graded alcohols and embedded in EPON resin (EMS). Semi-thin sections (~300-nm) were cut using an ultramicrotome (RMC 7000, RMC) equipped with a diamond knife and stained with toluidine blue, some of which were imaged. Ultrathin sections (~70-nm) were stained with uranyl acetate and lead citrate and visualized using a

Hitachi H-7600 electron microscope with a 4 megapixel digital monochrome camera and AMT-EM image acquisition software (Advanced Microscopy Techniques).

Platelet transfusion experiments. Platelets were isolated as described above and were intravenously injected into *Clec-2*$^{-/-}$BM chimaeras once a day for 4 consecutive days. For S1P^{less} mice, wild-type or *Clec-2*$^{-/-}$ platelets were isolated as before and resuspended in HBSS without Ca^{2+} and Mg^{2+}. Either wild-type or *Clec-2*$^{-/-}$ platelets were administered once a day for 4 consecutive days. One minute before euthanizing, mice were injected (i.v.) with 500 μg of FITC-Dextran (2,000 kDa, Invitrogen) in 100 μl HBSS without Ca^{2+} and Mg^{2+}.

Immune challenge. Three- to four-week-old mice received subcutaneous injections of 250 μg ovalbumin/complete Freunds' adjuvant (OVA/CFA, Sigma-Aldrich) into the hindlimb[34]. Control lateral hindlimbs were injected with an equivalent volume of saline as a control. Mice were killed 1 week after challenge and popliteal LNs were processed for immunofluorescence staining.

L-selectin blockade and *in vitro* leukocyte rolling assay. Three- to four-week-old mice were challenged with 250 μg OVA/CFA into the hindlimb. Mice were then administered i.p. 30 μg of an anti-mouse L-selectin antibody (Mel-14)[20] every other day for 1 week. Mice were killed and popliteal LNs were dissected and analysed.

To determine the blocking efficiency of Mel-14, an *in vitro* lymphocyte rolling assay was used. Briefly, 100 μg ml^{-1} streptavidin was coated on 35-mm polystyrene dishes overnight at 4 °C. After washing three times with HBSS, the dishes were blocked with 1% human serum albumin in HBSS for 2 h at 4 °C and then incubated with biotin conjugated 6-sulfo-sLe^x, a ligand for L-selectin, for 2 h at 4 °C[35]. Isolation of peripheral leukocytes was described previously[36]. Heparinized blood was obtained from wild-type or *Pdpn*$^{f/f}$;*CagCre* mice that were treated with Mel-14 or rat IgG. After lysis of red blood cells, leukocytes were centrifuged at 100*g* for 10 min. Leukocytes (0.1×10^6 per ml) were perfused over 6-sulfo-sLe^x coated surface under shear stress at 1.0 dyn per cm^2. The number of rolling cells was counted using the Element software (Nikon).

***In vivo* blockade of VE-cad and integrin αIIbβIIIa.** After hindlimb challenge with OVA/CFA, mice were i.p. injected with 30 μg every other day of a monoclonal antibody against VE-cad (clone 11D4.1, BD Pharmingen) or against activated integrin αIIbβIIIa (clone JON/A)[37]. One week after immune challenge, Evans blue dye permeability assays were performed as described above.

Peripheral lymphocyte counts. For measuring peripheral lymphocyte counts, whole blood obtained in EDTA-coated tubes was used to obtain a complete blood count using a Hemavet.

Development of PDPN$^+$ and PDPN$^-$ cell line. Parental murine melanoma cell line B16-F0, which contains PDPN$^+$ and PDPN$^-$ cells, was purchased from American Type Culture Collection (ATCC). Cells were cultured in DMEM containing 10% heat-inactivated FBS and 1% L-glutamine/penicillin/streptomycin (Cellgro) at 37°C in a humidified atmosphere of 5% CO_2. B16-F0 cells were stained with anti-murine PDPN monoclonal antibody (clone 8.1.1) and a DyLight488-conjugated secondary antibody. Stained cells were then sorted with a FACSAria III cell sorter (BD Biosciences) to generate the PDPN$^+$ and PDPN$^-$ B16-F0 melanoma cells.

Analysis of S1P concentration. Platelets were isolated from wild-type or *Clec-2*-deficient mice as before. After being washed with modified Tyrode's buffer containing 0.35% fatty acid-free BSA (Sigma-Aldrich), platelets (1×10^8 in 100 μl) were then incubated with a monoclonal antibody against CLEC-2 (clone INU1 at 10 μg ml^{-1}) for 10 min at room temperature. The supernatant was collected and S1P concentration was determined using an S1P ELISA kit (Echelon Biosciences) according to the manufacturer's protocol. Alternatively, platelets were incubated with 5×10^5 PDPN$^+$ or PDPN$^-$ cells for 10 min at room temperature. Then the supernatant was collected and S1P concentration was quantitated by electrospray ionisation tandem mass spectrometry (ESI-MS/MS) using stable isotope dilution with heptadeuterated S1P (Avanti Polar Lipids) as the internal standard as described previously[38].

***Ex vivo* LN slice culture.** MLNs were dissected from wild-type and *Pdpn*-deficient mice and embedded in low melting point agarose (Invitrogen) at 37°C. Blocks were cooled on ice and sectioned (~250 μm) using a Vibratome (McIlwain tissue chopper). LN slices were then co-incubated with washed platelets (2×10^7 in 100 μl of DMEM containing 10% charcoal stripped FBS; Invitrogen), isolated from wild-type, *Clec-2*$^{-/-}$, or S1P^{less} mice, for 1.5 h at 37 °C with gentle shaking. In some experiments, normal FBS or the S1PR1 antagonist, W146 (5 μM, Cayman Chemical), was included. MLN slices were then fixed in 4% PFA in PBS for 30 min and processed for cryopreservation. Cryosections were stained with antibodies against CD31 (clone 2H8, Abcam) and VE-cad (clone 11D4.1, BD Pharmingen).

Quantification of *Pdpn* deletion, dextran leakage or VE-cad intensity. PDPN levels were calculated using ImageJ software to compare the mean intensities between the groups from a minimum of 5 low magnification (×10) images. For quantification of FITC–dextran leakage, six cryosections were cut per mouse and each section was analysed for the total number of HEVs (~50 per section of MLNs) and non-HEVs based on CD31 or PNAd staining and morphology. The number of vessels that exhibited dextran outside the vessel was determined and the percentage of 'leaky' vessels compared to the total number of that vessel type was determined. Similarly, the percentage of VE-cad expression on vessels from MLN cryosections was determined by comparing the number of HEV and non-HEV that were VE-cad positive to the total number of vessel type. Alternatively, VE-cad intensities on HEVs were determined using ImageJ analysis software.

Statistical analysis. Statistical tests were performed using Prism software (GraphPad). Two-sided, Student's *t*-tests were performed after the data were confirmed to fulfil the criteria of normal distribution and equal variance. Differences were considered statistically significant when $P < 0.05$.

29. Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244,** 305–318 (2002).
30. Foo, S. S. *et al.* Ephrin-B2 controls cell motility and adhesion during blood-vessel-wall assembly. *Cell* **124,** 161–173 (2006).
31. Kisanuki, Y. Y. *et al.* Tie2-Cre transgenic mice: a new model for endothelial cell-lineage analysis *in vivo*. *Dev. Biol.* **230,** 230–242 (2001).
32. Mombaerts, P. *et al.* RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68,** 869–877 (1992).
33. Han, E. D. *et al.* Increased vascular permeability in C1 inhibitor-deficient mice mediated by the bradykinin type 2 receptor. *J. Clin. Invest.* **109,** 1057–1063 (2002).
34. Kamala, T. Hock immunization: a humane alternative to mouse footpad injections. *J. Immunol. Methods* **328,** 204–214 (2007).
35. Lou, J. *et al.* Flow-enhanced adhesion regulated by a selectin interdomain hinge. *J. Cell Biol.* **174,** 1107–1117 (2006).
36. Xia, L. *et al.* P-selectin glycoprotein ligand-1-deficient mice have impaired leukocyte tethering to E-selectin under flow. *J. Clin. Invest.* **109,** 939–950 (2002).
37. Bergmeier, W. *et al.* Flow cytometric detection of activated mouse integrin αIIbβ3 with a novel monoclonal antibody. *Cytometry* **48,** 80–86 (2002).
38. Selim, S. *et al.* Plasma levels of sphingosine 1-phosphate are strongly correlated with haematocrit, but variably restored by red blood cell transfusions. *Clin. Sci.* **121,** 565–572 (2011).

doi:10.1038/nature12507

Cyclin A regulates kinetochore microtubules to promote faithful chromosome segregation

Lilian Kabeche[1,2] & Duane A. Compton[1,2]

The most conspicuous event in the cell cycle is the alignment of chromosomes in metaphase. Chromosome alignment fosters faithful segregation through the formation of bi-oriented attachments of kinetochores to spindle microtubules. Notably, numerous kinetochore–microtubule (k–MT) attachment errors are present in early mitosis (prometaphase)[1], and the persistence of those errors is the leading cause of chromosome mis-segregation in aneuploid human tumour cells that continually mis-segregate whole chromosomes and display chromosomal instability[2–7]. How robust error correction is achieved in prometaphase to ensure error-free mitosis remains unknown. Here we show that k–MT attachments in prometaphase cells are considerably less stable than in metaphase cells. The switch to more stable k–MT attachments in metaphase requires the proteasome-dependent destruction of cyclin A in prometaphase. Persistent cyclin A expression prevents k–MT stabilization even in cells with aligned chromosomes. By contrast, k–MTs are prematurely stabilized in cyclin-A-deficient cells. Consequently, cells lacking cyclin A display higher rates of chromosome mis-segregation. Thus, the stability of k–MT attachments increases decisively in a coordinated fashion among all chromosomes as cells transit from prometaphase to metaphase. Cyclin A creates a cellular environment that promotes microtubule detachment from kinetochores in prometaphase to ensure efficient error correction and faithful chromosome segregation.

The correction of k–MT attachment errors relies on the detachment of microtubules from kinetochores[8], and current models for k–MT regulation involve either chromosome-autonomous[9] or chromosome-coordinated processes (Extended Data Fig. 1). We measured k–MT attachment stability using fluorescence dissipation after photoactivation in three vertebrate cell lines (Fig. 1 and Extended Data Fig. 2). Cells were defined as prometaphase and metaphase on the basis of chromosome alignment using differential interference contrast (DIC) optics. In each cell line the average stability of the stable MT population (for example, k–MTs) in prometaphase was significantly less than in metaphase ($P \leq 0.01$, Students *t*-test; Fig. 1a–c), and the k–MT half-lives in prometaphase and metaphase cells distributed into non-overlapping populations. The difference cannot be accounted for by differences in the initial intensity of green fluorescent protein (GFP) fluorescence after photoactivation (Extended Data Fig. 3a), the fraction of microtubules in the slowly decaying population (Extended Data Fig. 3b), or poleward microtubule flux (Extended Data Fig. 4). Notably, fluorescence decay of the activated region in both metaphase and prometaphase cells fits to a double exponential curve ($R^2 > 0.99$), indicating that only two populations of microtubules are identified by this method: non-k–MTs and k–MTs[10].

To test whether k–MT attachments become progressively stabilized during prometaphase we measured k–MT stability serially in the same

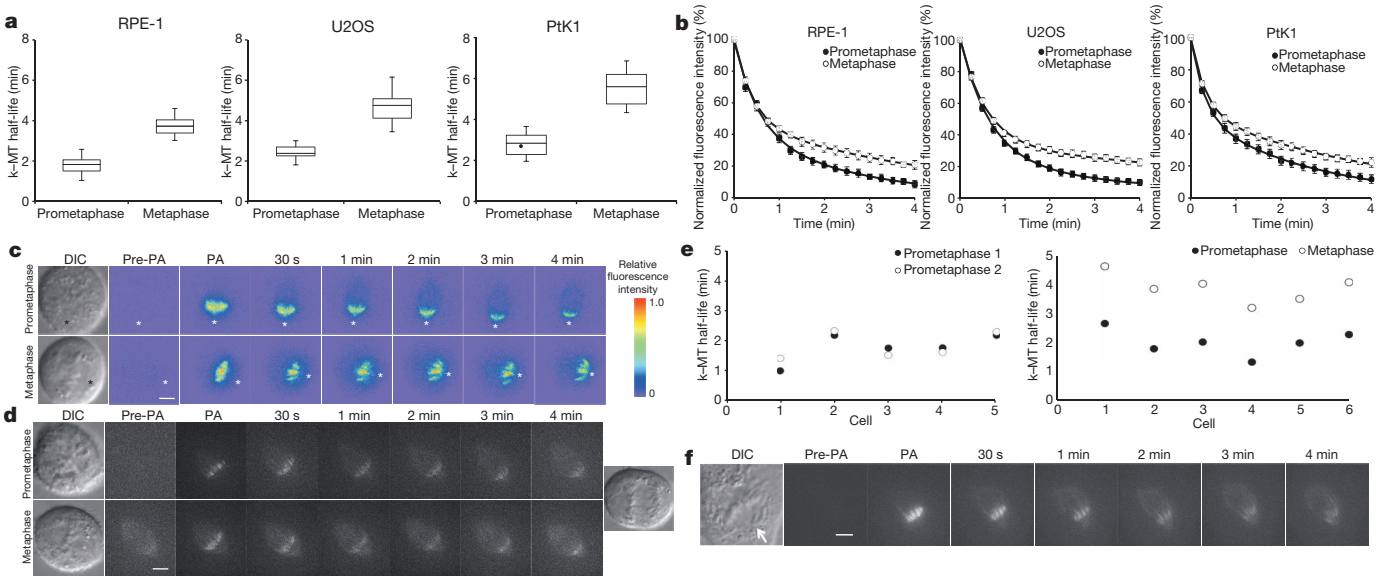


Figure 1 | The stability of k–MT attachments in prometaphase and metaphase. a, Box (s.d. of the mean) and whisker (range of data points) plots of k–MT half-lives of RPE-1, U2OS and PtK1 cells expressing photoactivatable GFP–tubulin in prometaphase and metaphase calculated from the fluorescence intensity decay curves ($R^2 > 0.99$); $n = 40$ cells for RPE-1 and U2OS, and 20 cells for PtK1 per condition. Black circle represents the cell in **f**. **b**, Normalized fluorescence intensity of prometaphase (filled circles) and metaphase (white circles) spindles. **c**, DIC and background-subtracted fluorescence images (pseudo-coloured heat maps) of U2OS cells in prometaphase and metaphase. Asterisks mark spindle poles. Scale bar, 5 μm. PA, photoactivation. **d**, DIC and fluorescence images of an RPE-1 cell in prometaphase and metaphase. Scale bar, 5 μm. **e**, k–MT half-life of individual RPE-1 cells photoactivated serially in prometaphase (left) or in prometaphase and then in metaphase (right). **f**, DIC and fluorescence images of a PtK1 cell in prometaphase. Arrow indicates unaligned chromosome. Scale bar, 5 μm.

[1]Department of Biochemistry, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA. [2]Norris Cotton Cancer Center, Lebanon, New Hampshire 03756, USA.

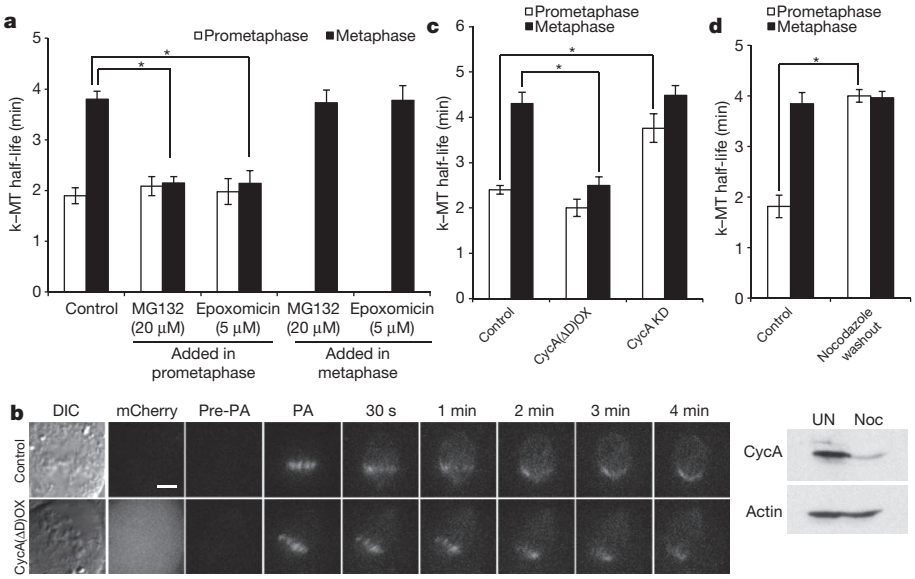


Figure 2 | k–MT stability relies on cyclin A. **a**, k–MT half-life of RPE-1 cells treated with 20 μM MG132 or 5 μM epoxomicin in prometaphase and metaphase; $n = 10$ cells per condition. **b**, DIC and fluorescence images of metaphase spindles in untreated (control) and cyclin-A-overexpressing (CycA(ΔD)OX; in which ΔD indicates mutant lacking the destruction box) U2OS cells, with cyclin A(ΔD) visualized by mCherry fluorescence. Scale bar, 5 μm. **c**, k-MT half-life of untreated (control), cyclin-A-overexpressing and cyclin-A-depleted (CycA knockdown (KD)) U2OS cells; $n = 13$ cells for control, $n = 10$ cells for CycA(ΔD)OX and CycA KD per condition. **d**, Top, k–MT half-life of RPE-1 cells untreated (control) or released from 12 h nocodazole treatment (nocodazole washout); $n = 10$ cells per condition. Bottom, cyclin A and actin immunoblot of untreated (UN, control) or nocodazole-arrested (Noc) cells. Graphs show mean ± s.e.m. $^*P \leq 0.01$, two-tailed t-test.

cell (Fig. 1d, e). Repeated photoactivation did not compromise cell viability as judged by successful progression to anaphase (Fig. 1d and Extended Data Fig. 5a). Photoactivation of RPE-1 cells twice in prometaphase yielded equivalent k–MT half-lives for each trial in each cell. By contrast, k–MT stability increased sharply between prometaphase and metaphase when measured in the same cell (Fig. 1e). The switch in k–MT stability was notably consistent at 1.9 ± 0.2 min. Similar results were obtained in U2OS cells (Extended Data Fig. 5b). The percentage of microtubules in the slowly decaying fraction did not change at different times in prometaphase cells (Extended Data Fig. 5c) as would be predicted by the chromosome-autonomous model (Extended Data Fig. 1). We also photoactivated the spindle microtubules of aligned

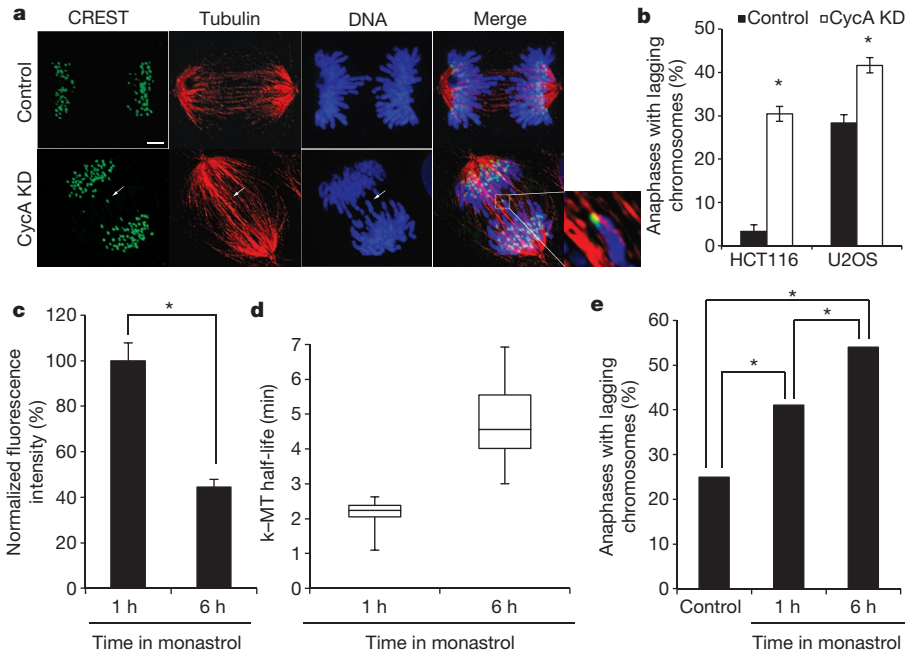


Figure 3 | Cyclin A deficiency increases chromosome mis-segregation. **a**, Anaphase spindles of untreated (control) or cyclin-A-depleted U2OS cells. White arrow highlights merotelic kinetochore. CREST sera is used to mark centromeres. Scale bar, 5 μm. **b**, Percentage of anaphase cells with lagging chromosomes; $n = 300$ cells per condition from three independent experiments. **c**, Fluorescence intensities of U2OS cells stained for cyclin A; $n = 100$ cells per condition from three independent experiments. **d**, Box and whisker plot of k–MT half-lives of U2OS cells incubated in monastrol for 1 h and 6 h; $n = 10$ cells per condition. **e**, Percentage of anaphase cells with lagging chromosomes that were untreated (control), or after recovery from monastrol incubation for 1 h or 6 h; $n = 100$ cells for 1 h and 123 cells for 6 h. Graphs show mean ± s.e.m. $^*P \leq 0.01$, two-tailed t-test.

chromosomes in a prometaphase PtK1 cell containing one unaligned chromosome (Fig. 1f). The half-life of k–MTs on these aligned chromosomes in this cell was 2.5 min (single data point identified in Fig. 1a) and within the population of other prometaphase cells. These data demonstrate a coordinated switch in k–MT attachment stability between prometaphase and metaphase cells (Extended Data Fig. 1).

Next, we tested whether the switch in k–MT attachment stability relies on protein turnover (Fig. 2a). Proteasome inhibition did not alter k–MT attachment stability during prometaphase. However, when cells transited from prometaphase to metaphase in the presence of the inhibitors the switch to stable k–MT attachments was prevented (Fig. 2a). Proteasome inhibition had no effect if the inhibitors were added after chromosome alignment in metaphase. Cells failed to progress to anaphase in all these conditions, verifying the effective inhibition of the proteasome. Thus, the proteasome-dependent destruction of protein substrates during prometaphase is required for the coordinated switch in k–MT stability in metaphase.

Cyclin A is degraded in prometaphase[11,12], and we tested whether cyclin A influenced the switch in k–MT attachment stability (Fig. 2b–d). Expression of an mCherry-tagged mutant version of cyclin A that lacks the degradation box was confirmed by immunoblot (Extended Data Fig. 6a) and shown to persist in mitotic cells by fluorescence microscopy (Fig. 2b). Expression of this mutant cyclin A did not change k–MT attachment stability in prometaphase (Fig. 2c), but prevented the switch to stable attachments in metaphase (Fig. 2b, c and Extended Data Fig. 6c). This mutant version of cyclin A did not impair chromosome bi-orientation as judged by the displacement of the BUB1B (also known as BUBR1) from kinetochores[13], the recruitment of astrin to kinetochores[14] or the relative inter-kinetochore distance (Extended Data Fig. 7). The quantity and activity of aurora B kinase[15,16] was slightly increased in cells overexpressing this mutant of cyclin A during metaphase (Extended Data Fig. 7). There was no apparent linear correlation between k–MT attachment stability and expression level of this protein (Extended Data Fig. 6b), indicating that the persistent expression of cyclin A is sufficient to exceed a threshold that prevents a switch from unstable to stable k–MT attachments.

Vertebrate cells can enter mitosis in the absence of cyclin A[17,18], and we depleted cyclin A expression using RNA interference (Extended Data Fig. 6a). Prometaphase cells lacking cyclin A displayed k–MT attachment stability equivalent to untreated metaphase cells (Fig. 2c and Extended Data Fig. 6c). k–MT attachment stability in metaphase cells lacking cyclin A was not different from control cells. There was a slight decrease in the quantity and activity of aurora B kinase during prometaphase (Extended Data Fig. 7). We also prolonged mitosis with nocodazole to allow the destruction of endogenous cyclin A (Fig. 2d), as cyclin A destruction is not prevented by the spindle assembly checkpoint[19]. k–MT attachments are significantly more stable in prometaphase cells recovering from nocodazole treatment than untreated control cells (Fig. 2d). Nocodazole recovery did not alter k–MT attachment stability once cells reached metaphase. Manipulation of cyclin A (either depletion or expression of the non-degradable mutant) had no effect on the stability of non-k–MT (Extended Data Fig. 8a) or MT stability in cells lacking the NUF2 protein that lack k–MT (Extended Data Fig. 8b, c) showing that cyclin A specifically influences k–MT. Poleward microtubule flux did not account for the difference in k–MT stability between control cells and cells overexpressing non-degradable cyclin A (Extended Data Fig. 9). Thus, cyclin A destabilizes k–MT attachments to regulate the switch from unstable k–MT in prometaphase to stable k–MT in metaphase.

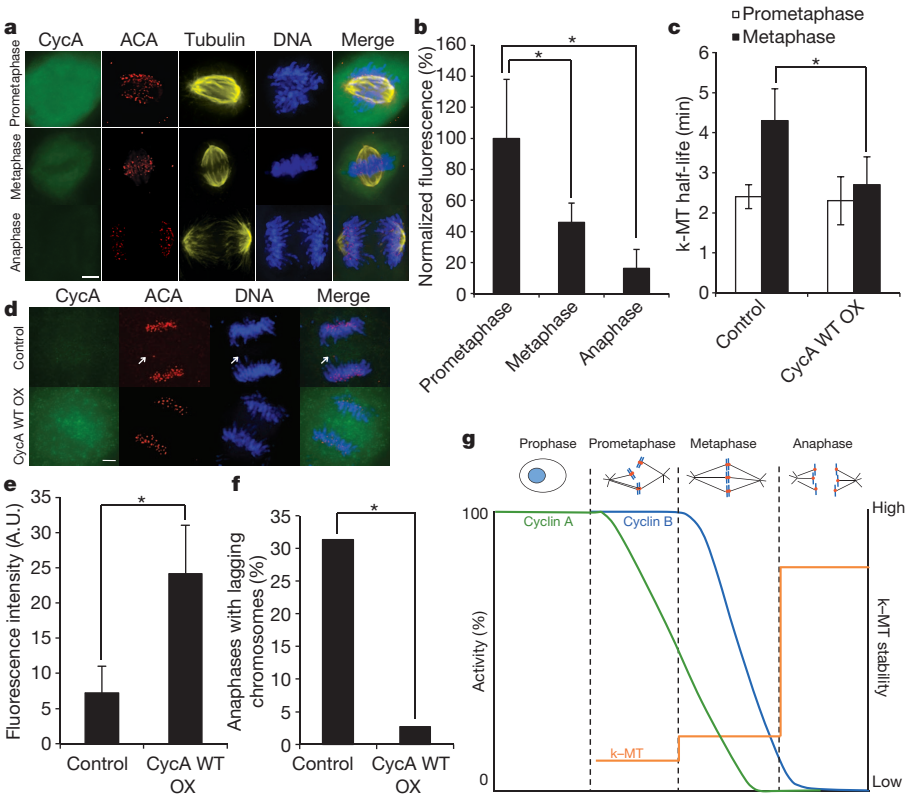


Figure 4 | Cyclin A promotes faithful chromosome segregation. **a**, Immunofluorescence of endogenous cyclin A, tubulin, centromeres (ACA) and DNA of U2OS cells in prometaphase, metaphase and anaphase. Scale bar, 5 μm. **b**, Fluorescence intensities of U2OS cells stained for cyclin A; $n = 150$ cells per condition from three independent experiments. **c**, k–MT half-life of untreated (control) and wild-type cyclin-A-overexpressing (CycA WT OX) U2OS cells; $n = 10$ cells per condition. **d**, Immunofluorescence of endogenous cyclin A, tubulin, centromeres and DNA of control or wild-type-cyclin-A-overexpressing U2OS cells. Scale bar, 5 μm. **e**, Fluorescence intensities of cyclin A of control U2OS cells or cells expressing wild-type cyclin A. **f**, Number of anaphases in U2OS cells with lagging chromosomes; $n = 150$ for control cells and 37 for wild-type cyclin A cells. **g**, Cells enter prometaphase with high cyclin A and cyclin B. The proteosome-dependent reduction of cyclin A levels below a critical threshold induces a coordinated increase in k–MT attachment stability at the prometaphase to metaphase transition. Graphs show mean ± s.e.m. $^*P \leq 0.01$, two-tailed t-test.

There is a direct relationship between k–MT attachment stability and chromosome segregation fidelity[2], and cyclin A deficiency led to a significant increase in the fraction of cells displaying lagging chromosomes in anaphase (Fig. 3b). The lagging chromosomes are caused by persistent merotelic kinetochore attachments, as judged by k–MT attachments oriented towards both spindle poles (Fig. 3a). These segregation errors were not caused by alterations in chromosome compaction (Extended Data Fig. 10a, b) or DNA double-strand breaks (Extended Data Fig. 10c). Furthermore, cells arrested in mitosis with monastrol for 6 h display a 60% reduction in cyclin A levels and have significantly more stable k–MT attachments relative to cells arrested for only 1 h (Fig. 3c, d). Accordingly, cells recovering from 6 h of monastrol treatment displayed significantly higher percentages of anaphase cells with lagging chromosomes compared to cells recovering from only 1 h of monastrol treatment (Fig. 3e). Thus, by destabilizing k–MTs cyclin A promotes faithful chromosome segregation, although disruption of the canonical functions of cyclin A during S phase (via short interfering RNA or protein overexpression) could also contribute to the observed chromosome mis-segregation.

Quantitative fluorescence imaging indicates that endogenous cyclin A levels drop to 40% and 20% in metaphase and anaphase, respectively, relative to prometaphase cells (Fig. 4a, b). Thus, the cyclin A level in the prometaphase cell shown in Fig. 1f has yet to dip below that threshold. Expression of wild-type cyclin A tagged with mCherry destabilizes k–MT attachments in metaphase akin to the non-degradable mutant cyclin A (Fig. 4c), but it does not prohibit anaphase entry[12] and we observe anaphase cells with cyclin A levels approximately fourfold higher than untreated cells in anaphase (Fig. 4e). Consequently, chromosome segregation fidelity is increased in cells expressing wild-type cyclin A (Fig. 4f), consistent with previous data showing that the destabilization of k–MT restores faithful chromosome segregation to chromosomally unstable cancer cells[2].

The stability of k–MT attachments must fall within a narrow permissible range to both satisfy the spindle assembly checkpoint and promote faithful chromosome segregation[2,20,21]. Our data show that initial k–MT attachments in prometaphase are unstable, yet sufficiently robust to promote chromosome alignment. By maintaining unstable k–MT attachments on bi-oriented chromosomes in prometaphase the system exploits the back-to-back geometry of sister kinetochores[22,23] to create optimal conditions for error correction needed to promote faithful chromosome segregation. The coordinated and decisive switch in k–MT stability that we show at the prometaphase to metaphase transition resembles the decisive switch in k–MT stability described previously during the metaphase to anaphase transition[10] (Fig. 4g). These switches in k–MT attachment stability occur as cyclin proteins drop below threshold levels, with the prometaphase to metaphase transition being regulated by cyclin A and the metaphase to anaphase transition being regulated by cyclin B[24,25]. Unlike cyclin B, there is no established feedback mechanism to prevent cyclin A destruction in the absence of k–MT attachment[19]. This indicates that cyclin A functions as a timer in prometaphase to ensure efficient error correction, consistent with previous data showing a linear relationship between the duration of prometaphase and the level of cyclin A[12]. These data define prometaphase and metaphase as biochemically distinct cellular states and show that the prometaphase to metaphase transition is a decisive, unidirectional biochemical event like other phase transitions in the cell cycle[26–28].

METHODS SUMMARY

k–MT attachment stability was measured in cultured cells expressing α-tubulin tagged with photoactivatable GFP. DIC microscopy was used to identify mitotic cells and fluorescent images were generated and acquired using Quorum WaveFX-X1 spinning disk confocal system (Quorum Technologies) equipped with Mosaic digital mirror for photoactivation (Andor Technology) and Hamamatsu ImageEM camera. Fluorescence intensities were normalized to the first time point after photoactivation for each cell following background subtraction and correction for photobleaching.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 April; accepted 26 July 2013.
Published online 8 September 2013.

1. Cimini, D., Moree, B., Canman, J. C. & Salmon, E. D. Merotelic kinetochore orientation occurs frequently during early mitosis in mammalian tissue cells and error correction is achieved by two different mechanisms. *J. Cell Sci.* **116,** 4213–4225 (2003).
2. Bakhoum, S. F., Thompson, S. L., Manning, A. & Compton, D. A. Genome stability is ensured by temporal control of kinetochore–microtubule dynamics. *Nature Cell Biol.* **11,** 27–35 (2009).
3. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilitites in human cancers. *Nature* **396,** 643–649 (1998).
4. Thompson, S. L. & Compton, D. A. Examining the link between chromosomal instability and aneuploidy in human cells. *J. Cell Biol.* **180,** 665–672 (2008).
5. Cimini, D. *et al.* Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J. Cell Biol.* **153,** 517–528 (2001).
6. Bakhoum, S. F., Genovese, G. & Compton, D. A. Deviant kinetochore microtubule dynamics underlie chromosomal instability. *Curr. Biol.* **19,** 1937–1942 (2009).
7. Thompson, S. L. & Compton, D. A. Chromosome missegregation in human cells arises through specific types of kinetochore–microtubule attachment errors. *Proc. Natl Acad. Sci. USA* **108,** 17974–17978 (2011).
8. Liu, D., Vader, G., Vromas, M. J., Lampson, M. A. & Lens, S. M. Sensing chromosome bi-orientation by spatial separation of aurora B kinase from kinetochore substrates. *Science* **323,** 1350–1353 (2009).
9. Nicklas, R. B. & Ward, S. C. Elements of error correction in mitosis: microtubule capture, release, and tension. *J. Cell Biol.* **126,** 1241–1253 (1994).
10. Zhai, Y., Kronebusch, P. J. & Borisy, G. G. Kinetochore microtubule dynamics and the metaphase-anaphase transition. *J. Cell Biol.* **131,** 721–734 (1995).
11. Pines, J. Mitosis: a matter of getting rid of the right protein at the right time. *Trends Cell Biol.* **16,** 55–63 (2006).
12. den Elzen, N. & Pines, J. Cyclin A is destroyed in prometaphase and can delay chromosome alignment and anaphase. *J. Cell Biol.* **153,** 121–136 (2001).
13. Skoufias, D. A., Andreassen, P. R., Lacroix, F. B., Wilson, L. & Margolis, R. L. Mammalian mad2 and bub1/bubr1 recognize distinct spindle-attachment and kinetochore-tension checkpoints. *Proc. Natl Acad. Sci. USA* **98,** 4492–4497 (2001).
14. Manning, A. L. *et al.* CLASP1, astrin and Kif2b form a molecular switch that regulates kinetochore-microtubule dynamics to promote mitotic progression and fidelity. *EMBO J.* **29,** 3531–3543 (2010).
15. Kelly, A. E. & Funabiki, H. Correcting aberrant kinetochore microtubule attachments: an Aurora B-centric view. *Curr. Opin. Cell Biol.* **21,** 51–58 (2009).
16. DeLuca, K. F., Lens, S. M. & DeLuca, J. G. Temporal changes in Hec1 phosphorylation control kinetochore–microtubule attachment stability during mitosis. *J. Cell Sci.* **124,** 622–634 (2011).
17. Gong, D. & Ferrell, J. E. The roles of cyclin A2, B1, and B2 in early and late mitotic events. *Mol. Biol. Cell* **21,** 3149–3161 (2010).
18. Mihaylov, I. S. *et al.* Control of DNA replication and chromosome ploidy by Germinin and Cyclin A. *Mol. Cell. Biol.* **22,** 1868–1880 (2002).
19. Di Fiore, B. & Pines, J. How cyclin A destruction escapes the spindle assembly checkpoint. *J. Cell Biol.* **190,** 501–509 (2010).
20. Musacchio, A. & Salmon, E. D. The spindle-assembly checkpoint in space and time. *Nature Rev. Mol. Cell Biol.* **8,** 379–393 (2007).
21. Bakhoum, S. F. & Compton, D. A. Kinetochore and disease: keeping microtubule dynamics *in check! Curr. Opin. Cell Biol.* **24** (2012).
22. Indjeian, V. B. & Murray, A. W. Budding yeast mitotic chromosomes have an intrinsic bias to biorient on the spindle. *Curr. Biol.* **17,** 1837–1846 (2007).
23. Lončarek, J. *et al.* The centromere geometry essential for keeping mitosis error free is controlled by spindle forces. *Nature* **450,** 745–749 (2007).
24. Holloway, S. L., Glotzer, M., King, R. W. & Myrray, A. W. Anaphase is initiated by proteolysis rather than by the inactivation of maturation-promoting factor. *Cell* **73,** 1393–1402 (1993).
25. Surana, U. *et al.* Destruction of the CDC28/CLB mitotic kinase is not required for the metaphase to anaphase transition in budding yeast. *EMBO J.* **12,** 1969–1978 (1993).
26. Pagliuca, F. W. *et al.* Quantitative proteomics reveals the basis for biochemical specificity of cell-cycle machinery. *Mol. Cell* **43,** 406–417 (2011).
27. Reed, S. I. Ratchets and clocks: the cell cycle, ubiquitylation and protein turnover. *Nature Rev. Mol. Cell Biol.* **4,** 855–864 (2003).
28. Trunnell, N. B., Poon, A. C., Kim, S. Y. & Ferrell, J. E. Ultrasensitivity in the regulation of Cdc25 and Cdk1. *Mol. Cell* **41,** 263–274 (2011).

Acknowledgements This work was supported by National Institutes of Health grants GM51542 (D.A.C.) and GM008704 (L.K.) and the John H. Copenhaver Jr and William H. Thomas, MD 1952 Junior Fellowship (L.K.). We thank J. Pines and J. DeLuca for providing reagents.

Author Contributions L.K. and D.A.C. were responsible for experimental design, data interpretation and writing the manuscript. L.K. conducted the experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.C. (duane.a.compton@dartmouth.edu).

METHODS

Cell culture. RPE-1 (ATCC, CRL-4000), PA (photoactivatable GFP–tubulin-expressing)–RPE-1 (ATCC, CRL-4000), U2OS (ATCC, HTB-96) and PA–U2OS (ATCC, HTB-96) cells, and PA–PtK1 (ATCC, CCL-35) cells were grown in Dulbecco's modified Eagle's medium, (DMEM; Invitrogen) supplemented with 10% FBS (Mediatech), 50 IU ml^{-1} penicillin and 50 μg ml^{-1} streptomycin (Mediatech) at 37 °C in a humidified atmosphere with 5% CO_2. All cell lines are validated as mycoplasma free. Media for cells expressing photoactivatable GFP–tubulin was supplemented with G418 (Mediatech). Cells were incubated with 150 μg ml^{-1} of nocodazole (Millipore) for 12 h and then either released into 5 μM MG132 and analysed through live-cell imaging, or collected through mitotic shake-off and prepared for immunoblots. Cells were incubated with 100 μM monastrol (TOCRIS Bioscience) for 1 h or 6 h and then either analysed through live-cell imaging, prepared for immunofluorescence or released into DMEM for 40 min and prepared for immunofluorescence.

Cell transfection. Plasmid transfections were done with FuGENE 6 (Roche Diagnostics), and cells analysed 12 h later by live-cell imaging, immunofluorescence or preparation for immunoblots. Short interfering RNA transfections were conducted using Oligofectamine (Invitrogen), and cells analysed 48 h later. RNA duplexes for *CCNA2* (5′-CTATGGACATGTCAATTGT-3′) and *NUF2* (5′-GCAUGCCGUGAAACGUAUA-3′) were purchased from Applied Biotechnologies.

Antibodies. Antibodies used for this study were: ACA (anti-centromere antibody) (CREST; Geisel School of Medicine), actin (Seven Hills Bioreagents, LMAB-C4), astrin[13], aurora B kinase (Novus Biologicals, NB100-294), BUBR1 (Abcam, ab4637), HEC1 (Novus Biologicals, NB100-338), phospho (p)-HEC1 (provided by Jennifer DeLuca), cyclin A (Santa Cruz Antibodies, sc-751), tubulin (Sigma-Aldrich, T9026), p-H3 (Cell Signaling Technologies, 3377S) and γ-H2AX (Novus Biologicals, NB100-384). Secondary antibodies were conjugated to fluorescein isothyocyante (FITC) (Jackson ImmunoResearch; anti-mouse, 715-096-151; anti-rabbit, 111-005-003), Texas Red (Jackson ImmunoResearch; anti-mouse, 715-076-020; anti-rabbit, 111-605-045), Cy5 (Invitrogen, anti-human, A-11013) and horseradish peroxidase (Jackson ImmunoResearch; anti-mouse, 715-036-151; anti-rabbit, 211-002-171). Immunoblots were detected using Lumiglow (KPL).

Photoactivation. Images were acquired using Quorum WaveFX-X1 spinning disk confocal system (Quorum Technologies) equipped with Mosaic digital mirror for photoactivation (Andor Technology) and Hamamatsu ImageEM camera. DIC microscopy was used to identify prometaphase and metaphase cells with bipolar spindles. Microtubules were locally activated in one half spindle. Fluorescence images were captured every 15 s for 4 min with a 100× oil-immersion 1.4 numerical aperture objective. For measurement of unstable MTs, fluorescence images were captured every 5 s for 1 min. DIC microscopy was then used to verify that a bipolar spindle was maintained throughout image acquisition and that cells had not entered anaphase.

For double photoactivation experiments, prometaphase cells were identified using DIC. Microtubules were locally activated in one half spindle, and DIC was then used to verify that the cell was still in prometaphase. After which, microtubules were again locally activated in either prometaphase or metaphase (identified through DIC). Using DIC, cells were observed continuing through anaphase to ensure cell survival.

To quantify fluorescence dissipation after photoactivation, pixel intensities were measured within a 1-μm rectangular area surrounding the region of highest fluorescence intensity and background subtracted using an equal area from the non-activated half spindle. The values were corrected for photobleaching by treating cells with 10 μM taxol and determining the percentage of fluorescence loss during 4 min of image acquisition after photoactivation. Fluorescence values were normalized to the first time point after photoactivation for each cell and the average intensity at each time point was fit to a double exponential curve $A1 \times \exp(-k_1 t) + A2 \times \exp(-k_2 t)$ using MatLab (Mathworks), in which t is time, A1 represents the less stable non-kinetochore microtubule population and A2 the more stable kinetochore microtubule population with decay rates of k_1 and k_2, respectively. The turnover half-life for each process was calculated as $\ln 2/k$ for each population of microtubules.

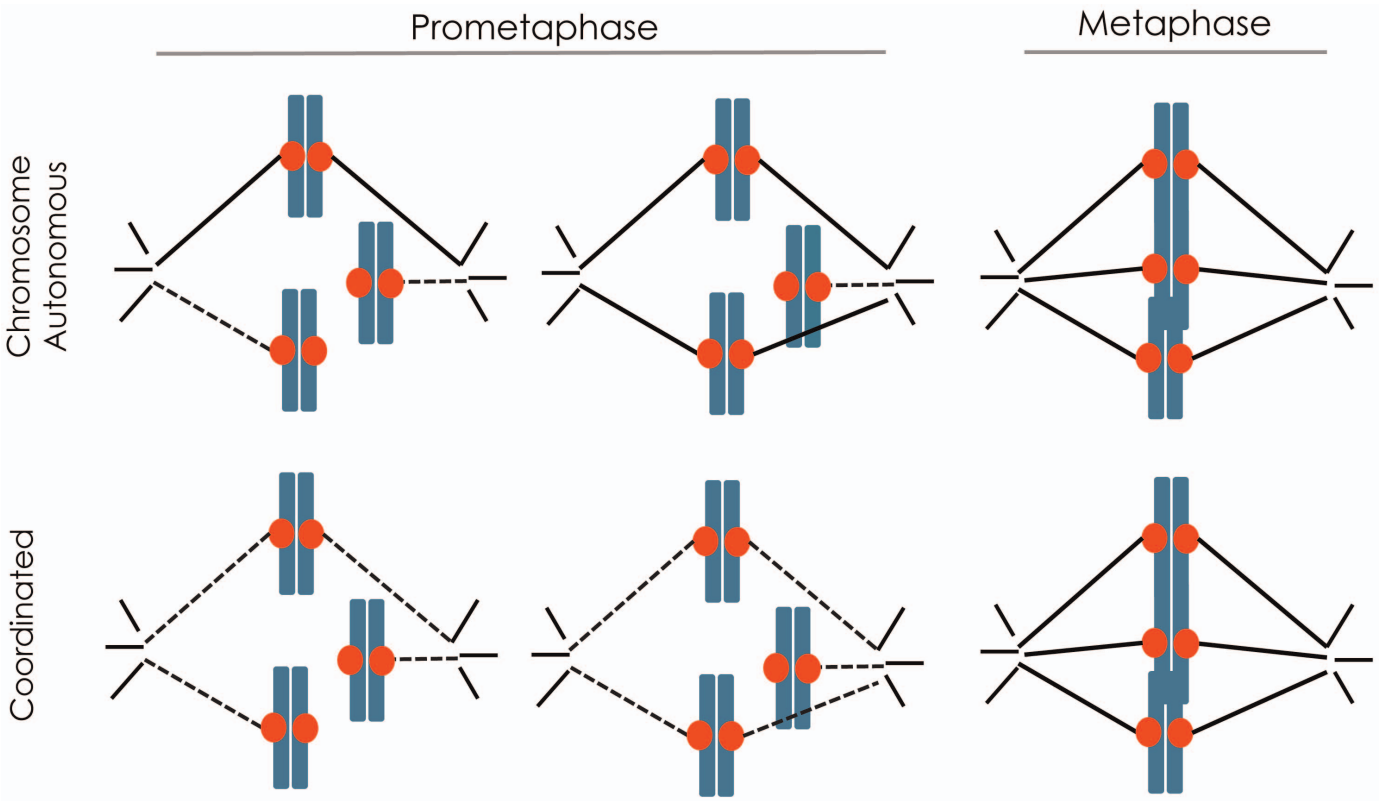
Immunofluorescence microscopy. Cells were fixed with 3.5% paraformaldehyde for 15 min, washed with Tris-buffered saline with 5% bovine serum albumin (TBS-BSA) and 0.5% Triton X-100 for 5 min and TBS-BSA for 5 min. Antibodies were diluted in TBS-BSA plus 0.1% Triton X-100 and coverslips incubated for 12 h at 4 °C. After which, cells were washed with TBS-BSA for 5 min with shaking. Secondary antibodies were diluted in TBS-BSA plus 0.1% Triton X-100 and coverslips incubated for 1 h at room temperature (20–25 °C). For p-HEC1, all wash buffers were supplemented with 80 nM okadaic acid and 40 nM microcystin. Images were acquired with Orca-ER Hamamatsu cooled charge-coupled device camera mounted on an Eclipse TE 2000-E Nikon microscope. 0.2 μm optical sections in the z-axis were collected with a plan Apo 60 ×1.4 numerical aperture oil immersion objective at room temperature. Iterative restoration was performed using Phylum Live software (Improvision). Anaphase chromatids were counted as lagging if they contained centromere staining (using CREST antibody) in the spindle midzone separated from centromeres at the poles. The scoring of lagging chromosomes in anaphase was performed blinded. The investigator was not aware of which sample they were counting until all samples were completed and subsequently unblended.

For quantitative assessments, cells were fixed and stained for aurora B/p-HEC1, CREST and DNA. Pixel intensities for CREST and aurora B/p-HEC1 staining were measured in approximately 15 regions over the entire cell. Background fluorescence was subtracted and the ratio of intensities were calculated and averaged over multiple kinetochores from multiple mitotic cells ($n \geq 10$ cells). To quantify DNA condensation, DAPI fluorescence was measured in approximately five regions over the entire cell. Background fluorescence was subtracted, and the ration was averaged over multiple cells ($n \geq 10$ cells per condition).

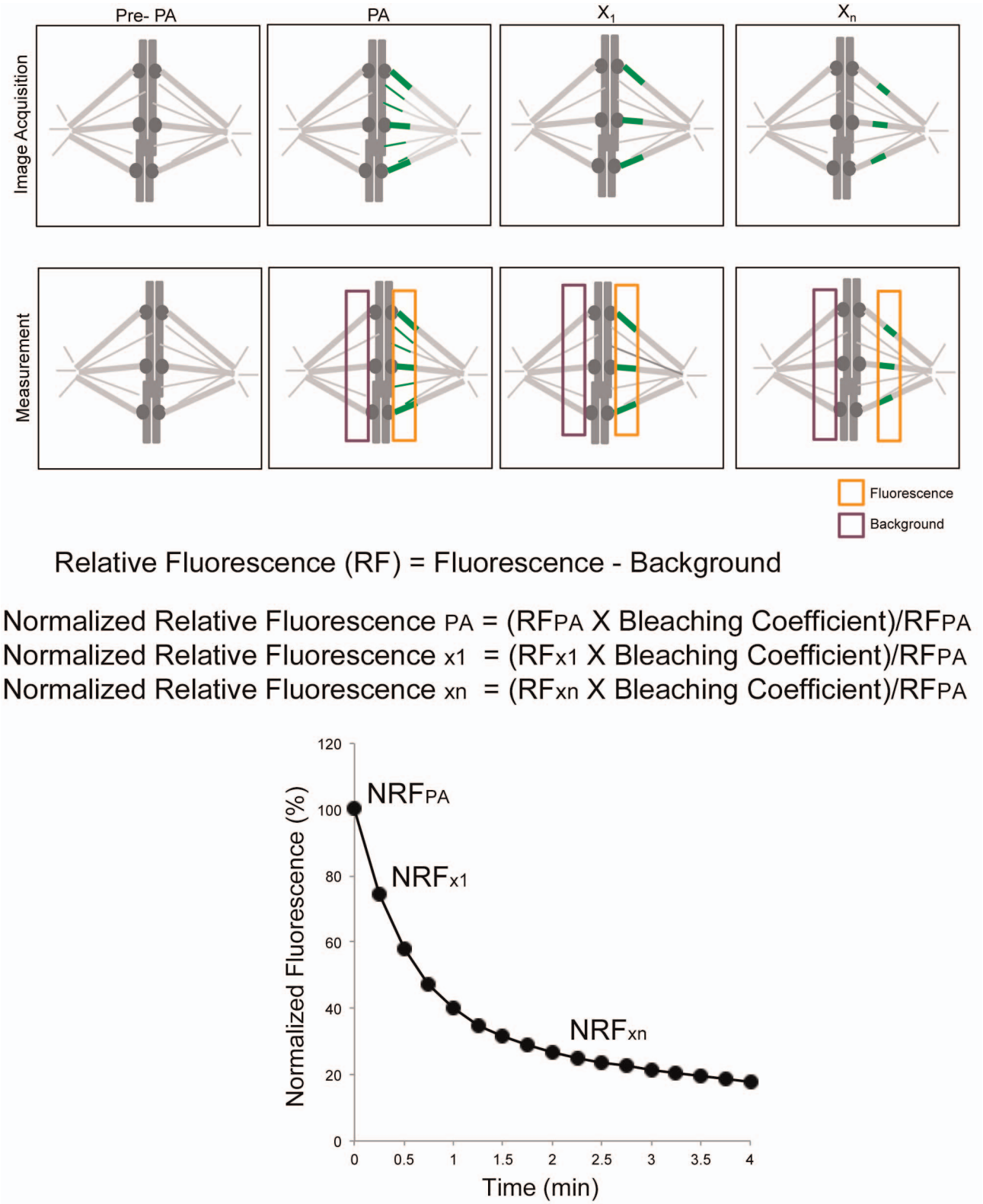
Measurements of intercentromere distances were made with Huygens Essential software. All measurements were performed for three independent experiments. Error bars represent standard errors (s.e.m.). The Student's t-test was used to calculate the significance of differences between samples.

For endogenous cyclin A staining and quantification, Cells were fixed with PBS with 3.5% paraformaldehyde and 2% sucrose for 10 min. After which, cells were permeabilized with ice-cold methanol for 5 min and subsequently washed with 500 mM ammonium chloride in PBS for 20 min twice. Cells were then incubated with PBS with 2% donkey serum for 1 h. After which, cells were incubated with primary and secondary antibodies. Quantification of cyclin A levels were done using ImageJ (National Institutes of Health). The cytoplasm of a G1 cell (identified by the lack of cyclin A in the nucleus) was used to measure fluorescence intensity and was then used for background subtraction.

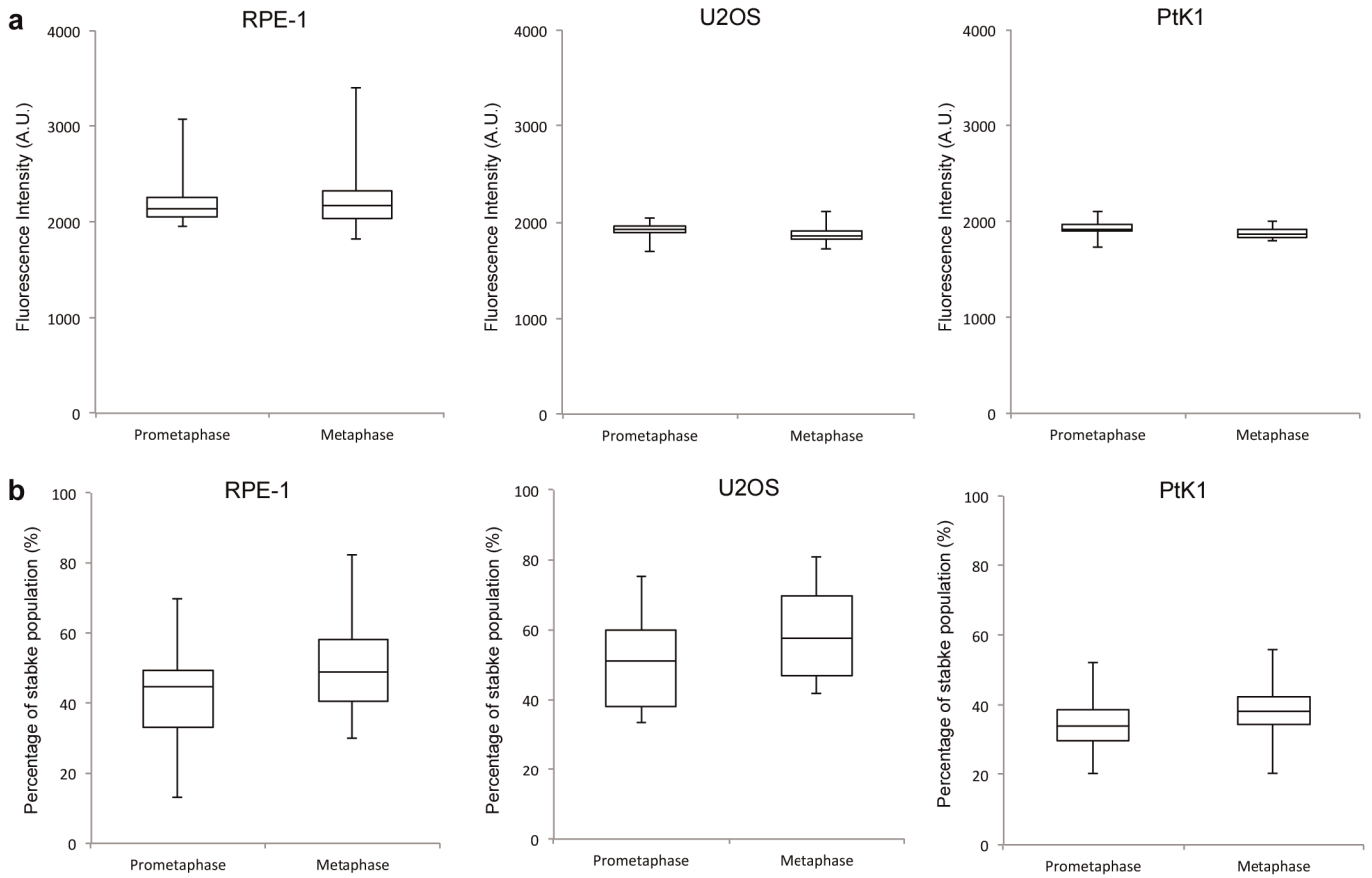
Statistical analysis. For photoactivation, no fewer than ten cells were used for each condition, which is sufficient to detect significant differences when the effect size is twofold or more. For scoring lagging chromosomes and measuring differences in fluorescence intensity, no fewer than 20 cells per condition, which is sufficient to detect significant differences between samples, were used. Data analysis was performed blind. The investigator was unaware of which sample they were counting until all samples were completed and subsequently unblinded. All data had a normal distribution, with similar variance between all conditions tested. Two-tailed t-tests were conducted where indicated in the figure legends.



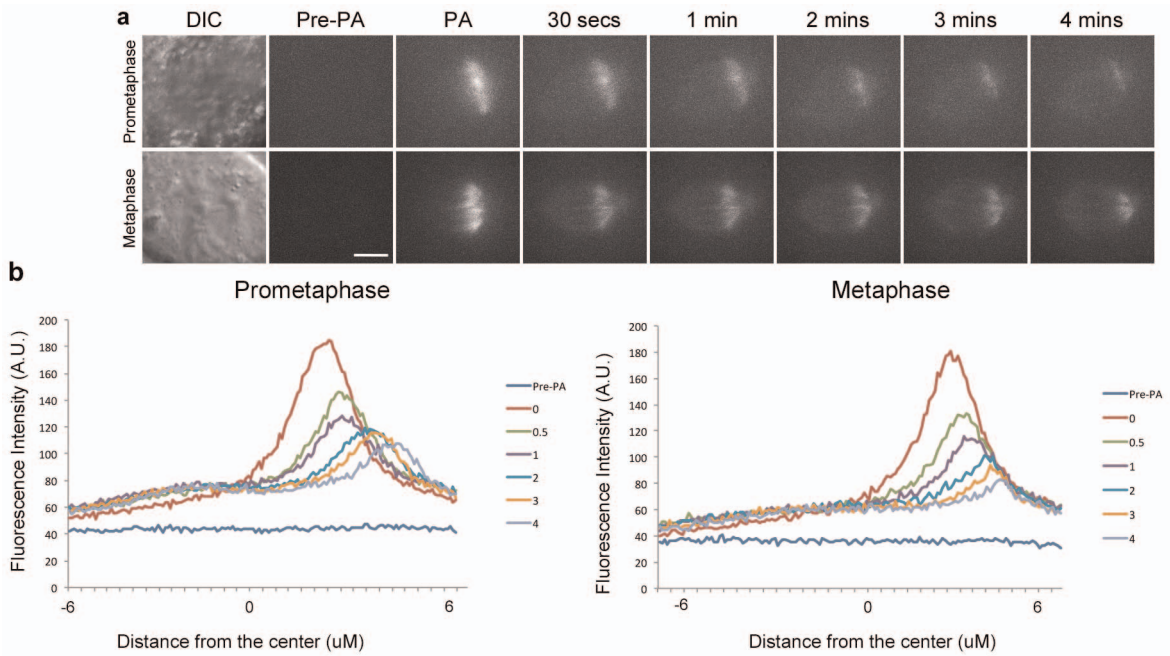
Extended Data Figure 1 | Models of k–MT attachment stability. Unstable (dotted lines) and stable (solid lines) k–MT depict the differences in the chromosome-autonomous and -coordinated models of regulating k–MT attachment stability.



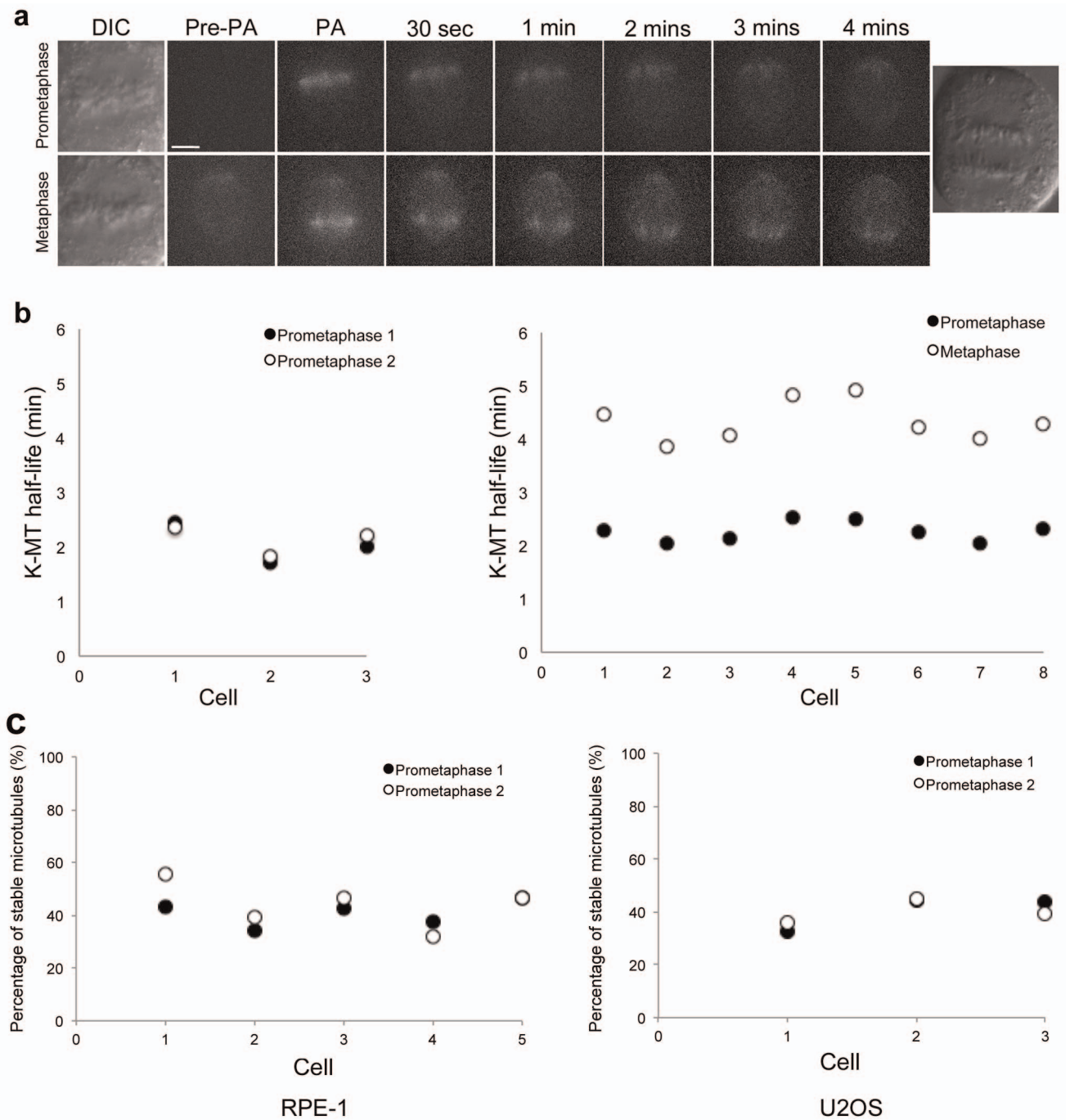
Extended Data Figure 2 | Schematic of methods for quantification of photoactivation. Relative fluorescence (RF) is calculated by subtracting fluorescence of an equal size region on the non-photoactivated half-spindle (background) from the fluorescence intensity of the photoactivated half-spindle (fluorescence). Bleaching coefficient is determined for each time point using taxol-treated cells. Normalized relative fluorescence is then calculated by multiplying the relative fluorescence of individual time points by the bleaching coefficient, divided by the relative fluorescence of the first photoactivated time point (RF_{PA}). The data fits into a double exponential curve.



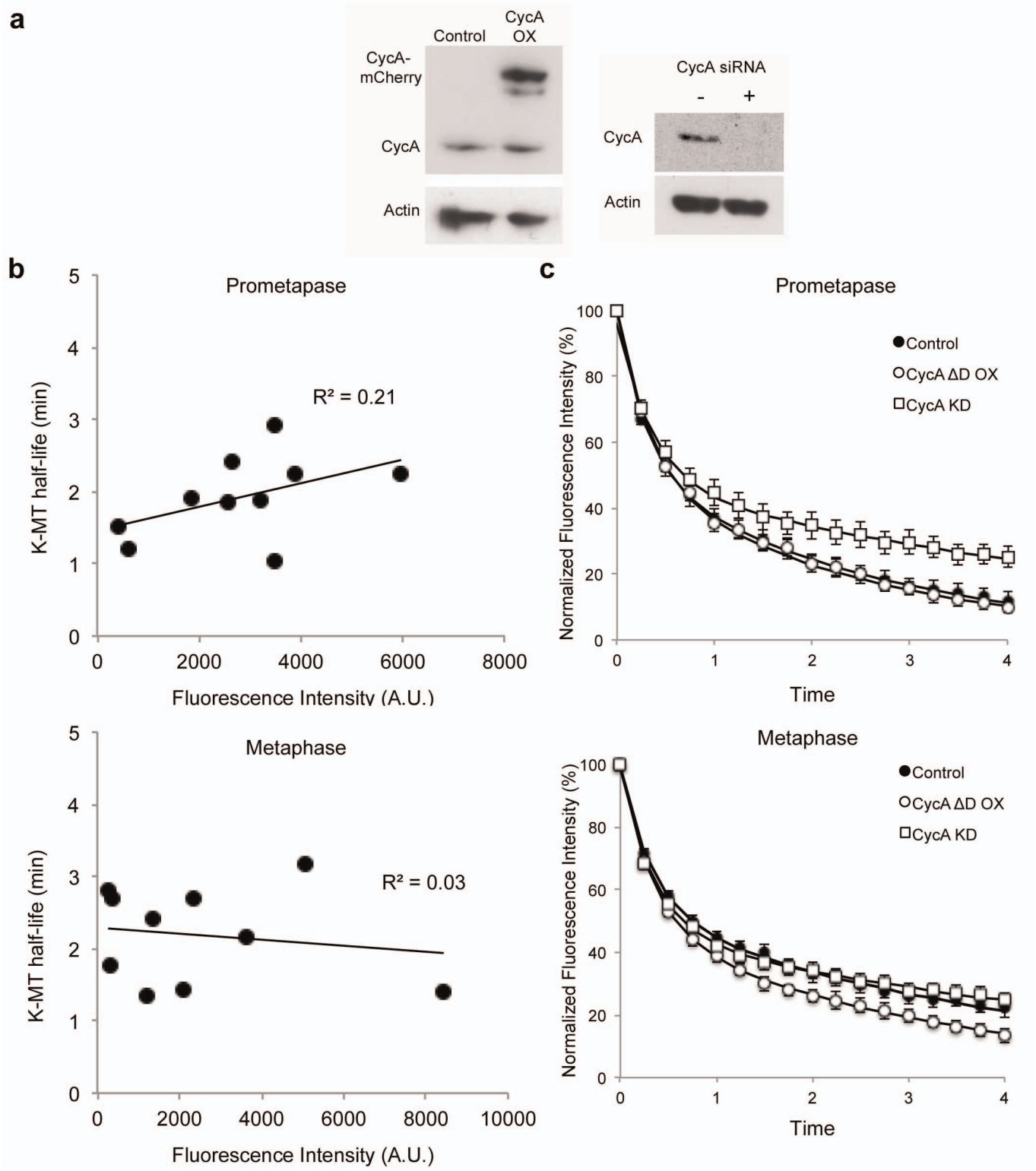
Extended Data Figure 3 | Fluorescence intensity and percentage of MTs in the stable population. **a**, Box and whisker plots of fluorescence intensity of photoactivatable GFP–tubulin after photoactivation. **b**, Percentage of MTs in the stable population (for example, k–MTs) calculated from the exponential decay curve of photoactivated fluorescence ($R^2 > 0.99$); $n = 40$ cells for RPE-1 and U2OS, and 20 cells for PtK1 per condition.



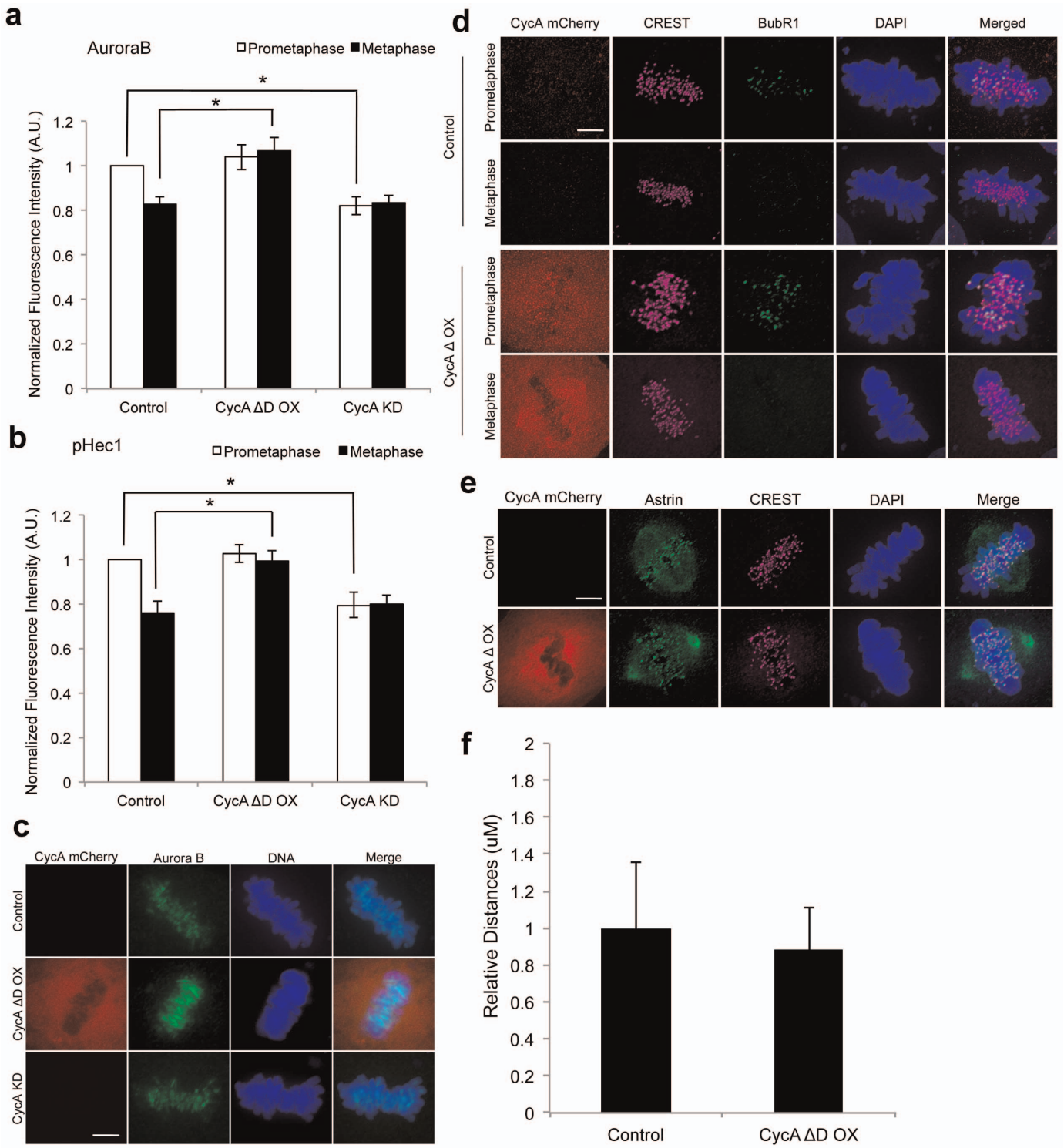
Extended Data Figure 4 | Poleward microtubule flux. **a**, DIC and time-lapse fluorescence images of an individual U2OS cell in prometaphase (top) and metaphase (bottom). Scale bar, 5 μm. **b**, fluorescence intensity linescan of spindles shown in **a**.



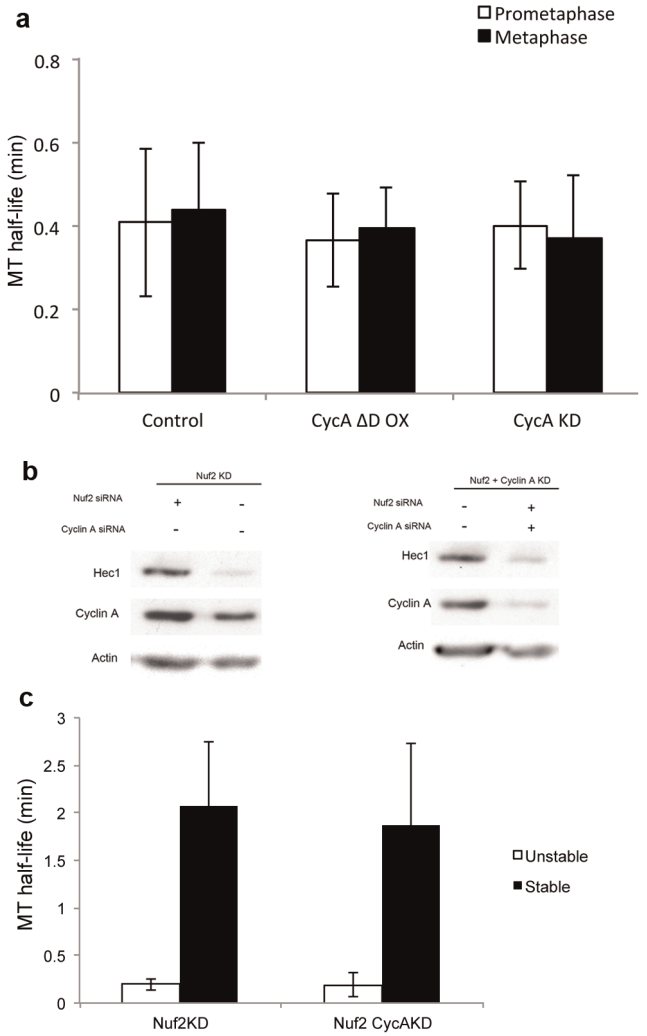
Extended Data Figure 5 | Single-cell measurements in U2OS cells. **a**, DIC and time-lapse fluorescence images of an individual U2OS cell in prometaphase and then in metaphase. Scale bar, 5 μm. **b**, k–MT half-life of individual U2OS cells photoactivated serially in prometaphase (left) or in prometaphase and then again in metaphase (right). **c**, Percentage of MTs in the stable population (for example, k–MTs) calculated from the exponential decay curve of serially photoactivated prometaphase fluorescence ($R^2 > 0.99$).



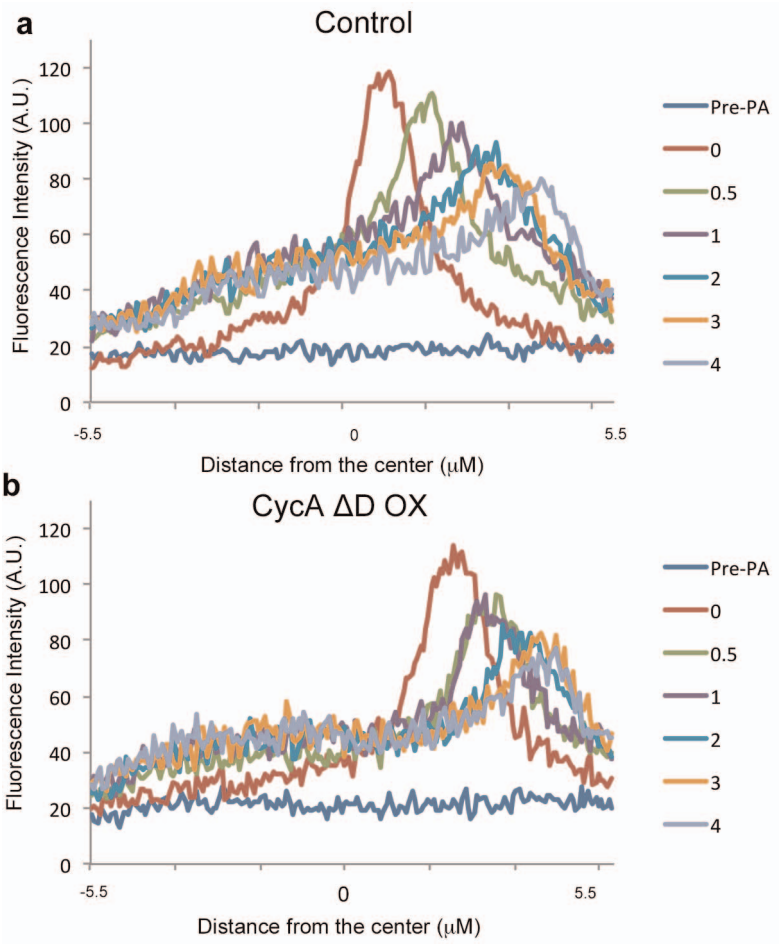
Extended Data Figure 6 | Manipulation of cyclin A levels. **a**, Western blots of cyclin-A-overexpressing (left) and cyclin-A-depleted (right) U2OS cells compared to control. siRNA, short interference RNA. **b**, Cyclin A(ΔD)–mCherry fluorescence intensity and respective k–MT half-life of U2OS cells photoactivated in prometaphase (top) and metaphase (bottom). Linear fit with R^2 value. **c**, Normalized fluorescence over time after photoactivation of untreated U2OS (control), U2OS cells overexpressing cyclin A(ΔD)–mCherry (CycA(ΔD)OX) and depleted of cyclin A (CycA KD); $n = 13$ cells for control and 10 cells for CycA(ΔD)OX and CycA KD per condition.



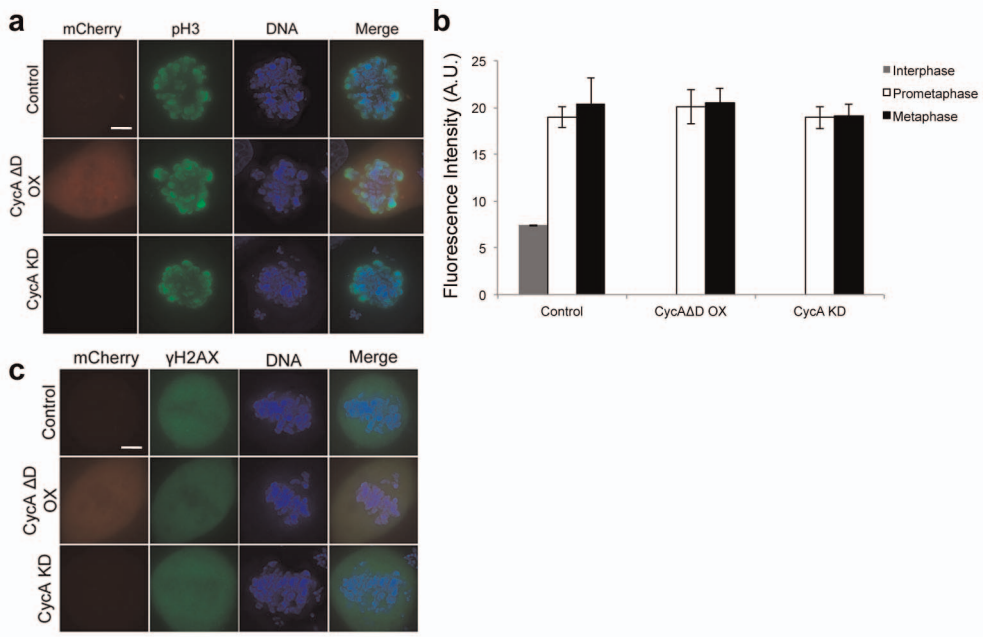
Extended Data Figure 7 | Mitotic properties of cells with altered levels of cyclin A. Immunofluorescence of prometaphase and metaphase untreated U2OS (control) and cyclin A(ΔD)-overexpressing U2OS cells. **a**, Fluorescence intensities of centromeres stained for aurora B kinase in U2OS cells normalized using CREST in both prometaphase and metaphase. **b**, Fluorescence intensities of centromeres stained for phosphor (p)-HEC1 in U2OS cells normalized using CREST in both prometaphase and metaphase. **c**, Immunofluorescence of untreated metaphase U2OS (control), U2OS cells overexpressing cyclin A(ΔD)–mCherry and depleted of cyclin A. Scale bar, 5 μm. Graphs show mean ± s.e.m. from 20 cells per condition from three independent experiments. $^*P \leq 0.01$, two-tailed *t*-test. **d**, Localization of BUB1B in prometaphase and metaphase cells with and without (control) expression of cyclin A(ΔD). **e**, Localization of astrin in metaphase cells with and without (control) expression of cyclin A(ΔD); $n = 30$ cells per condition from three independent experiments. Scale bar, 5 μm. **f**, Intercentromere distances of untreated and cyclin A(ΔD)-overexpressing U2OS cells; $n = 30$ cells per condition from three independent experiments. Graphs show mean ± s.e.m. $^*P \leq 0.01$, two-tailed *t*-test.



Extended Data Figure 8 | k–MT are selectively influenced by cyclin A. **a**, MT half-life of untreated (control), cyclin-A-overexpressing and cyclin-A-depleted prometaphase and metaphase U2OS cells measured at 5-s intervals for 1 min; $n = 10$ cells per condition. **b**, western blots of NUF2-depleted (left) and NUF2- and cyclin-A-depleted (right) U2OS cells compared to control U2OS cells. **c**, MT half-life of NUF2-depleted and NUF2- and cyclin-A-depleted U2OS cells; $n = 10$ cells per condition. Graphs show mean ± s.e.m.



Extended Data Figure 9 | Poleward flux in cells expressing mutant cyclin A. Linescan analysis measuring fluorescence intensity of metaphase spindles in untreated (control) and cyclin A overexpression in photoactivatable GFP–tubulin-expressing U2OS cells.



Extended Data Figure 10 | Properties of mitotic cells expressing mutant cyclin A. Immunofluorescence of untreated metaphase U2OS cells (control), U2OS cells overexpressing cyclin A(ΔD)–mCherry and U2OS cells depleted of cyclin A (CycA KD). Scale bar, 5 μm. **b**, Fluorescence intensities of DNA stained with DAPI in U2OS cells in both prometaphase and metaphase. Scale bar, 5 μm; $n = 60$ cells per condition from three independent experiments. Graphs show mean ± s.e.m.

LETTER

doi:10.1038/nature12265

Transport dynamics in a glutamate transporter homologue

Nurunisa Akyuz[1], Roger B. Altman[1], Scott C. Blanchard[1] & Olga Boudker[1]

Glutamate transporters are integral membrane proteins that catalyse neurotransmitter uptake from the synaptic cleft into the cytoplasm of glial cells and neurons[1]. Their mechanism of action involves transitions between extracellular (outward)-facing and intracellular (inward)-facing conformations, whereby substrate binding sites become accessible to either side of the membrane[2]. This process has been proposed to entail transmembrane movements of three discrete transport domains within a trimeric scaffold[3]. Using single-molecule fluorescence resonance energy transfer (smFRET) imaging[4], we have directly observed large-scale transport domain movements in a bacterial homologue of glutamate transporters. We find that individual transport domains alternate between periods of quiescence and periods of rapid transitions, reminiscent of bursting patterns first recorded in single ion channels using patch-clamp methods[5,6]. We propose that the switch to the dynamic mode in glutamate transporters is due to separation of the transport domain from the trimeric scaffold, which precedes domain movements across the bilayer. This spontaneous dislodging of the substrate-loaded transport domain is approximately 100-fold slower than subsequent transmembrane movements and may be rate determining in the transport cycle.

In the brain, glutamate mediates excitatory synaptic transmission, responsible for learning, memory formation and cognition[1,7]. Glutamate transporters are electrochemically driven pumps that maintain a low neurotransmitter background at glutamatergic synapses, allowing for repeated rounds of signalling and preventing excitotoxicity[8]. The sodium/aspartate symporter from *Pyrococcus horikoshii*, Glt_{Ph}, is the only glutamate transporter homologue with known three-dimensional structures of both outward- and inward-facing states[3,9]. Correspondingly, this system has served as a valuable model for establishing the structural and dynamic underpinnings of the transport cycle[10–12]. Because Glt_{Ph} originates from a hyper-thermophilic archaeon, it has a slow turnover time of ~200 s at room temperature[12], indicating that the dynamic processes required for transport may fall within the time regime accessible through smFRET imaging[4,13].

Crystal structures have shown that Glt_{Ph} is a homotrimer in which each protomer comprises two domains: a rigid trimerization domain[14], which serves as a scaffold and mediates the inter-protomer interactions, and a transport domain, which can move over 15 Å across the bilayer[3]. High extracellular and low intracellular chemical potentials of sodium ions (Na^+) drive substrate binding and unbinding to the transporter, respectively[15]. By contrast, large-scale transport domain movements, thought to mediate the translocation of substrate aspartate (Asp) and coupled Na^+ ions across the cellular membranes[3,16], are expected to occur spontaneously[3].

We probed the dynamics of the transport domain movements using wide-field, total internal reflection smFRET imaging[17]. To detect the relative motions of two transport domains, we introduced single cysteine mutations at positions of low sequence conservation and high solvent accessibility that exhibited inter-protomer distance changes >20 Å in the crystal structures of the outward- and inward-facing Glt_{Ph} (Fig. 1a). The three cysteines present in each Glt_{Ph} trimer were derivatized with a mixture of maleimide-activated Cy3, Cy5 and biotin–$(PEG)_{11}$ (Methods). This labelling strategy ensured that only trimers containing one of each of these moieties were immobilized on streptavidin-decorated surfaces and yielded detectable FRET signals (Fig. 1b). The constructs pursued showed selective labelling reactivity, >80% functional activity and biotin-specific surface attachment (Supplementary Figs 1 and 2). To parallel previous crystallographic studies, Glt_{Ph} mutants were first characterized in the presence of saturating concentrations of Na^+ and Asp (200 mM and 100 μM, respectively).

For a control Glt_{Ph}(K55C) construct labelled within the rigid trimerization domain, individual smFRET trajectories exhibited stable, high FRET before photobleaching (τ_{FRET} ~40 s) (Supplementary Fig. 3). FRET efficiency distributions, calculated from hundreds of trajectories, yielded a single, narrow peak centred at 0.8 (Fig. 1c, Supplementary Table 1). These data indicated robust fluorophore performance[18] and confirmed that the trimerization domain lacks significant dynamics[3,14]. By contrast, Glt_{Ph} labelled on I294C or N378C, located on the cytoplasmic and extracellular surfaces of the transport domain, respectively, exhibited clear evidence of discrete FRET states. In both constructs, low- and high-FRET states were observed (I294C: 0.35 and 0.55; N378C: 0.35 and 0.8) (Fig. 1d, e), consistent with the inter-protomer distances in outward- and inward-facing Glt_{Ph} structures (Fig. 1a and Supplementary Table 1). Here, the intracellular Glt_{Ph}(I294C) construct preferentially populated a high-FRET state, whereas the extracellular Glt_{Ph}(N378C) construct favoured a low-FRET state (Fig. 1d, e). These data indicate that the individual Glt_{Ph} protomers preferentially reside in outward-facing conformations, but also adopt less favoured inward-facing orientations. In both systems, intermediate-FRET states were also observed (Fig. 1d, e). This state was better resolved in Glt_{Ph}(N378C) (0.55), where it could be attributed to a configuration in which neighbouring subunits adopt inward- and outward-facing orientations (Supplementary Methods, Supplementary Fig. 4 and Supplementary Table 1). Such a configuration is consistent with a recent crystal structure of an asymmetric Glt_{Ph} trimer[19]. Notably, similar FRET states and FRET state occupancies were obtained when Glt_{Ph}(N378C) was reconstituted into liposomes (Supplementary Methods, Supplementary Fig. 5).

To validate our FRET state assignments, we introduced mutations distal from the site of fluorophore attachment in Glt_{Ph}(N378C) to either stabilize or destabilize the outward-facing state (Fig. 1a, Supplementary Fig. 6). A G221A mutation was introduced within the hinge region between the transport and trimerization domains, to hinder local rearrangements accompanying inward transport domain motions[3]. A K290A mutation was introduced to disrupt a network of polar interactions at the interface between the transport and trimerization domains in the outward-facing structure[11]. Consistent with expectations, we observed dramatic increases and decreases in the low-FRET, outward-facing state population for the G221A and K290A mutants, respectively (Fig. 1f, g). We conclude that transitions between FRET states in Glt_{Ph}(N378C) reflect motions of the transport domains between outward- and inward-facing orientations. This construct was pursued to examine the effect of ligand binding on conformational dynamics.

In the absence and in the presence of Na^+ and Asp, the FRET state values were largely similar, whereas ligands shifted the state occupancies

[1]Department of Physiology and Biophysics, Weill Cornell Medical College, 1300 York Avenue, New York, New York 10064, USA.

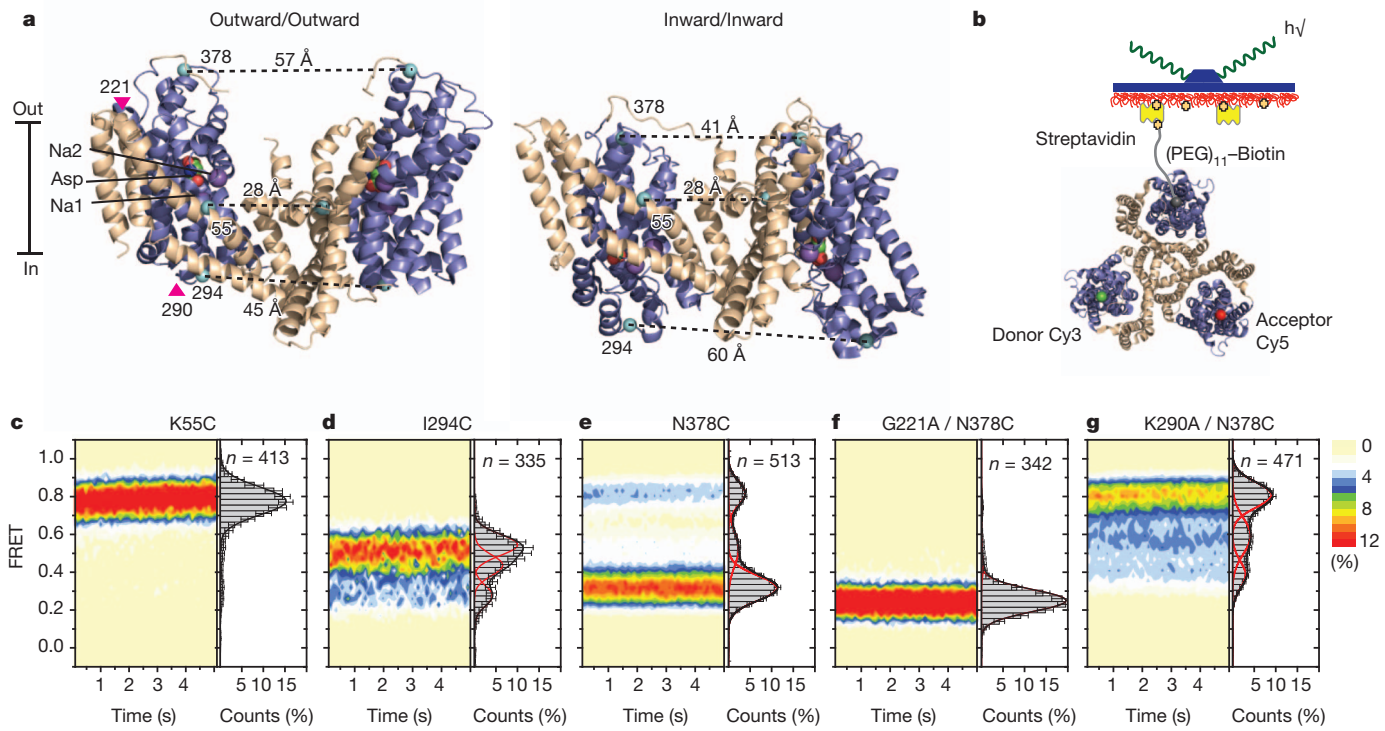


Figure 1 | FRET efficiency changes reflect relative orientations of the transport domains. **a**, Glt_{Ph} protomer pairs in symmetrical outward- and inward-facing states viewed within the membrane plane. Trimerization and transport domains are coloured wheat and blue, respectively. Bound Asp and Na^+ ions are emphasized as spheres and coloured by atom type. Introduced cysteines are highlighted in cyan with inter-protomer distances above the dotted lines. Magenta arrows mark sites of mutations altering state distributions. **b**, Labelling and surface-immobilization strategies. **c–g**, FRET efficiency population histograms for Asp/Na^+-bound transporters. Introduced mutations are indicated above the panels. The number of molecules analysed (n) is shown. Population contour plots (left) are colour-coded from tan (lowest) to red (highest population) with the colour scale shown beside the graphs. The cumulative population histogram (right) displays the time-averaged values and standard deviations. The solid black lines are fits to the sums of individual Gaussian functions (red lines).

in favour of the low-FRET state (Supplementary Fig. 5). Strikingly, individual smFRET traces exhibited much more frequent transitions in the apo compared to the bound transporter (Fig. 2a, b). To quantify the dynamics, we idealized the smFRET trajectories using a model containing three kinetically linked, non-zero FRET states (Methods, Supplementary Fig. 7). The quality of idealizations and the average rates of the conformational transitions were assessed by inspection of individual trajectories as well as transition density plots, which report on the frequencies of transitions between distinct FRET states[20]. In apo Glt_{Ph}, a comparable number of transitions occurred between low-, intermediate- and high-FRET states with an average frequency of 0.5 s^{-1} (Fig. 2c), suggesting that the apo transporter samples outward- and inward-facing states relatively rapidly. For the substrate-loaded transporter, the frequency of transitions was reduced by more than an

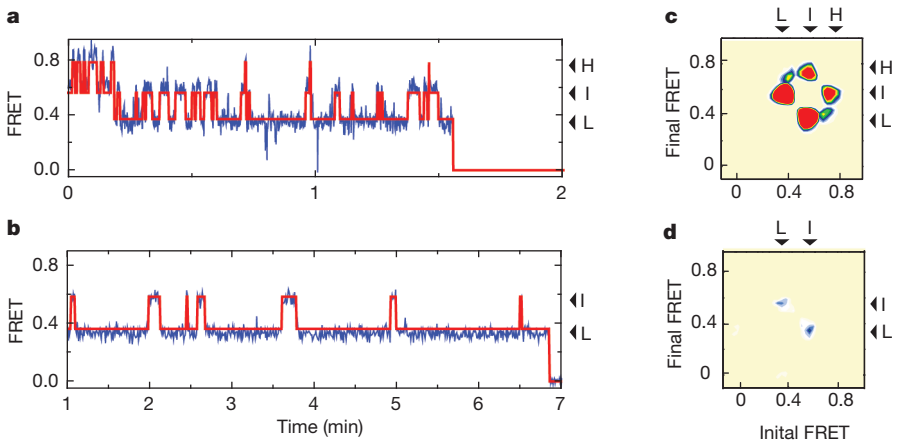


Figure 2 | Dynamics in the apo and substrate bound transporter. **a**, **b**, Shown are smFRET trajectories (blue), acquired for apo (**a**) and Na^+/Asp-bound (**b**) Glt_{Ph}(N378C). Overlaid are idealizations generated in QuB (red). Arrows mark population averages for the low- (L), intermediate- (I) and high- (H) FRET efficiencies. Data in panel **b** were collected using 400 ms integration time. **c**, **d**, Transition density plots for the apo (**c**) and Na^+/Asp-bound (**d**) transporters show that transitions occur between three distinct FRET states (L, I and H) with an average frequency of ~0.5 s^{-1} and ~0.02 s^{-1}, respectively. Initial and final FRET values for each transition are accumulated into two-dimensional histograms. Colour scale is from tan (lowest frequency) to red (highest frequency).

order of magnitude ($0.02\,s^{-1}$). Under these conditions, we primarily observed transitions between low- and intermediate-FRET states (Fig. 2d), consistent with transport domains moving inward one at a time. The paucity of direct transitions from low- to high-FRET states in the presence and in the absence of Na^+ and Asp is in line with previous functional studies[14,21], suggesting that individual protomers undergo conformational transitions, and thus function independently of each other.

To assess the concentration dependence of the ligand-induced effects, we performed two titrations on Glt_{Ph}(N378C): a Na^+ titration in the presence of 10 μM Asp and an Asp titration in the presence of 2 mM Na^+. At the population level, we observed a gradual stabilization of the low-FRET, outward-facing Glt_{Ph} configuration as concentrations of both ligands increased (Fig. 3). Consistent with previous bulk experiments[10], these binding isotherms yielded apparent dissociation constants (K_d) of approximately 2 mM for Na^+ and 7 μM for Asp (Supplementary Methods). The titration data also showed a gradual decrease in the average transition frequency. At intermediate Asp concentrations, smFRET trajectories showed molecules reversibly switching between periods of quiescence and periods of rapid transitions, probably due to apo protomers (Fig. 4a, top panel). Increased substrate concentrations shortened the durations of the dynamic periods and increased the lifetimes of the stable states (Fig. 4a, lower panels; Fig. 4b). The midpoints of these changes (~10 μM) were consistent with the K_d of Asp binding obtained from the population data. From the apparent lifetimes of the dynamic and non-dynamic modes, we estimate the ligand binding and dissociation rates of $k_{on} \sim 10^4\,M^{-1}\,s^{-1}$ and $k_{off} \sim 0.1\,s^{-1}$, respectively, at 2 mM Na^+ (Supplementary Methods). A substrate residence time of 10 s is consistent with previous estimates[12].

Remarkably, dynamic and non-dynamic periods were also observed in the absence of both Na^+ and Asp. Visual inspection of smFRET trajectories revealed molecules in which (1) rapid dynamics persisted throughout the entire observation periods; (2) stable FRET states predominated; and (3) dynamic and non-dynamic periods were reversibly visited (Supplementary Fig. 8). Consequently, biphasic FRET state lifetimes were observed (Supplementary Fig. 9), wherein the longer dwell times reflected the durations of non-dynamic periods and the shorter dwell times reflected the lifetimes of the FRET states during rapid dynamics. On average, dynamic and non-dynamic periods were approximately equally represented (Fig. 4b). The addition of Na^+ and Asp shortened the dynamic periods. However, periods of dynamics persisted, even as the substrate binding sites became saturated (Fig. 4b). In the presence of 200 mM Na^+ (a Na^+ concentration at which the K_d for Asp is ~1 nM (ref. 10)) and 100 μM Asp, when we expect the fraction of the unbound transporters to be 1 in 10^5, we still observed brief dynamic periods. Such bursts were typified by one or two transient excursions to an inward-facing state with an average dwell time of 1 s (Fig. 4c and Supplementary Fig. 8), before returning to outward-facing conformations (Fig. 4d). 'Flickers' of this kind occurred every ~200 s and constituted about half of all transitions observed out of the outward-facing, low-FRET state. The remainder of the transitions out of the low-FRET state lead to long-lasting intermediate- or high-FRET states (Supplementary Fig. 8). To confirm that such flickers were due to the rapid transitions of the substrate-loaded transport domains, we performed identical experiments in the presence of saturating concentrations of D,L-threo-β-benzyloxyaspartate (TBOA)[22], which binds to the substrate-binding site in a manner that blocks transport[10]. Compared to Asp, TBOA reduced flickers by >80% (Fig. 4e), suggesting that these excursions reflect translocation of Asp across the bilayer. Notably, similar dynamic signatures were also observed for Glt_{Ph} within the context of proteoliposomes (Supplementary Fig. 8).

We conclude that dynamic heterogeneity, which manifests kinetically as periods of quiescence punctuated by periods of rapid transitions, is an intrinsic property of Glt_{Ph} arising from spectroscopically hidden isomerization of the protein[23]. We posit that quiescent periods of substrate-bound Glt_{Ph} reflect stable conformations that closely resemble crystal structures of the outward- and inward-facing states[3,9]. In these states, the transport and trimerization domains are closely packed. Dynamic periods reflect transmembrane movements of individual transport domains. These movements are enabled by a critical protein isomerization, which involves dislodging of the transport domain from the trimeric scaffold (Supplementary Fig. 10a). Such an 'unlocked' configuration has been captured crystallographically[19] showing disrupted packing and polar interactions at the domain interface. Consistent with this model, the interfacial K290A mutation (Supplementary Fig. 6) dramatically increases the population of molecules in the dynamic mode (Supplementary Fig. 10b–e). The existence of unlocked intermediates with potentially increased hydration at the domain interface[24] may help explain the well known capacity of glutamate transporters and Glt_{Ph} to catalyse bidirectional anion fluxes[25,26]. Our data also show that binding of Na^+ and Asp significantly increases the lifetimes of stable outward- and inward-facing states, suggesting that there is an allosteric coupling between the substrate-binding site and the domain interface.

Mammalian glutamate transporters are architecturally similar to Glt_{Ph}, and their transport cycle is anticipated to proceed through similar intermediates. Current kinetic models for the mammalian transporters suggest that substrate translocation is relatively fast[27,28]. However, in Glt_{Ph} we find that rapid translocation is preceded by a slow isomerization step that occurs on a timescale comparable to the transporter turnover rate ~$0.005\,s^{-1}$ (every 200 s). Therefore, under our experimental conditions, isomerization of the substrate-loaded Glt_{Ph} may constitute a key, rate-limiting step in the transport cycle, whereas return of the unloaded transport domain to an outward-facing state is relatively fast (~$0.5\,s^{-1}$). The apparent differences between Glt_{Ph} and the mammalian transporters may, at least in part, be because only ensemble properties have thus far been obtained in the latter system. Here, the application of the glutamate substrate to cell membranes expressing the mammalian transporters induced transient electric currents decaying to

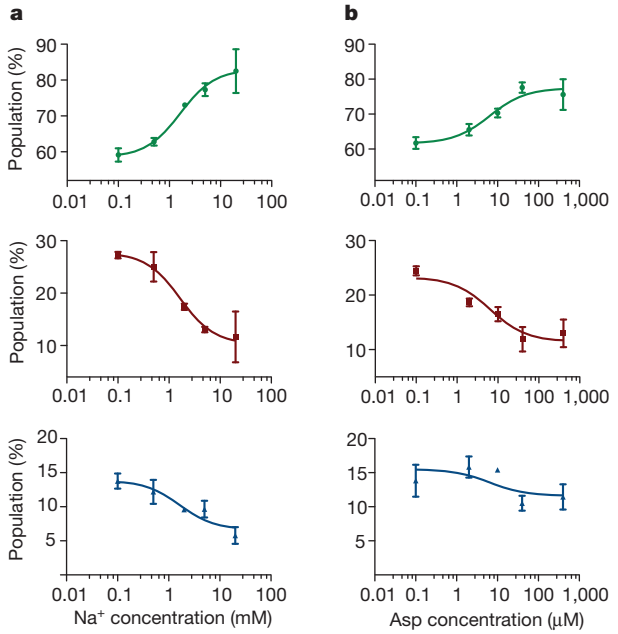


Figure 3 | Na^+ ions and Asp favour the outward facing state. Populations of low- (top, green), intermediate- (middle, red) and high- (bottom, blue) FRET states as a function of Na^+ ions added in the presence of 10 μM Asp (**a**) and Asp added in the presence of 2 mM Na^+ (**b**). The titrations yielded dissociation constants (K_d) of 1.6 ± 0.3 mM and 6.5 ± 2.5 μM and Hill coefficients of 1.3 ± 0.3 and 0.9 ± 0.3, respectively. Shown are averages and standard deviations from at least three independent data sets (each containing at least 250 molecules). A few error bars were too small to be clearly visible. Solid lines through the data points are the results of global fitting of the data to the Hill equation.

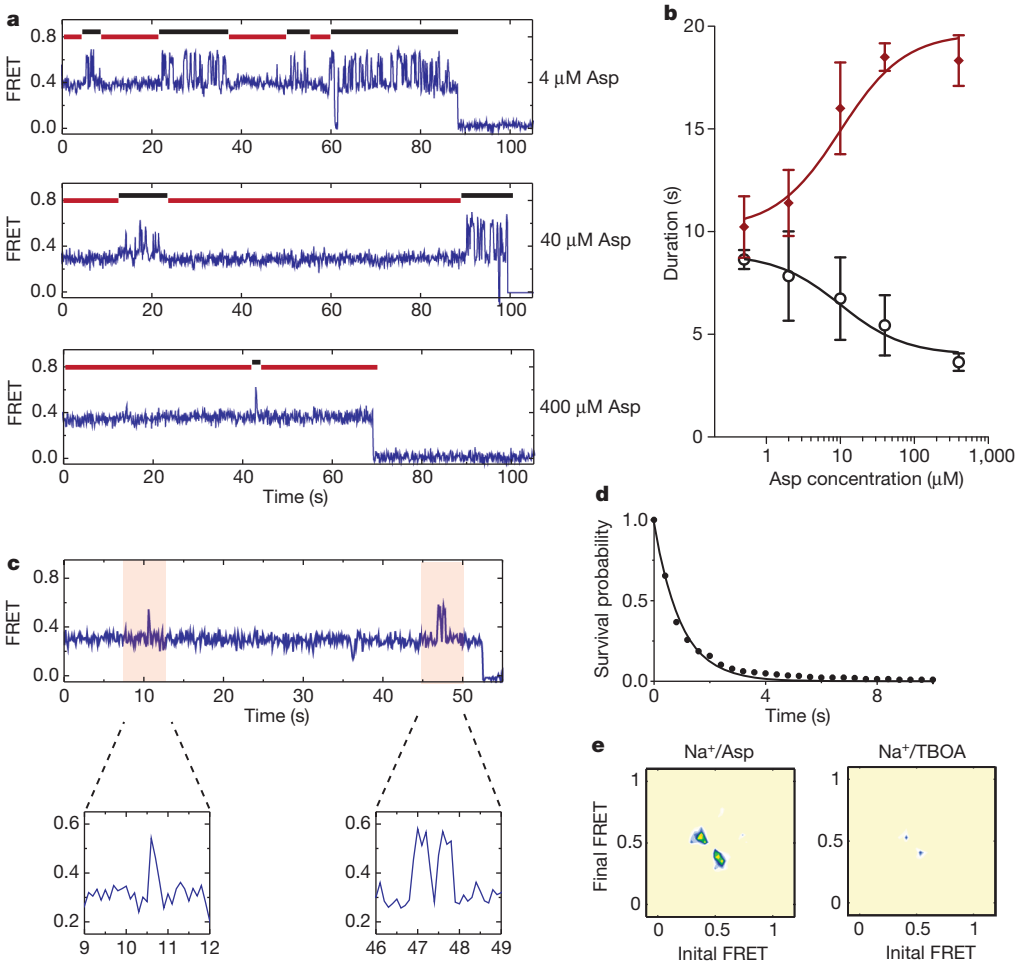


Figure 4 | Modulation of dynamics by substrate and inhibitor binding. **a**, smFRET traces obtained in the presence of 2 mM Na^+ and increasing Asp concentrations. **b**, The apparent durations of the dynamic (black) and quiescent (red) periods as a function of Asp concentration. Shown are averages and standard deviations for three independent data sets containing at least 250 molecules each. **c**, A smFRET trace obtained in the presence of 200 mM Na^+ and 100 μM Asp. Expanded views of the flicker events (shaded in pink) are shown below the trace. **d**, Survival plot of the observed flickers. Solid line is a fit to a single exponential decay. **e**, Transition density plots for flicker events in saturating Asp (left) or TBOA (right). Average transition frequencies are $0.02\ s^{-1}$ and $<0.005\ s^{-1}$, respectively. Data in panels **d** and **e** were collected with 400 ms integration time.

lower steady-state levels. These currents are thought to reflect rapid substrate translocation followed by slower transport cycle events, including relocation of the unloaded transporter. We speculate that the transient currents observed in these experiments reflect translocation of the already unlocked transporter, whereas the slower process of dislodging of the transport domain is masked by steady-state, asynchronous events.

Future efforts must be aimed at delineating the temporal relationship between transport domain translocation and substrate release and the correspondence between transport domain dynamics and anion conductance. Such efforts will be greatly aided by the establishment of imaging strategies wherein stable transmembrane gradients can be maintained over extended periods, the advent of fluorophores that are highly stable in the lipid environment, and technologies enabling the detection of single transport events.

METHODS SUMMARY

Glt_{Ph} mutants were generated within a Glt_{Ph} variant lacking cysteines in which seven unconserved residues have been replaced with histidines resulting in improved expression levels (termed Glt_{Ph} hereafter for brevity)[9]. Proteins were expressed in DH10-B *E. coli* cells as C-terminal $(His)_8$ fusion proteins, solubilized in *n*-dodecyl β-D-maltopyranoside and purified by metal affinity chromatography, followed by proteolytic removal of the $(His)_8$-tag and further purification by size exclusion chromatography[9]. Protein samples at 40 μM were labelled with a mixture of maleimide-activated Cy3, Cy5 and biotin–$(PEG)_{11}$ at respective concentrations of 50, 100 and 25 μM for 30 min at room temperature. Labelled proteins were purified away from the excess reagents by size exclusion chromatography, and their purity and specificity of labelling were assessed by SDS–PAGE followed by fluorescent imaging and Coomassie staining. The labelled proteins were reconstituted into liposomes and their Asp uptake activity measured as previously described[10,12]. For smFRET experiments, labelled proteins were surface-immobilized through the biotin-streptavidin bridge within passivated, streptavidin-derivatized microfluidic devices as previously described[29] and imaged using a home-built total internal reflection fluorescence microscope under oxygen-scavenging conditions[18]. Determination of FRET efficiencies and selection of traces were performed using automated analysis software developed in the laboratory[30].

Full Methods and any associated references are available in the online version of the paper.

Received 2 November 2012; accepted 30 April 2013.
Published online 23 June 2013.

1. Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65,** 1–105 (2001).
2. Krishnamurthy, H., Piscitelli, C. L. & Gouaux, E. Unlocking the molecular secrets of sodium-coupled transporters. *Nature* **459,** 347–355 (2009).
3. Reyes, N., Ginter, C. & Boudker, O. Transport mechanism of a bacterial homologue of glutamate transporters. *Nature* **462,** 880–885 (2009).
4. Weiss, S. Fluorescence spectroscopy of single biomolecules. *Science* **283,** 1676–1683 (1999).
5. Sakmann, B., Patlak, J. & Neher, E. Single acetylcholine-activated channels show burst-kinetics in presence of desensitizing concentrations of agonist. *Nature* **286,** 71–73 (1980).
6. Cull-Candy, S. G. & Parker, I. Rapid kinetics of single glutamate-receptor channels. *Nature* **295,** 410–412 (1982).

7. Tzingounis, A. V. & Wadiche, J. I. Glutamate transporters: confining runaway excitation by shaping synaptic transmission. *Nature Rev. Neurosci.* **8,** 935–947 (2007).
8. Zerangue, N. & Kavanaugh, M. P. Flux coupling in a neuronal glutamate transporter. *Nature* **383,** 634–637 (1996).
9. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
10. Boudker, O. *et al.* Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445,** 387–393 (2007).
11. Boudker, O. & Verdon, G. Structural perspectives on secondary active transporters. *Trends Pharmacol. Sci.* **31,** 418–426 (2010).
12. Ryan, R. M., Compton, E. L. & Mindell, J. A. Functional characterization of a Na^+-dependent aspartate transporter from *Pyrococcus horikoshii*. *J. Biol. Chem.* **284,** 17540–17548 (2009).
13. Blanchard, S. C. Single-molecule observations of ribosome function. *Curr. Opin. Struct. Biol.* **19,** 103–109 (2009).
14. Groeneveld, M. & Slotboom, D. J. Rigidity of the subunit interfaces of the trimeric glutamate transporter GltT during translocation. *J. Mol. Biol.* **372,** 565–570 (2007).
15. Reyes, N., Oh, S. & Boudker, O. Binding thermodynamics of a glutamate transporter homolog. *Nature Struct. Mol. Biol.* **20,** 634–640 (2013).
16. Groeneveld, M. & Slotboom, D. J. Na^+:aspartate coupling stoichiometry in the glutamate transporter homologue Glt_{Ph}. *Biochemistry* **49,** 3511–3513 (2010).
17. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature Methods* **5,** 507–516 (2008).
18. Dave, R., Terry, D. S., Munro, J. B. & Blanchard, S. C. Mitigating unwanted photophysical processes for improved single-molecule fluorescence imaging. *Biophys. J.* **96,** 2371–2381 (2009).
19. Verdon, G. & Boudker, O. Crystal structure of an asymmetric trimer of a bacterial glutamate transporter homolog. *Nature Struct. Mol. Biol.* **19,** 355–357 (2012).
20. McKinney, S. A., Joo, C. & Ha, T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91,** 1941–1951 (2006).
21. Grewer, C. *et al.* Individual subunits of the glutamate transporter EAAC1 homotrimer function independently of each other. *Biochemistry* **44,** 11913–11923 (2005).
22. Shimamoto, K. *et al.* DL-*threo*-β-benzyloxyaspartate, a potent blocker of excitatory amino acid transporters. *Mol. Pharmacol.* **53,** 195–201 (1998).
23. Liu, S., Bokinsky, G., Walter, N. G. & Zhuang, X. Dissecting the multistep reaction pathway of an RNA enzyme by single-molecule kinetic "fingerprinting". *Proc. Natl Acad. Sci. USA* **104,** 12634–12639 (2007).
24. Stolzenberg, S., Khelashvili, G. & Weinstein, H. Structural intermediates in a model of the substrate translocation path of the bacterial glutamate transporter homologue Glt_{Ph}. *J. Phys. Chem. B* **116,** 5372–5383 (2012).
25. Fairman, W. A. *et al.* An excitatory amino-acid transporter with properties of a ligand-gated chloride channel. *Nature* **375,** 599–603 (1995).
26. Wadiche, J. I., Amara, S. G. & Kavanaugh, M. P. Ion fluxes associated with excitatory amino acid transport. *Neuron* **15,** 721–728 (1995).
27. Grewer, C., Watzke, N., Wiessner, M. & Rauen, T. Glutamate translocation of the neuronal glutamate transporter EAAC1 occurs within milliseconds. *Proc. Natl Acad. Sci. USA* **97,** 9706–9711 (2000).
28. Otis, T. S. & Kavanaugh, M. P. Isolation of current components and partial reaction cycles in the glial glutamate transporter EAAT2. *J. Neurosci.* **20,** 2749–2757 (2000).
29. Blanchard, S. C. *et al.* tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101,** 12893–12898 (2004).
30. Zhao, Y. *et al.* Single-molecule dynamics of gating in a neurotransmitter transporter homologue. *Nature* **465,** 188–193 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank D. S. Terry for his help with the design of smFRET experiments and discussions; Z. Zhou for the synthesis of cyanine fluorophores; E. Georgieva for initial biochemical characterization of the single-cysteine mutants; G. Verdon, A. Accardi and N. Reyes for helpful discussions and comments on the manuscript. The work was supported in part by the National Institute of Health grants 5U54GM087519 and R01NS064357.

Author Contributions N.A. purified Glt_{Ph} mutants, carried out the experiments and analysed the data. R.B.A. prepared reagents for smFRET experiments. N.A., O.B. and S.C.B. together designed, analysed and interpreted the experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.C.B. (scb2005@med.cornell.edu) or O.B. (olb2003@med.cornell.edu).

METHODS

DNA manipulations, protein expression, purification and labelling. Single cysteine mutations were introduced within a Glt_{Ph} variant lacking cysteines, in which seven unconserved residues have been replaced with histidines resulting in improved expression levels (termed Glt_{Ph} hereafter for brevity)[9], using a QuikChange kit (Strategene). Constructs were verified by DNA sequencing and transformed into *E. coli* DH10-B cells (Invitrogen). Proteins were expressed as C-terminal $(His)_8$ fusions as described previously[3,9]. Briefly, isolated cell membranes were resuspended in buffer A, containing 20 mM HEPES/NaOH, pH 7.4, 200 mM NaCl, 0.1 mM L-aspartate, 0.1 mM Tris (2-carboxyethyl)phosphine (TCEP). Membranes were solubilized in the presence of 40 mM *n*-dodecyl β-D-maltopyranoside (DDM) for 1 h at 4 °C. Solubilized transporters were purified by metal-affinity chromatography in buffer A supplemented with 1 mM DDM and eluted in 250 mM imidazole. The $(His)_8$-tag was cleaved by thrombin and proteins were further purified by size exclusion chromatography in buffer A supplemented with 1 mM DDM. Glt_{Ph} was labelled at a concentration of 40 μM in buffer A with a mixture of maleimide-activated Cy3, Cy5 and biotin–$(PEG)_{11}$ at 50, 100 and 25 μM final concentrations, respectively (molar ratio of ~ 1:2:0.5). Labelled proteins were quenched with 10 mM 2-mercaptoethanol and subsequently purified away from excess reagents by size exclusion chromotography. The extent of labelling was determined by measuring absorbance at 552 nm and 650 nm for Cy3 and Cy5, respectively.

Protein reconstitution into liposomes and transport assays. Labelled and unlabelled Glt_{Ph} variants were reconstituted into liposomes and assayed as previously described[10,12]. Briefly, liposomes, prepared from 3:1 (w/w) mixture of *E. coli* total lipid extract and egg yolk phosphotidylcholine (Avanti Polar Lipids) in a buffer containing 20 mM Tris/HEPES, pH 7.4 and 200 mM KCl, were destabilized by addition of Triton X-100 at a detergent to lipid ratio of 0.5:1 (w/w). For transport assays, proteins were added at final protein to lipid ratio of 1:400 (w/w) and incubation for 30 min at room temperature. Detergents were removed by repeated incubations with Biobeads as described[12]. The proteoliposomes were extruded through 400-nm filters before analysis[12]. To measure the uptake of radioactive substrate, proteoliposomes were diluted into reaction buffer containing 20 mM Tris/HEPES, pH 7.4, 200 mM NaCl and 0.3 μM [^{3}H]Asp at room temperature. In control experiments, NaCl was replaced with KCl. For smFRET experiments, protein to lipid ratio of 1:1,000 (w/w) was used and proteoliposomes were extruded through 100-nm filters.

smFRET experiments. All experiments were performed using a home-built prism-based total internal reflection fluorescence microscope constructed around a Nikon TE2000 Eclipse inverted microscope body. The samples were illuminated with a 532 nm laser (Laser Quantum), Cy3 and Cy5 fluorescence were separated using a 650DCXR dichroic filter (Chroma) mounted in a DualView apparatus (Photometrics). Imaging data were acquired using MetaMorph acquisition software (Molecular Devices) and an Evolve 512 EMCCD (Photometrics). Before the experiments, the passivated microfluidic imaging chambers were prepared and coated with streptavidin as previously described[29,31]. Briefly, the microfluidic channel was prepared by incubating with a buffer solution containing 10 mM Tris/acetate, pH 7.5 and 50 mM KCl, 1 μM BSA, 1 μM 25 nucleotide DNA duplex, 0.8 μM streptavidin (Invitrogen) and 0.1% (v/v) glycerol for 5 min. The channel was then rinsed thoroughly in buffer A with 1 mM DDM. All imaging experiments were performed in buffer A (unless otherwise stated), supplemented with 1 mM DDM, 2 mM cyclooctatetraene (Sigma), 5 mM 2-mercaptoethanol, an enzymatic oxygen scavenger system comprising 1 U per ml glucose oxidase (Sigma), 8 U per ml catalase (Sigma) and 0.1% glucose[18]. If not otherwise specified, all data were collected at an imaging rate of 10 s^{-1} (100 ms integration time).

Analysis of smFRET data. Fluorescence trajectories were selected for analysis using custom-made MATLAB- (MathWorks) encoded software[30,31] according to the following criteria: a single catastrophic photobleaching event; >8:1 signal-to-background noise ratio; a FRET lifetime of at least 15 frames. FRET trajectories were calculated from the acquired intensities, I_{Cy3} and I_{Cy5}, using the formula $FRET = I_{Cy5}/(I_{Cy3} + I_{Cy5})$. Population contour plots were constructed by superimposing the FRET data from individual traces. Histograms of these population data were fitted to the sum of three Gaussian functions for all mutants, except K55C, for which only a single Gaussian function was used. The Gaussian means, widths and amplitudes were optimized in Origin (OriginLab). The relative populations, the dwell times of each FRET state and the transition frequencies between states were obtained by idealizing the smFRET traces using QuB[32,33] (Supplementary Methods). Transition density plots were generated as previously described[20]. The global fits of the titration curves to Hill equations were performed in Prism[34] (Supplementary Methods). The dwell time survival plots were fitted to exponential decay functions and the logarithmic histograms of the dwell times were fitted to transformed probability density functions[35], respectively, using Origin (Supplementary Methods).

Preparation of structural figures. All structural renderings were generated using PyMOL (version 1.5.0.4, Schrödinger) and coordinates deposited in Protein Data Base: accession number 2NWX (ref. 10) for the outward-facing state and 3KBC (ref. 3) for the inward-facing state.

31. Munro, J. B., Altman, R. B., O'Connor, N. & Blanchard, S. C. Identification of two distinct hybrid state intermediates on the ribosome. *Mol. Cell* **25,** 505–517 (2007).
32. Qin, F., Auerbach, A. & Sachs, F. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* **79,** 1915–1927 (2000).
33. Qin, F. Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling. *Biophys. J.* **86,** 1488–1501 (2004).
34. Motulsky, H. & Christopoulos, A. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting* 256–311 (Oxford Univ. Press, 2004).
35. Sigworth, F. J. & Sine, S. M. Data transformations for improved display and fitting of single-channel dwell time histograms. *Biophys. J.* **52,** 1047–1054 (1987).

doi:10.1038/nature12538

Unsynchronised subunit motion in single trimeric sodium-coupled aspartate transporters

Guus B. Erkens[1], Inga Hänelt[1,2]†, Joris M. H. Goudsmits[1], Dirk Jan Slotboom[1,2,3] & Antoine M. van Oijen[1,2,3]

Excitatory amino acid transporters (EAATs) are secondary transport proteins that mediate the uptake of glutamate and other amino acids[1]. EAATs fulfil an important role in neuronal signal transmission by clearing the excitatory neurotransmitters from the synaptic cleft after depolarization of the postsynaptic neuron. An intensively studied model system for understanding the transport mechanism of EAATs is the archaeal aspartate transporter Glt_{Ph}[2–6]. Each subunit in the homotrimeric Glt_{Ph} supports the coupled translocation of one aspartate molecule and three Na^+ ions[2] as well as an uncoupled flux of Cl^- ions[7]. Recent crystal structures of Glt_{Ph}[3,5,6,8] revealed three possible conformations for the subunits, but it is unclear whether the motions of individual subunits are coordinated to support transport. Here, we report the direct observation of conformational dynamics in individual Glt_{Ph} trimers embedded in the membrane by applying single-molecule fluorescence resonance energy transfer (FRET). By analysing the transporters in a lipid bilayer instead of commonly used detergent micelles, we achieve conditions that approximate the physiologically relevant ones. From the kinetics of FRET level transitions we conclude that the three Glt_{Ph} subunits undergo conformational changes stochastically and independently of each other.

Ion-coupled transporters such as Glt_{Ph} couple the free energy of an ion gradient to the transport of solutes across the membrane. Several crystal structures of Glt_{Ph} show the transporter either in a symmetric state with all three subunits in an outward-facing[5] conformation (with the substrate-binding site exposed to the extracellular side), an inward-facing[3] one (with the substrate-binding site facing the intracellular side) (Fig. 1a), or in an asymmetric conformation[8] with one subunit in an intermediate and the two others in an inward-facing conformation. The asymmetry of the latter structure raises important mechanistic questions about the coordination of subunits during transport. Do the conformational changes in the subunits take place in a synchronised, co-directional manner? Or can the subunits assume different conformations within the trimer? The inter-subunit distances in Glt_{Ph} have been determined using electron paramagnetic resonance (EPR) and were found to be broadly distributed[9,10]. Whether this heterogeneity has a basis in the behaviour of individual subunits within the trimer or arises from the protein ensemble is difficult to resolve with population-averaging techniques. Only single-molecule techniques will allow one to distinguish between truly stochastic subunit dynamics or a more subtle form of coordination within the trimer (for example, rotary coupling of the transport cycles of individual subunits).

The application of single-molecule FRET techniques to visualize conformational dynamics devoid of population averaging[11] has only recently been pioneered to study membrane transporters[12–14]. Such single-molecule studies (including one on Glt_{Ph}[15]) have so far focused on detergent-solubilised rather than membrane-embedded transport proteins. Because ion gradients and membrane potentials cannot be generated in detergent solution and membrane proteins can behave radically different in detergent micelles rather than in a lipid membrane[16], it is important to develop methods for the single-molecule characterization of membrane-reconstituted transporters. To observe directly the transport dynamics and conformational changes of Glt_{Ph}, we have applied single-molecule FRET to liposome-reconstituted Glt_{Ph}.

To follow conformational changes in Glt_{Ph}, we created a single-cysteine mutant (S331C) at a position where the intersubunit distance differs strongly between the conformations observed in the crystal structures (Fig. 1a). The cysteines were labelled with donor (Alexa Fluor 555) and acceptor (Alexa Fluor 647) dyes, which allow the detection of the inter-subunit distance as a change in the FRET signal. To follow conformational dynamics of membrane-embedded Glt_{Ph} and to study conditions that require a vectorial system, we developed a single-molecule assay based on liposome-reconstituted Glt_{Ph} in which we could generate the Na^+ gradient required for substrate translocation as well as a membrane potential (see Methods). We used a 1:10,000 (w/w) protein:lipid ratio to ensure a small probability of 0.25% of trapping multiple Glt_{Ph} trimers in a single 100-nm liposome (see Methods). The liposomes were subsequently surface-tethered using the interactions between biotinylated lipids and a streptavidin-covered glass slide (Fig. 1b) and imaged with total internal reflection fluorescence (TIRF) microscopy (Fig. 1c). By evaluating the total fluorescence intensities and donor:acceptor fluorescence ratios, we were able to select trimers labelled with a single donor and acceptor (see Methods).

We performed single-molecule FRET experiments under three different conditions: in the absence of Na^+ and aspartate, in the presence of external Na^+ and a membrane potential but without aspartate, and finally in the presence of external Na^+, aspartate and a membrane potential. After subjecting several thousands of liposomes to the stringent selection criteria described in the supplementary information, we retained in total 30 FRET traces each with a duration of several minutes and reflecting liposomes containing a single trimer with the correct 1:1 labelling stoichiometry. These data represent a total of 2.2 h of FRET trajectories, containing 1,179 transitions. In the absence of aspartate and Na^+, the time-dependent FRET signal from individual Glt_{Ph} trimers shows a dynamic behaviour and alternates between a FRET efficiency of ~0.6 and ~0.9 (Fig. 2a). The timescale of the dynamics (several seconds for each conformation) is similar to the timescale of aspartate transport as determined previously in bulk experiments[4]. Notably, such dynamics would appear as a broad distance distribution in an ensemble-averaging experiment. Next, we added 50 mM NaP_i and the ionophore valinomycin, but not aspartate, to the external medium to create a Na^+ gradient ($\Delta[Na^+]$, inward directed) and an electrochemical membrane potential ($\Delta\Psi$, K^+ diffusion potential, negative inside). These conditions resulted in two populations of molecules that either alternate between ~0.6 and ~0.9 FRET efficiency, or ~0.4 and ~0.6 FRET efficiency (Fig. 2b). It has been determined previously that Glt_{Ph} has an equal probability of being reconstituted in an inside-out or inside-in orientation[4]. Because both the membrane potential and the $\Delta[Na^+]$ are directional, an asymmetric situation is thus created in which the membrane potential and the $\Delta[Na^+]$ act differently on either orientation. We therefore assign

[1]University of Groningen, Zernike Institute for Advanced Materials, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [2]University of Groningen, Groningen Biomolecular Science and Biotechnology Institute, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [3]University of Groningen, Centre for Synthetic Biology, Nijenborgh 4, 9747 AG Groningen, The Netherlands. †Present address: Goethe-University Frankfurt, Molecular Microbiology and Bioenergetics, Institute of Molecular Biosciences, Max-von-Laue-Strasse 9, 60438 Frankfurt am Main, Germany.

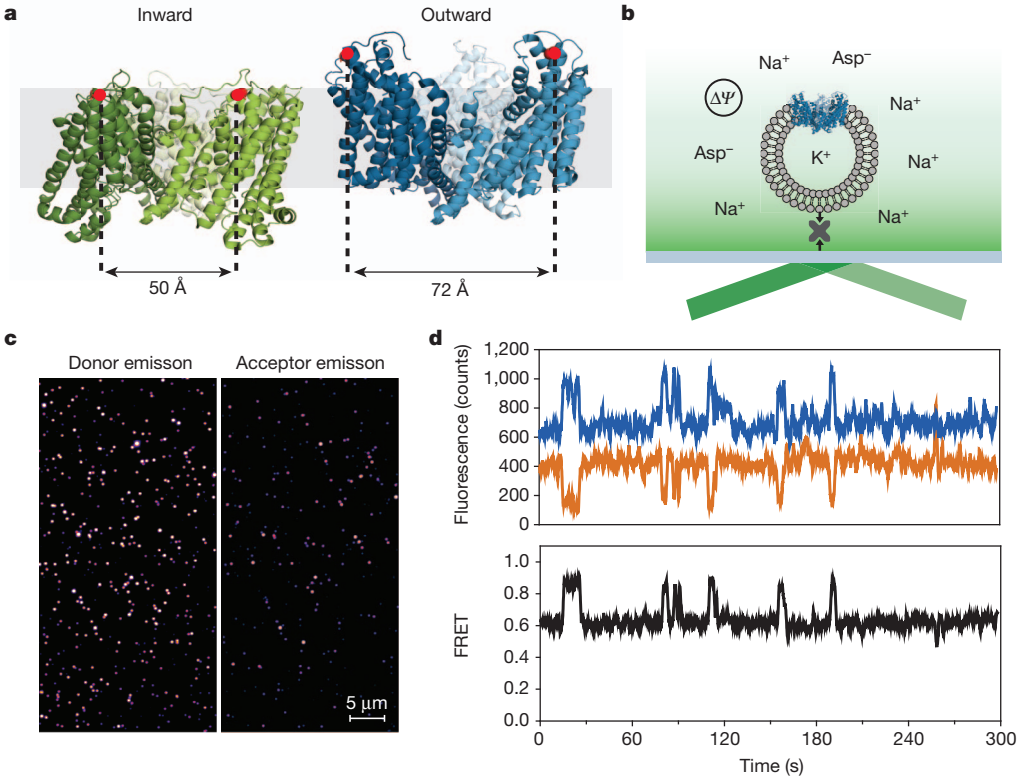


Figure 1 | Experimental setup. **a**, Structures of Glt_{Ph} with all three subunits in the inward-facing (left, green, Protein Data Bank (PDB) ID 3KBC) or outward-facing (right, blue, PDB ID 1XFH) conformation. The individual subunits have differently coloured shades. A cysteine at position 331 (red dots) was used for labelling with the donor and acceptor dyes. The inter-subunit distances are ~50 Å and ~70 Å in the all-inward-facing and in the all-outward-facing conformation, respectively. The grey bar indicates the approximate position of the lipid bilayer. **b**, Proteoliposomes with an embedded Glt_{Ph} trimer are attached to a coverslip using biotin–streptavidin interactions and imaged with TIRF microscopy. **c**, Fluorescence of double-labelled Glt_{Ph} upon excitation of the donor dye. The emission in the acceptor channel is an indication of FRET. **d**, Example trace of Glt_{Ph} single-molecule dynamics with donor (orange) and acceptor (blue) fluorescence (top panel) and the calculated FRET efficiencies (bottom panel).

the two populations of FRET state distributions to the two possible orientations of Glt_{Ph} in the liposomes. To exclude the possibility that the two populations result from a non-saturating concentration of Na^+, we performed the same experiments with a Na^+ concentration of 1 M. The resulting single-molecule FRET data show the same two populations as in the 50 mM Na^+ experiments. The appearance of a population that visits a lower FRET efficiency (FRET = 0.4) is also consistent with EPR measurements of the same mutant that demonstrated an increase of the average interspin distances[10]. In a third experiment, we added besides a $\Delta[Na^+]$ and $\Delta\Psi$ a 100 µM gradient of aspartate to create the conditions in which aspartate can be transported. In addition to small populations that alternated between 0.4/0.6 or 0.6/0.9 FRET efficiency, we now observed a dominant fraction (73%) that visits all three FRET levels and always transits from low (0.4) to high (0.9) FRET efficiency through the intermediate FRET (0.6) or vice versa (Fig. 2c). The timescale of the FRET transitions seems not to depend strongly on the presence or absence of substrates, that is, the transporter is always dynamic.

To address the question whether or not the subunits in a single Glt_{Ph} trimer can be simultaneously in different conformations, we considered single-molecule FRET traces from trimers that were labelled with three fluorescent dyes (1:2 or 2:1 donor:acceptor). A synchronous movement of the three subunits such that all subunits remain in identical conformations should give rise to the same pattern as we see for the 1:1 labelled complexes, because it would yield a convolution of identical FRET signals. In contrast, a large fraction of the triple-labelled Glt_{Ph} (71%) gave rise to a complex FRET trajectory containing a multitude of FRET levels (Fig. 3). We therefore conclude that the subunits in a single trimer can be in different conformations. However, this conclusion does not exclude coordination of their transitions within the trimer.

To gain a more detailed understanding on the coordination between subunits, we quantitatively analysed and modelled FRET traces from the 1:1 labelled complexes, obtained in the presence of external Na^+, aspartate and a membrane potential. Following the observation that asymmetric combinations of subunit conformations are possible, we considered the inter-subunit distances for asymmetric pairs of conformations (for example inward-outward or intermediate-outward) as derived from the available crystal structures[3,5,6,8]. All combinations fell in three distance categories: a short distance (~50 Å), intermediate distances (57–61 Å) and long distances (64–72 Å, see Supplementary Information). With multiple trimer configurations corresponding to similar inter-subunit distances, it follows that multiple configurations will give rise to the same FRET level. The distribution of lifetimes from a single FRET level is therefore not only necessarily determined by the lifetime of one particular conformation, but will depend on the lifetimes of several subsequent conformations that give rise to the same FRET level[17].

The exact relation between the observed FRET level dynamics and the underlying conformational dwell times is determined by the manner with which individual Glt_{Ph} subunits are coordinated in the trimer. For example, the relative number of times a transition from intermediate FRET to high FRET is observed versus a transition from intermediate to low FRET strongly depends on whether there is a preferred order in which trimer configurations follow each other in time (that is, whether the subunits are coordinated or not). To extract this information, we created an analytical model of the Glt_{Ph} transport cycle (Fig. 4a, Supplementary Information), which allows us to simulate the shape of the dwell-time distributions. Furthermore, we can calculate the relative number of times one FRET level follows after another (branching ratio) to check whether our model predicts the

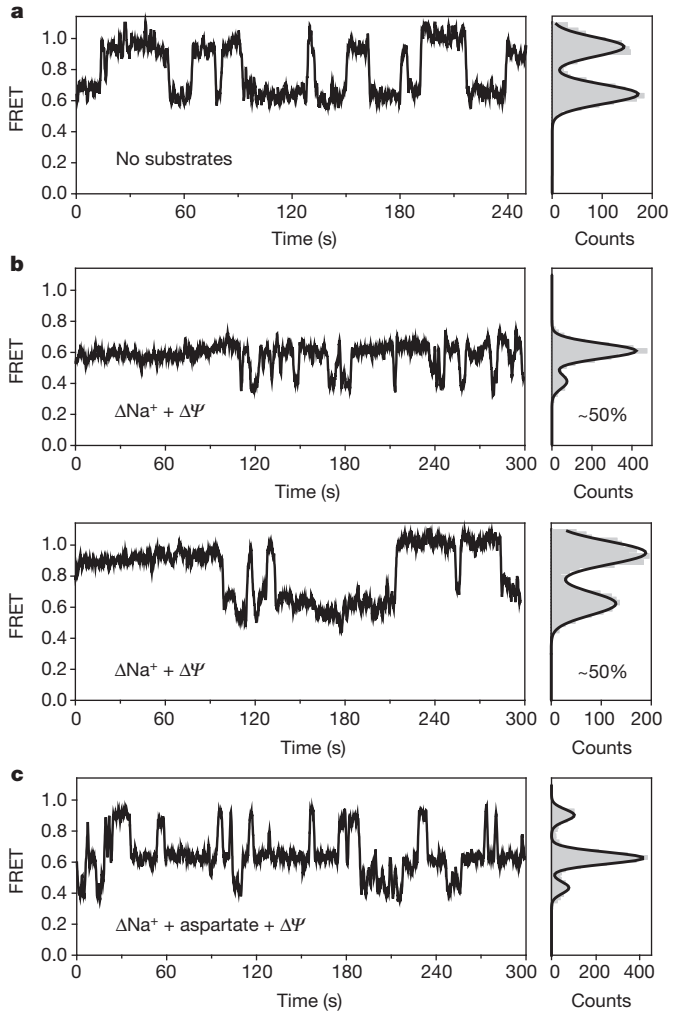


Figure 2 | Single-molecule FRET dynamics. **a**–**c**, FRET efficiencies for single Glt_{Ph} trimers without membrane gradients (**a**), with $\Delta[Na^+]$ and $\Delta\Psi$ (**b**) and with $\Delta[Na^+]$, $\Delta\Psi$ and an aspartate gradient (**c**). A 10-frame moving average filter was applied to the traces to reduce noise. In total, 30 FRET traces were analysed with a combined length of 2.2 h.

asymmetric branching ratios that we experimentally observed (Fig. 4b). As a starting point, we took the simplest set of rules to describe the behaviour of the individual subunits: entirely uncoordinated and stochastic conformational changes, in which the individual subunits move from the inward-facing to the outward-facing conformation through the intermediate conformation. We then simulated the Glt_{Ph} transport cycle using only the conformational dwell times for individual subunits as an input. We could uniquely assign the high FRET efficiency to a conformation with the two labelled subunits both in the inward-facing conformation based on the distances we calculated from the crystal structures (see Supplementary Information). In this configuration, the only transition that gives rise to a change in FRET efficiency is the transition to inward-facing/intermediate (see Supplementary Information). The average dwell time for the highest FRET level (Fig. 4c) is therefore directly related to the probability of a subunit changing from the inward-facing to the intermediate conformation and thus can be used directly as an input parameter for our model.

Using this information, we then calculated theoretical branching ratios and average FRET level dwell times with various conformational dwell times as an input, to test the consistency of our model with the experimental data. As listed in Table 1 we could closely reproduce the FRET level dwell times and branching ratios we observed in our experiments. Importantly, our model robustly predicts a strong preference for Glt_{Ph} to go to the lowest FRET level after visiting the intermediate FRET level, regardless of the exact conformational dwell times that we used as an input.

Our single-molecule data agree well with a simple model for the Glt_{Ph} transport mechanism, in which the subunits transport aspartate uncoordinated and independent of each other. Nevertheless, this observation does not formally exclude other models with a certain degree of coordination within the trimer. We have, however, demonstrated that the Glt_{Ph} subunits can be in different conformations during aspartate transport (Fig. 3). A possible model with subunits that are in different conformations but move in a coordinated fashion could include a rotary coordination, in which each subunit is one or more steps out of phase in the transport cycle compared to its neighbours. A consequence of rotary coordination is that conformational changes in the subunits are coupled and always occur simultaneously and in a specific order. Any transition between configurations with the same FRET levels that is repeated during the transport cycle would in such a situation no longer be described by a single-exponential decay, but by a rise-and-decay function[17]. Such

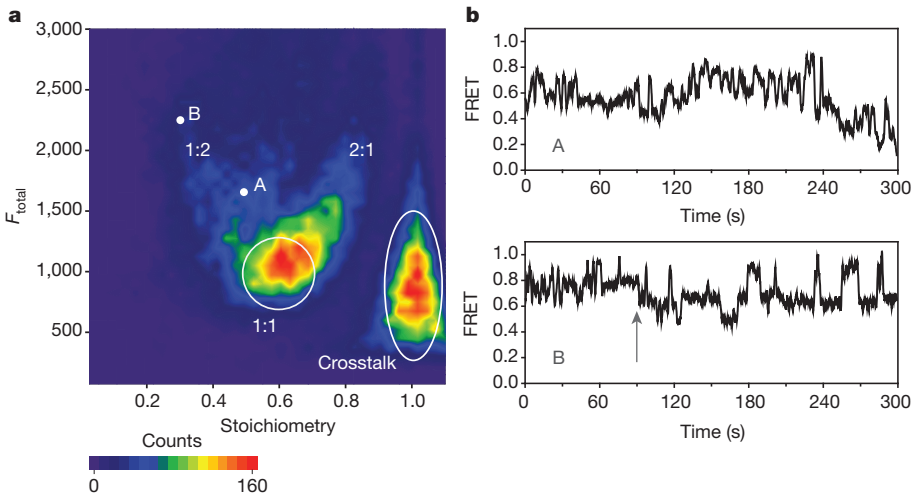


Figure 3 | Independent subunit dynamics within Glt_{Ph} trimmers. **a**, A two-dimensional histogram with the donor:acceptor fluorescence stoichiometry on the horizontal axis and the total donor/acceptor fluorescence on the vertical axis (see Supplementary Information for more details). The white circle indicates the area where 1:1 labelled complexes can be found with high confidence. **b**, Single-molecule FRET dynamics for the trimers that are marked as 'A' and 'B' in panel **a** containing one donor and two acceptor dyes. Marked by the arrow is the point where one of the acceptor dyes presumably undergoes photobleaching; from this point onwards the FRET dynamics appear to be the same as for other 1:1 labelled trimers (see Fig. 2).

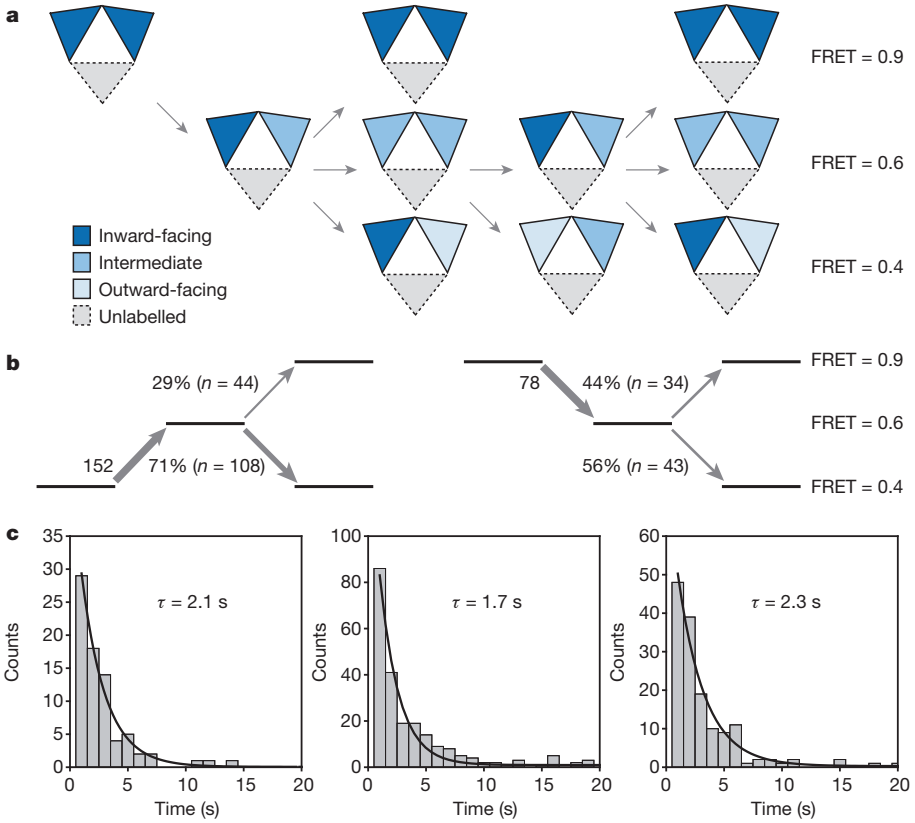


Figure 4 | Branching ratios and dwell times for various Glt_{Ph} conformations. **a**, Visual representation of a part of the kinetic model that describes unsynchronised conformational transitions of individual subunits. Starting from the high FRET level (0.9; uniquely corresponding to two subunits in the inward-facing conformation), the trimer progresses through the different FRET levels in a manner that is determined by the stochastic transition of individual subunits. A more detailed description can be found in the supplementary information. **b**, The diagram represents the number of times a certain FRET level transition is observed in Glt_{Ph} when coming from the lowest (left) or highest (right) FRET level. **c**, Distribution of dwell times of the highest (left), intermediate (middle) and lowest (right) FRET level. The exponential lifetime from a single-exponential fit is indicated.

rise-and-decay features are absent in our experimental FRET-level dwell-time distributions (Fig. 4c). Furthermore, a simultaneous conformational change in all three Glt_{Ph} subunits would give rise to identical transition kinetics in the 1:1 and 2:1/1:2 labelled molecules. The much faster transition kinetics in the 2:1/1:2 complexes (Fig. 3b) precludes any such coordination.

A recent[15] single-molecule study of detergent-solubilised Glt_{Ph} showed long periods without protein dynamics followed by short bursts of fast dynamics. We did not observe such on-and-off behaviour in our single-molecule FRET experiments. The discrepancy may be caused by differences in the environment of the membrane proteins (detergent or lipid bilayer) or by the method with which the proteins were immobilized during the experiments. In the recent study detergent-solubilised Glt_{Ph} was anchored directly to the coverslip, whereas in our experiments the immobilization was lipid-mediated, without additional protein modification. Further study is needed to reconcile and understand these differences.

We have directly observed the transport dynamics in a membrane-reconstituted transporter at the single-molecule level. These observations allow us to conclude that the Glt_{Ph} subunits move independently and in an unsynchronised fashion during aspartate transport. Our results provide a mechanistic explanation for biochemical experiments on mammalian members from the EAAT family that indicated independent transport activity of the individual subunits[18–20]. Our data also suggest a previously unknown but important function for the intermediate conformation of Glt_{Ph}. In the absence of aspartate, the transporter cycles between either the inward-facing and intermediate conformation, or the outward-facing and intermediate conformation. Under these conditions the dynamics of the conformational changes are on the same timescale as the dynamics during aspartate transport, indicating a relatively small difference in free energy between the conformations that are probed by Glt_{Ph}, as proposed before[9,10]. The intermediate conformation might therefore be regarded as barrier that prevents Glt_{Ph} from probing both the inward- and the outward-facing conformation in the absence of aspartate. Such a mechanism is not trivial; continuous switching between the inward- and outward-facing conformations in the absence of aspartate could result in dissipation of the $\Delta[Na^+]$. The helical hairpin motifs that act as lids for the occlusion of the aspartate binding site may provide the structural basis for this barrier. Incomplete closure of the binding site lids could result in a conformation that is incompatible with the full transport cycle. Further structural and functional studies are needed to test this hypothesis.

Table 1 | Experimental parameters compared with calculations based on the model

	Model input	Calculation	Experimental
$\tau_{I\to X}$	4.0 s	4.2 s	4.2 s
$\tau_{X\to I}$	4.0 s	–	–
$\tau_{O\to X}$	2.0 s	–	–
$\tau_{X\to O}$	4.0 s	–	–
$\tau_{0.9}$	–	2.0 s	2.1 s
$\tau_{0.6}$	–	2.3 s	1.7 s
$\tau_{0.4}$	–	1.7 s	2.3 s
R_1	–	0.89	0.79
R_2	–	0.44	0.41

τ indicates the average dwell time of the state preceding the specific conformational transition (I, inward-facing; X, intermediate and O, outward-facing) or FRET level. $\tau_{0.9}$, $\tau_{0.6}$ and $\tau_{0.4}$ are the average dwell times for the three FRET levels. R_1 is the branching ratio for the intermediate FRET level (high FRET/low FRET) coming from the highest FRET level, R_2 is the branching ratio for the intermediate FRET level (high FRET/low FRET) coming from the lowest FRET level.

METHODS SUMMARY

Glt_{Ph}-His_8 was expressed in *Escherichia coli* and purified using Ni-Sepharose chromatography. While immobilized on the Ni-Sepharose column material, the protein was labelled overnight with a mixture of Alexa Fluor 555 and Alexa Fluor 647 maleimide. Next, labelled Glt_{Ph} was further purified using size-exclusion chromatography (SEC) and directly used for reconstitution at a 1:10,000 protein: lipid ratio in liposomes containing 1% (w/w) biotinylated lipids. The proteoliposomes (extruded through a 100-nm pore-size filter) were subsequently immobilized on biotinylated microscopy coverslip in a home-built flow-cell that was pre-incubated with streptavidin. The acquisition was started shortly after flushing in activation buffer (desired buffer composition + 1 mM Trolox and enzymatic oxygen scavenging system[21]). Each experiment started with a short (10 s) period of alternating donor and acceptor excitation (using 532 and 637 nm laser excitation, respectively) that was later used to calculate the donor:acceptor stoichiometry, followed by a 300 s acquisition during which only the donor dyes were excited. Donor and acceptor fluorescence were simultaneously recorded on an EM-CCD camera with a frame rate of 10 Hz. From each acquisition, peaks were selected in the acceptor channel (indicative of FRET) that were dynamic and had a corresponding fluorescence signal in the donor channel that showed an anti-correlation with the acceptor fluorescence dynamics. We then selected the 1:1 labelled Glt_{Ph} trimers based on their donor:acceptor fluorescence stoichiometry and total fluorescence. FRET level dwell-times were determined using a home-written algorithm, where we used a minimum dwell-time length of 5 frames (0.5 s).

Full Methods and any associated references are available in the online version of the paper.

Received 5 April; accepted 13 August 2013.

1. Kanner, B. I. & Zomot, E. Sodium-coupled neurotransmitter transporters. *Chem. Rev.* **108,** 1654–1668 (2008).
2. Groeneveld, M. & Slotboom, D. J. Na^+:aspartate coupling stoichiometry in the glutamate transporter homologue Glt_{Ph}. *Biochemistry* **49,** 3511–3513 (2010).
3. Reyes, N., Ginter, C. & Boudker, O. Transport mechanism of a bacterial homologue of glutamate transporters. *Nature* **462,** 880–885 (2009).
4. Ryan, R. M., Compton, E. L. & Mindell, J. A. Functional characterization of a Na^+-dependent aspartate transporter from *Pyrococcus horikoshii. J. Biol. Chem.* **284,** 17540–17548 (2009).
5. Yernool, D. *et al.* Structure of a glutamate transporter homologue from *Pyrococcus horikoshii. Nature* **431,** 811–818 (2004).
6. Boudker, O. *et al.* Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445,** 387–393 (2007).
7. Ryan, R. M. & Mindell, J. A. The uncoupled chloride conductance of a bacterial glutamate transporter homolog. *Nature Struct. Mol. Biol.* **14,** 365–371 (2007).
8. Verdon, G. & Boudker, O. Crystal structure of an asymmetric trimer of a bacterial glutamate transporter homolog. *Nature Struct. Mol. Biol.* **19,** 355–357 (2012).
9. Georgieva, E. R. *et al.* Conformational ensemble of the sodium-coupled aspartate transporter. *Nature Struct. Mol. Biol.* **20,** 215–221 (2013).
10. Hänelt, I. *et al.* Conformational heterogeneity of the aspartate transporter Glt_{Ph}. *Nature Struct. Mol. Biol.* **20,** 210–214 (2013).
11. Joo, C. *et al.* Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.* **77,** 51–76 (2008).
12. Verhalen, B. *et al.* Dynamic ligand-induced conformational rearrangements in P-glycoprotein as probed by fluorescence resonance energy transfer spectroscopy. *J. Biol. Chem.* **287,** 1112–1127 (2012).
13. Zhao, Y. *et al.* Substrate-modulated gating dynamics in a Na^+-coupled neurotransmitter transporter homologue. *Nature* **474,** 109–113 (2011).
14. Zhao, Y. *et al.* Single-molecule dynamics of gating in a neurotransmitter transporter homologue. *Nature* **465,** 188–193 (2010).
15. Akyuz, N. *et al.* Transport dynamics in a glutamate transporter homologue. *Nature* http://dx.doi.org/10.1038/nature12265 (23 June 2013).
16. Dorwart, M. R. *et al. S. aureus* MscL is a pentamer *in vivo* but of variable stoichiometries *in vitro*: implications for detergent-solubilized membrane proteins. *PLoS Biol.* **8,** e1000555 (2010).
17. Floyd, D. L., Harrison, S. C. & van Oijen, A. M. Analysis of kinetic intermediates in single-particle dwell-time distributions. *Biophys. J.* **99,** 360–366 (2010).
18. Koch, H. P. & Larsson, H. P. Small-scale molecular motions accomplish glutamate uptake in human glutamate transporters. *J. Neurosci.* **25,** 1730–1736 (2005).
19. Grewer, C. *et al.* Individual subunits of the glutamate transporter EAAC1 homotrimer function independently of each other. *Biochemistry* **44,** 11913–11923 (2005).
20. Koch, H. P., Brown, R. L. & Larsson, H. P. The glutamate-activated anion conductance in excitatory amino acid transporters is gated independently by the individual subunits. *J. Neurosci.* **27,** 2943–2947 (2007).
21. Blanchard, S. C. *et al.* tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101,** 12893–12898 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to thank M. Punter for developing software and algorithms for data analysis, V. Krasnikov for his help with the design, construction and maintenance of the single-molecule fluorescence microscopes and I. Küsters for sharing experience on liposome tethering and single-molecule imaging of membrane proteins. A.M.v.O. acknowledges funding from the Netherlands Organization for Scientific Research (NWO; *Vici* 680-47-607) and the European Research Council (ERC Starting 281098). D.J.S. acknowledges funding from the Netherlands Organization for Scientific Research (NWO; *Vidi* 700.54.423, *Vici* 865.11.001) and the European Research Council (ERC Starting 282083). I.H. acknowledges the Deutsche Forschungsgemeinschaft (HA 6322/1-1) for providing funding.

Author Contributions G.B.E., I.H., J.M.H.G., D.J.S. and A.M.v.O. designed experiments, G.B.E. and I.H. performed experiments, J.M.H.G. performed the FRET level transition calculations, G.B.E., D.J.S. and A.M.v.O. wrote the manuscript, all authors contributed to the interpretation of the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.M.v.O. (a.m.van.oijen@rug.nl) and D.J.S. (d.j.slotboom@rug.nl).

METHODS

Purification, labelling and reconstitution of Glt_{Ph}–His. For all experiments, we used the Glt_{Ph}–7H–His_8 derivate[5] with additional mutations C321S and S331C (introduced with standard cloning techniques). Glt_{Ph} was expressed in *Escherichia coli* MC1061 cells and purified Na^+-free as described[2,10]. While immobilized on Ni-sepharose (GE healthcare), Glt_{Ph} was labelled with a mixture of ~80 μM donor (Alexa Fluor 555 C_2-maleimide, Life technologies) and ~80 μM acceptor (Alexa Fluor 647 C_2 maleimide, Life technologies) dissolved in 50 mM KP_i, pH 7.0, 300 mM KCl, 0.04% dodecyl maltoside (DDM, Anatrace) (buffer A) at 4 °C, for 20 h. After this step, unreacted dye molecules were removed by washing the column with 20 column volumes of buffer A and labelled Glt_{Ph} was eluted in 500 μl fractions with buffer A + 500 mM imidazole. To obtain a monodisperse protein fraction and to remove remaining unreacted dye molecules, the second Glt_{Ph} elution fraction was further purified using size-exclusion chromatography (SEC) on a Superdex-200 column (GE healthcare), equilibrated with buffer A. The peak fraction from SEC was used immediately for reconstitution into liposomes.

Liposomes were prepared from synthetic lipids and had a final composition of 40% (w/w) 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC), 29% (w/w) 1,2-dioleoyl-*sn*-glycero-3-phosphoethanolamine (DOPE), 30% (w/w) 1,2-dioleoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) (DOPG) and 1% (w/w) 1,2-dioleoyl-*sn*-glycero-3-phosphoethanolamine-N-(cap biotinyl) (biotin-DOPE) and were suspended in 50 mM KP_i, pH 7.0. Preparation of the liposomes and protein reconstitution was performed according to previously published procedures[22]. Glt_{Ph} was reconstituted in a 1:10,000 protein:lipid (w/w) ratio to ensure a high probability of having only a single Glt_{Ph} trimer per liposome. Assuming that a 100-nm liposome consists of 80,000 lipids with an average molecular mass of 750 Da, we calculated that approximately 95% of the liposomes should be empty with this reconstitution ratio. The probability of finding two Glt_{Ph} trimers in a single liposome would thus be $0.05^2 = 0.0025$ (0.25%). After reconstitution, the proteoliposomes were subjected to four freeze/thaw cycles and stored in liquid nitrogen.

Single-molecule fluorescence microscopy. Microscope coverslips (no. 1.5 precision coverslips, Karl Roth) were extensively cleaned and functionalised with PEG-5000 and Biotin-PEG-5000 (Laysan Bio) and subsequently used to construct a flow-cell as described previously[23]. Prior to use, the flow-cell was incubated with 50 mM KP_i, pH 7.0 (buffer B) + 100 μg ml^{-1} streptavidin (streptavidin from *Streptomyces avidinii*, Sigma-Aldrich) for 5 min. Unbound streptavidin was subsequently washed out with buffer B. Next, a suspension of proteoliposomes (lipid concentration 200 μg ml^{-1}) extruded through a 100-nm pore-size filter was flushed in, followed by a 5-min incubation to allow the liposomes to adhere to the surface. Before starting the experiment, unbound liposomes were washed out with buffer B.

Image acquisition was started shortly (<1 min) after the flow-cell was equilibrated with activating buffer. Depending on the nature of the experiment, we used the following compositions for the activating buffer: 50 mM KP_i, pH 7.0 (no substrates), 50 mM NaP_i, pH 7.0 (Na^+ gradient) and 50 mM NaP_i, pH 7.0 + 100 μM Na-aspartate. In addition, the activating buffer contained the following components: 0.5 μM valinomycin (to establish a membrane potential where appropriate), 1 mM Trolox (antiblinking agent, prepared and aged according to a previously published procedure[24]), the GODCAT oxygen scavenging system[21] consisting of 7.5 U ml^{-1} glucose oxidase (glucose oxidase from *Aspergillus niger* type VII, Sigma-Aldrich), 450 U ml^{-1} catalase (catalase from bovine liver, Sigma-Aldrich) and 0.8% (w/v) glucose monohydrate. Immediately before the experiments, the activating buffer was degassed. Two consecutive experiments were performed on a single flow-cell.

Glt_{Ph} was visualized using TIRF microscopy on an Olympus IX-71 inverted fluorescence microscope equipped with a high numerical aperture ×100 TIRF objective (Olympus) and a ×1.6 additional magnification insert. Dye molecules were excited with 532 nm (donor) or 637 nm (acceptor) lasers at 100–150 W cm^{-2} and the resulting fluorescence signal was imaged on an EM-CCD camera (Hamamatsu Photonics) at a frame rate of 10 Hz. The donor and acceptor fluorescent signals were separated and co-projected on the EM-CCD chip using a Dual View simultaneous imaging system (Photometrics).

Selecting Glt_{Ph} trimers labelled with a single donor and acceptor. At the start of each experiment, a 10 s (100 frames) movie was recorded with alternating 532-nm and 637-nm excitation (2 Hz alternation rate, 10 Hz frame rate). This alternating laser excitation (ALEX) scheme allowed us to determine the separate fluorescence contributions from both the green and red dyes and thus allowed us to identify those complexes that had only a single green and a single red dye[25]. In the 532-nm excitation frames, co-localizing peaks (indicative of FRET) were identified using a discoidal peak-finding algorithm (see later). In each frame, and for each peak pair, two values were calculated: the total fluorescence (F_{total}) and the stoichiometry (S) according to the following equations:

$$F_{total} = F_{DD} + F_{DA} + F_{AA}$$

$$S = (F_{DD} + F_{DA})/(F_{DD} + F_{DA} + F_{AA})$$

where F_{DD} is the donor fluorescence upon donor excitation, F_{DA} is the acceptor fluorescence upon donor excitation and F_{AA} is the acceptor fluorescence upon acceptor excitation. Using these data, a three-dimensional histogram was constructed with the values for F_{total} and S on separate axes (Supplementary Fig. 1a). The histograms show the existence of three major populations separated by their S and F_{total} values: 1:1 donor:acceptor labelled trimers and 1:2 or 2:1 labelled trimers. In the next step, we placed a cut-off for the F_{total} values that roughly separates the 1:1 labelled pool from the 1:2 or 2:1 labelled pools of trimers. However, the relatively large variation of the F_{total} and S values within each pool did not allow us to exclude non-1:1 labelled trimers based on the F_{total} values alone. To get a more confident assignment, all S values from either below or above the F_{total} cut-off were used to assemble two new histograms (Supplementary Fig. 1b). Both histograms should now represent a linear combination of four distributions: a distribution with an average S value of 1.0, corresponding to crosstalk from donor fluorescence in the acceptor channel, and three distributions with average S values corresponding to a 1:1, 2:1 or 1:2 donor:acceptor labelling stoichiometry. In the next step, both histograms were fitted with a sum of four Gaussian distributions. The same set of distributions was used to fit both histograms and only the amplitude was varied. From these fits, the relative contribution of the 1:1 labelled population was calculated for every value of S. As a criterion for assigning a FRET signal to a 1:1 donor:acceptor stoichiometry, we used a stoichiometry window of at least half the σ-value (from the Gaussian fit) positioned such that it maximises the relative contribution of the 1:1 stoichiometry peak. In this way we could assign each FRET signal with a confidence of 75–90% to a 1:1 labelled Glt_{Ph} trimer. Overall, we observed a clear difference between the 1:1 and 2:1 or 1:2 labelled populations in terms of the kinetics and number of FRET levels (similar to the differences in Fig. 3) that are visited.

Data analysis. A typical FRET experiment resulted in a 3,000-frame (300 s) image stack. First, the effects of microscopy stage drift were removed from the stack by applying Fourier correlation[26,27] of the first image with every subsequent image in the stack. Using this approach, every frame was translated a certain number of pixels in the x or y direction to overlay perfectly with the first frame. To correct for fluorescence background and the electronic offset from the EM-CCD camera, an image was created with the average values from the whole stack. This image was converted in a background-mask that was subtracted from every individual frame by applying a median filter with a 24-pixel radius three times consecutively. To find the position of the fluorescently labelled Glt_{Ph} molecules, we applied a discoidal filtering algorithm[28] on an image with the average pixel values from the whole image stack. We used a threshold of at least two standard deviations above the average pixel value for the whole image to select peaks in the acceptor channel. Although these settings lead to the selection of a significant amount of non-FRET peaks (that is, crosstalk of donor-only labelled Glt_{Ph} in the acceptor channel, Supplementary Fig. 1a) we could in this way also detect donor-acceptor pairs with low FRET efficiencies, or pairs for which one of the molecules were photobleached after only part of the 300 s acquisition time.

For every selected peak, the average fluorescence of a 7 × 7 pixel box was plotted over time and dynamic acceptor fluorescence signals were selected manually. The subset of selected donor traces was subsequently checked for anti-correlation with the donor fluorescence, before calculating the FRET efficiency according to the following equation:

$$E_{FRET} = F_A/(F_A + F_D)$$

Where F_A is the fluorescence signal in the acceptor channel (corrected for background) and F_D is the fluorescence in the donor channel (corrected for background). Traces that showed for example very rapid photobleaching, interference of strongly fluorescent neighbouring peaks, or additional molecules (transiently) absorbing to the surface were also excluded at this point.

To extract dwell times from the FRET traces we used the following procedure: first, the exact average values and number of FRET levels were determined for each individual trace by fitting the distribution of FRET values with a sum of Gaussian distributions. This information was subsequently used to categorise each data point in the FRET trace according to one of the average FRET levels. For example, in the case of two FRET levels, each data point was associated with one of these two levels, depending on which one is closest to the value of the data point. In the next step, all neighbouring points that were associated with the same FRET level were joined to a single step. Finally, to exclude short dwell times as a result of fluctuations in the noise, we set a minimum step length of 5 data points. All steps shorter than 5 points were joined with a neighbouring step that had a FRET level that was closest to their average value.

22. van der Heide, T. & Poolman, B. Osmoregulated ABC-transport system of *Lactococcus lactis* senses water stress via changes in the physical state of the membrane. *Proc. Natl Acad. Sci. USA* **97,** 7102–7106 (2000).
23. Tanner, N. A. & van Oijen, A. M. Visualizing DNA replication at the single-molecule level. *Methods Enzymol.* **475,** 259–278 (2010).
24. Cordes, T., Vogelsang, J. & Tinnefeld, P. On the mechanism of Trolox as antiblinking and antibleaching reagent. *J. Am. Chem. Soc.* **131,** 5018–5019 (2009).
25. Kapanidis, A. N. *et al.* Alternating-laser excitation of single molecules. *Acc. Chem. Res.* **38,** 523–533 (2005).
26. Guizar-Sicairos, M., Thurman, S. T. & Fienup, J. R. Efficient subpixel image registration algorithms. *Opt. Lett.* **33,** 156–158 (2008).
27. Reddy, B. S. & Chatterji, B. N. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Process.* **5,** 1266–1271 (1996).
28. Hedde, P. N. *et al.* Online image analysis software for photoactivation localization microscopy. *Nature Methods* **6,** 689–690 (2009).

doi:10.1038/nature12588

Vinylogous chain branching catalysed by a dedicated polyketide synthase module

Tom Bretschneider[1], Joel B. Heim[2], Daniel Heine[1], Robert Winkler[1], Benjamin Busch[1], Björn Kusebauch[1], Thilo Stehle[2,3], Georg Zocher[2] & Christian Hertweck[1,4]

Bacteria use modular polyketide synthases (PKSs) to assemble complex polyketides, many of which are leads for the development of clinical drugs, in particular anti-infectives and anti-tumoral agents[1]. Because these multifarious compounds are notoriously difficult to synthesize, they are usually produced by microbial fermentation. During the past two decades, an impressive body of knowledge on modular PKSs[2,3] has been gathered that not only provides detailed insight into the biosynthetic pathways but also allows the rational engineering of enzymatic processing lines to yield structural analogues[4,5]. Notably, a hallmark of all PKS modules studied so far is the head-to-tail fusion of acyl and malonyl building blocks, which leads to linear backbones. Yet, structural diversity is limited by this uniform assembly mode. Here we demonstrate a new type of PKS module from the endofungal bacterium *Burkholderia rhizoxinica* that catalyses a Michael-type acetyl addition to generate a branch in the carbon chain. *In vitro* reconstitution of the entire PKS module, X-ray structures of a ketosynthase-branching didomain and mutagenesis experiments revealed a crucial role of the ketosynthase domain in branching the carbon chain. We present a trapped intermediary state in which acyl carrier protein and ketosynthase are covalently linked by the branched polyketide and suggest a new mechanism for chain alkylation, which is functionally distinct from terpenoid-like β-branching. For the rice seedling blight toxin rhizoxin, one of the strongest known anti-mitotic agents, the non-canonical polyketide modification is indispensable for phytotoxic and anti-tumoral activities. We propose that the formation of related pharmacophoric groups follows the same general scheme and infer a unifying vinylogous branching reaction for PKS modules with a ketosynthase-branching–acyl-carrier-protein architecture. This study unveils the structure and function of a new PKS module that broadens the biosynthetic scope of polyketide biosynthesis and sets the stage for rationally creating structural diversity.

In light of the highly diverse architectures of polyketide metabolites, it is surprising to learn that their carbon backbones are produced by rather basic enzymatic mechanisms that are reminiscent of fatty acid biosynthesis[6,7]. The minimal requirement is an acyl carrier protein (ACP) that serves as an anchor for the growing chain, an acyl transferase (AT) that loads activated malonyl building blocks onto the ACP and a ketosynthase (KS) that catalyses the head-to-tail carbon bond formation between the malonyl and acyl units tethered to the KS and ACP. Ketoreductase, dehydratase and enoyl reductase domains may optionally be involved in processing the β-keto group that results from the KS-mediated Claisen condensation[6]. Apart from the number of chain propagations and the degree of β-keto processing, alkyl side chains largely contribute to polyketide diversity. Alkyl substituents at α-positions to a former carbonyl residue may result from the incorporation of substituted malonyl units or α-methylation[8]. Furthermore, terpenoid-like alkylations at β-positions have been observed in various polyketide pathways, in which a set of freestanding enzymes act *in trans* to modify the β-keto group in analogy to mevalonate biosynthesis[9] (Fig. 1c).

Such polyketide chain branches may play a key part in the biological activity of polyketides. A particularly noteworthy example is rhizoxin (**1b**), the phytotoxin produced in a rare fungal–bacterial symbiosis of the rice seedling blight fungus *Rhizopus microsporus* and its endofungal bacterium *B. rhizoxinica*[10,11]. Owing to its efficient binding to β-tubulin subunits, rhizoxin congeners represent some of the strongest anti-mitotic agents known and are considered as promising anti-tumoral agents[12]. Yet, structure–activity relationships and modelling studies have revealed that the anti-mitotic activity of the molecule crucially depends on the short (C2) carbon chain that branches off from the macrolide ring at C6 (refs 13–15). The formation of this pharmacophoric residue could not be rationalized by any biosynthetic precedence as genes for terpenoid-like β-alkylation were absent in the bacterial genome[16,17]. Analysis of the gene cluster coding for the rhizoxin D1 (**1a**) polyketide synthase[18] (Fig. 1a and Supplementary Fig. 1) and structural elucidation of prematurely released pathway intermediates strongly suggested that the β-branch is introduced by a Michael addition of a C2 unit to an α,β-unsaturated thioester[19] (Fig. 1b).

Bioinformatic analyses and *in vivo* dissection of the rhizoxin pathway indicated that an unusual PKS module located on RhiE would have a role in the chain-branching event[19]. The designated module consists of a typical KS domain, an ACP domain and a cryptic branching ('B') domain in between, which does not show any homology to known enzyme domains. To determine the essential components for polyketide chain branching and to gain insight into the reaction mechanism, we fully reconstituted the PKS module *in vitro*. Reaching this goal was challenging because it afforded a variety of pure, functional and, in part, post-translationally modified proteins as well as the appropriate substrates (Fig. 2A and Supplementary Information). In short, we heterologously produced tagged ketosynthase-branching (KS-B) didomain (RhiE$^{\#}$) of the branching module, the requisite ACP from the same module (Fig. 2B), which was transformed into its *holo* form by an endogenous phosphopantetheine transferase (PPTase) encoded in the *B. rhizoxinica* genome[17] (Fig. 2C), and the *trans*-acting AT (RhiG) for loading malonyl units onto the ACP. The module was split into KS-B and ACP domains to allow for flexible substrate loading and for analysis of ACP-bound products. To provide a substrate for the enzyme assay, we synthesized *N*-acetylcysteamine (SNAC) thioester **2**, a truncated surrogate of the polyketide intermediate produced upstream of the branching module. The enzymatic chain branching reaction was performed *in vitro* by mixing the His-tagged KS-B didomain (His-KS-B), His-tagged *holo*-ACP (His-ACP), His-tagged AT (AT-His), malonyl-CoA, and the synthetic surrogate substrate **2**. By matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) analysis of the ACP adduct (**3**) we detected a new compound attached to the ACP with the expected molecular mass. Product formation was confirmed by liquid chromatography–mass spectrometry (LC–MS) analysis of the product (Fig. 2D) liberated by the PPant ejection method[20]. To confirm unequivocally the identity of the lactone (**4**), we synthesized a synthetic

[1]Department of Biomolecular Chemistry, Leibniz Institute for Natural Product Research and Infection Biology (HKI), Jena 07745, Germany. [2]Interfaculty Institute of Biochemistry, Eberhard Karls University Tübingen, Tübingen 72076, Germany. [3]Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [4]Chair for Natural Product Chemistry, Friedrich Schiller University, Jena 07737, Germany.

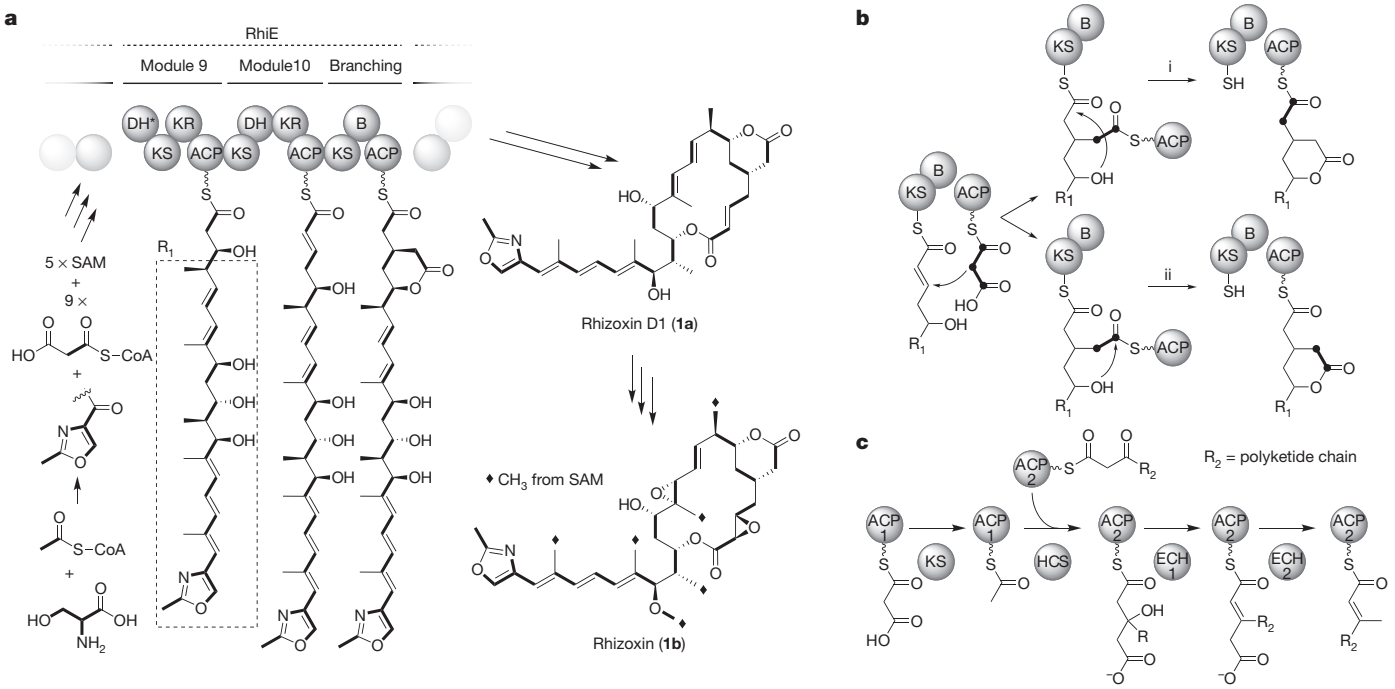


Figure 1 | Model of rhizoxin biosynthesis and chain branching mechanisms. **a**, Type I PKS for assembly of the rhizoxin backbone and structure of rhizoxin D1 (**1a**) (the box designates residue R_1 in **b**). ACP, acyl carrier protein; AT, acyl transferase; B, branching domain; DH, dehydratase; KR, ketoreductase; KS, ketosynthase. **b**, Possible course of chain-branching reaction highlighting two different lactonization routes. **c**, Isoprenoid-like β-branching reaction catalysed by *trans*-acting enzymes. ECH, enoyl-CoA hydratase/crotonase; HCS, 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase.

reference and compared it with the hydrolysed enzyme product (**4**) by high-resolution LC–MS (HRMS). Notably, the enzyme reaction performed with heat-inactivated His-KS-B did not yield any product (Fig. 2E), thus proving the catalytic activity of the PKS module.

To gain deeper insight into this unusual reaction we studied the structure of the non-canonical PKS module. For improved protein stability we generated a truncated version of His-KS-B (RhiE*), which crystallized and led to a structural model of the KS-B module at 2.14 Å resolution (Fig. 3 and Supplementary Table 1). The KS-B module is a dimeric protein and folds into three domains (Fig. 3a and Supplementary Fig. 3). The KS domain shows a typical thiolase fold[21] and is flanked by a small linker region. A long linker runs across the entire KS-B RhiE module (Fig. 3a) and connects the *N*-terminal domains with the B domain. The B domain itself shares no noteworthy sequence homology (sequence identity is below 15%) to any known PKS domains described so far, but features a double hot dog (DHD) fold that is present in dehydratase domains[22]. A structural comparison (Supplementary Fig. 4) revealed highest structural homology to the product template domain (dehydratase-fold domain) of a fungal PKS[23] (Supplementary Table 2).

Binding of the substrate mimic **2** to the KS was confirmed by LC–MS. Moreover, X-ray crystallography revealed positive electron density in one chain next to the catalytic cysteine of the active site of the KS (Fig. 3c), thus suggesting that the KS domain is involved in the branching reaction.

To clarify the role of the KS domain in the branching reaction we intended to study the KS domain as a standalone enzyme. However, despite various efforts it was not possible to obtain a catalytically active KS domain outside the module. Thus, as an alternative, we inspected the active site of the KS domain and generated point mutations of the catalytic triad (Cys-His-His) in the intact KS-B didomain (Supplementary Fig. 5). Specifically, we replaced the cysteine with serine and alanine, and mutated each of the two histidines into alanine. Notably, none of the mutants showed any chain branching activity in the established assay. This finding provided clear evidence that the intact active site of the KS is crucial for the β-branching reaction.

Next, we investigated the role of the B domain in chain branching. We noted that the histidine-aspartate dyad, a hallmark of typical dehydratase domains, is altered in the B domain of RhiE. Notably, point mutation of the conserved aspartate residue amino acid in the B domain (Asp3876Ala mutant) did not affect the branching activity. Furthermore, we heterologously produced a freestanding, soluble B domain. In the absence of the KS domain, however, no branching activity was observed, indicating that this domain alone is not sufficient to catalyse the Michael addition.

There are several possible routes to the enzyme-mediated course of the vinylogous attack and lactone formation. In principle, the δ-lactone ring could be formed by installing the C–C bond before esterification, or vice versa. Furthermore, in the more likely case of an initial C–C bond formation, the δ-hydroxyl group (at C5) of the thioester intermediate could potentially attack the C1 carbonyl bound to the KS (Fig. 1b, route i) or the carbonyl of the added malonyl unit bound to the ACP (Fig. 1b, route ii). The course of the reaction has implications for the resulting polyketide chain because the C2 unit introduced by the branching module would be found in either the polyketide backbone or the side chain.

To solve this riddle we performed stable isotope labelling experiments *in vitro*. After addition of ^{13}C-labelled malonyl-CoA to the enzyme mixture and the synthetic surrogate, the product (**4**) was analysed by HRMS. MS/MS fragmentation, however, did not reveal the site of isotope incorporation (**4**) because fragments with the same molecular formula can arise from different parts of the symmetric substructure. To overcome this limitation we pursued an NMR analysis of the labelled reaction product still tethered to the entire ACP domain. After performing the enzyme assay, the ACP domain was purified and subjected to NMR measurements. By this approach we observed two strong NMR signals that correspond to the C1 and C2 positions of the thioester bound to the ACP (Fig. 4a). To confirm unequivocally the structure of the ACP-bound product, we synthesized the corresponding SNAC thioester and verified the identity of the chemical shifts. Consequently, the carbons of the malonyl unit added by the branching module are found in the linear polyketide chain, whereas the C1 and C2 carbons of the precursor molecule constitute the side chain (Fig. 4a).

This reaction scheme is quite intriguing as it suggests an unprecedented model in which a polyketide intermediate would be covalently tethered to both the KS and the ACP (Fig. 4b). To test this model we

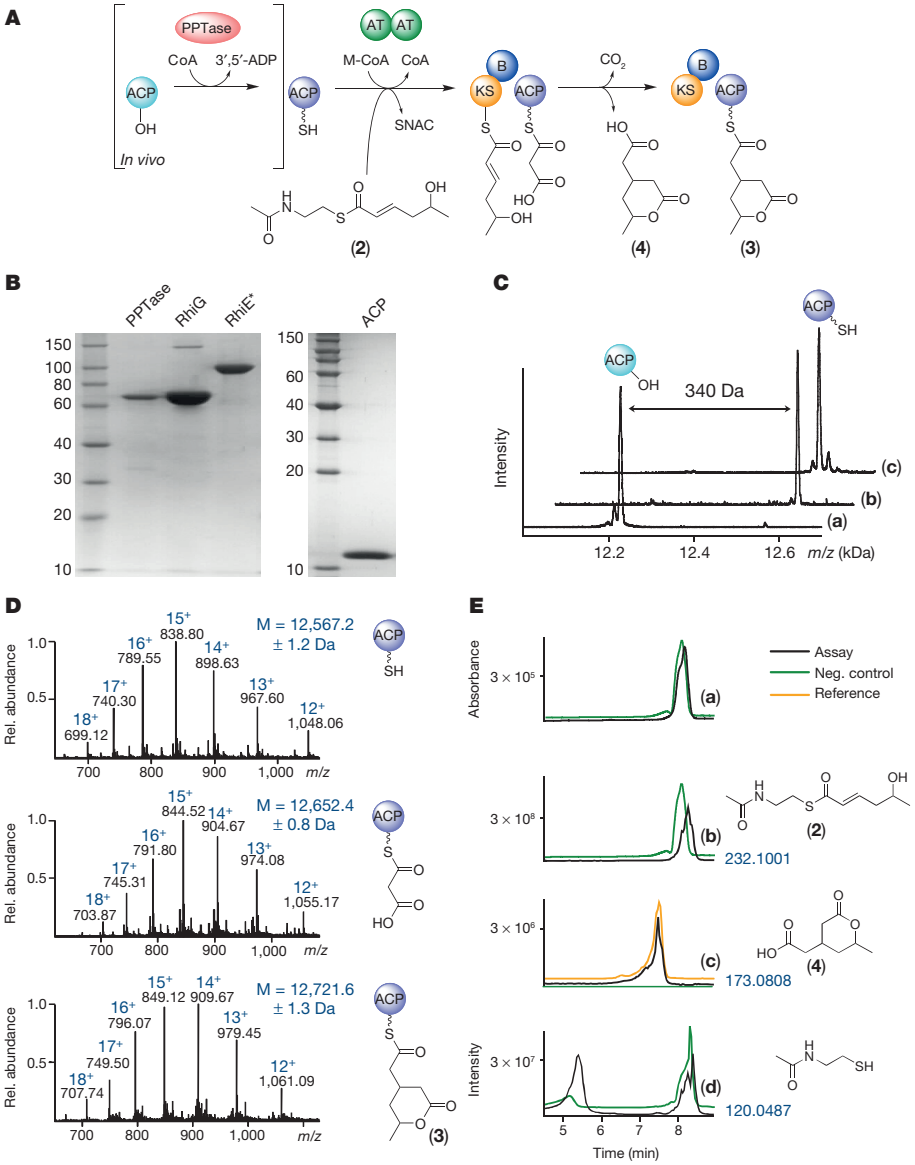


Figure 2 | *In vitro* reconstitution of the chain branching reaction. **A**, General experimental set-up for *in vitro* analysis of enzyme functions. **B**, Recombinant MalE-PPTase (70.7 kilodaltons (kDa)), RhiG-His (75.9 kDa, dimer at 152 kDa) and His-RhiE* (104.7 kDa) on a 10% SDS gel; His-ACP (12.2 kDa) on a 15% SDS gel. **C**, MALDI mass spectrum of *apo*-ACP (**a**) and *holo*-ACP, phosphopantetheinylated (wavy line) *in vitro* (**b**) and *in vivo* (**c**), indicated by mass shift (340 Da). **D**, Analysis of ACP species by LC–MS. Rel., relative. **E**, LC–HRMS analysis of enzymatic chain branching. HPLC trace ($UV_{220\,nm}$) (**a**), and extracted *m/z* chromatograms of 232.1001 (**2**) (**b**), 173.0808 (**4**) (**c**), and 120.0487 (liberated *N*-acetylcysteamine) (**d**). Retention times of diasteromers of **4** are identical. Neg., negative.

designed an experiment that allows trapping the proposed intermediary state. We reasoned that a substrate analogue lacking the hydroxyl moiety (**5**) would still undergo the Michael addition of the ACP-bound nucleophile, but the subsequent lactonization of the thioester would be prevented. For this purpose we synthesized the deoxy variant (**5**) and performed the multi-enzyme assay as described above. SDS-gel electrophoresis of the reaction mixture revealed a band that corresponds to the size of KS-B and ACP. Notably, this band is absent in an assay using **2**, where only the band for KS-B can be detected (Fig. 4c). Analysis of this band by tryptic mass fingerprinting revealed the presence of the ACP in the top band, whereas it was missing in the bottom band and the negative control (Supplementary Fig. 7). We repeated this assay with the point-mutated KS mutants and found that protein cross-linking does not take place in the Cys3227Ala, Cys3227Ser and His3404Ala KS mutants, once again confirming the crucial role of a catalytically intact KS domain. It is interesting to note that the His3364Ala mutant retains weak activity for vinylogous addition using deoxy variant **5**, whereas it cannot catalyse lactonization when surrogate **2** is used (see above). Taken together, the KS clearly has a key role in forming an enolate for the vinylogous addition to the unsaturated thioester. Chain propagation is only possible by lactonization involving the δ-hydroxy group of the resulting intermediate and the C1 carbonyl, and C1 and C2 of the former linear chain constitute the side chain. It is generally conceivable that the unusual B domain assists in the formation of the δ-lactone ring. However, an interaction of the B domain with the KS-ACP-linked intermediate is rather unlikely given the large distance between the active site cavities of the KS and B domains (Fig. 3). Because mutation of the conserved Asp residue resulted in an enzyme species that is still able to catalyse the vinylogous branching reaction, one may conclude that the B domain seems to have a rather structural role in the module.

A vast number of modular PKS systems have been investigated at the molecular and biochemical levels. Yet, in all studied examples PKS modules catalyse the head-to-tail fusion of acyl and malonyl units, thus invariably yielding a linear polyketide chain[6]. Our results show, for the first time, direct evidence for a vinylogous addition of a malonyl building block to a polyketide chain, which results in chain branching (Fig. 4b). This non-canonical branching reaction is catalysed by a novel type of module consisting of KS, B and ACP domains, as shown by the successful *in vitro* reconstitution of the entire module and transformation of a substrate mimic. As a proof of concept, we used NMR analysis of an ACP-bound, branched reaction product to complement mass spectrometry and MALDI analyses. The results reported here will have broad implications for the field of polyketide biosynthesis. First, we unveil a novel role for a ketosynthase in polyketide biosynthesis[24,25]. As point mutations and the chemical cross-linking reaction revealed, a catalytically active KS domain is indispensable for the vinylogous

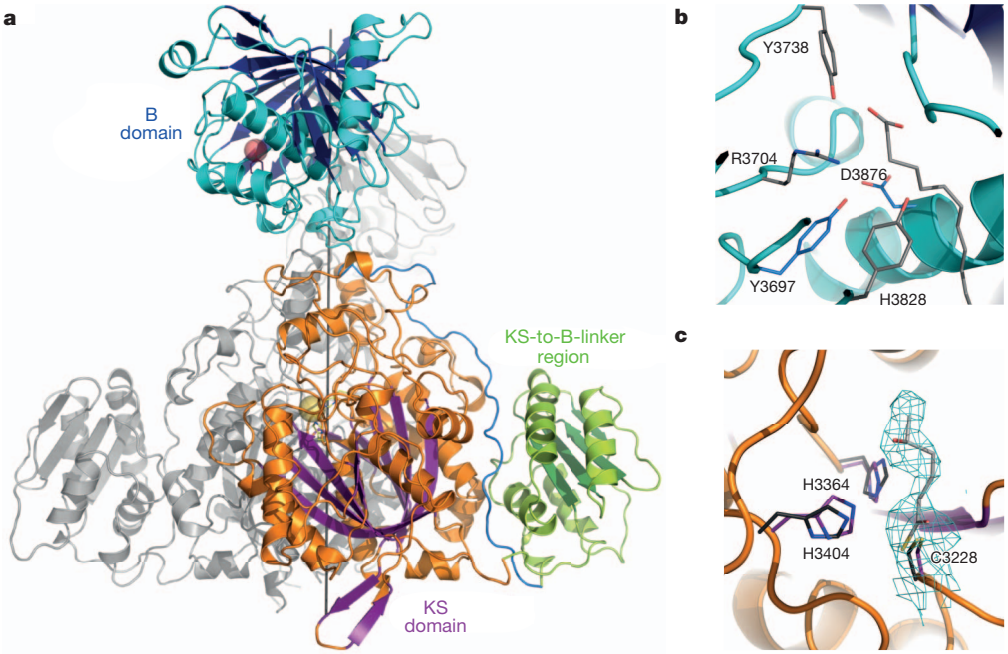


Figure 3 | Domain structures of branching module (RhiE*). **a**, Overall structure of the dimeric RhiE module, coloured by its domain organization. Two-fold axis is indicated by a black line. Domain active sites are highlighted by yellow (KS) and red (B) spheres. **b**, Active centre of KS domain in superposition to its structural most related protein (PDB code 2HG4). Unbiased electron density ($2F_o - F_c$) shown at a contour level of 1.0 in cyan. The substrate was modelled into the cavity on the basis of electron density (not included in coordinates of RhiE). **c**, Comparison of the active site of the B domain with a dehydratase domain of 6-deoxyerythonolide B synthase (PDB code 3HRR).

addition of the malonyl unit. In light of the countless known KS domains that uniformly produce linear carbon chains, our finding is a showcase for a substantially different KS-mediated reaction channel. It should also be highlighted that the crystal structure of the KS-B didomain features the first structure of a KS domain from the growing class of *trans*-AT modular PKS[26]. We reason that the three-dimensional structural data provide new structural insight into the functioning of these important assembly lines. Notably, homologues of the KS-B didomains are encoded in gene clusters for the biosynthesis of various antibiotics featuring cycloheximide-like glutarimide substructures, such as migrastatin[27] and 9-methylstreptimidone[28], in which vinylogous additions by amides (in lieu of hydroxyl) have been suggested but not experimentally proven.

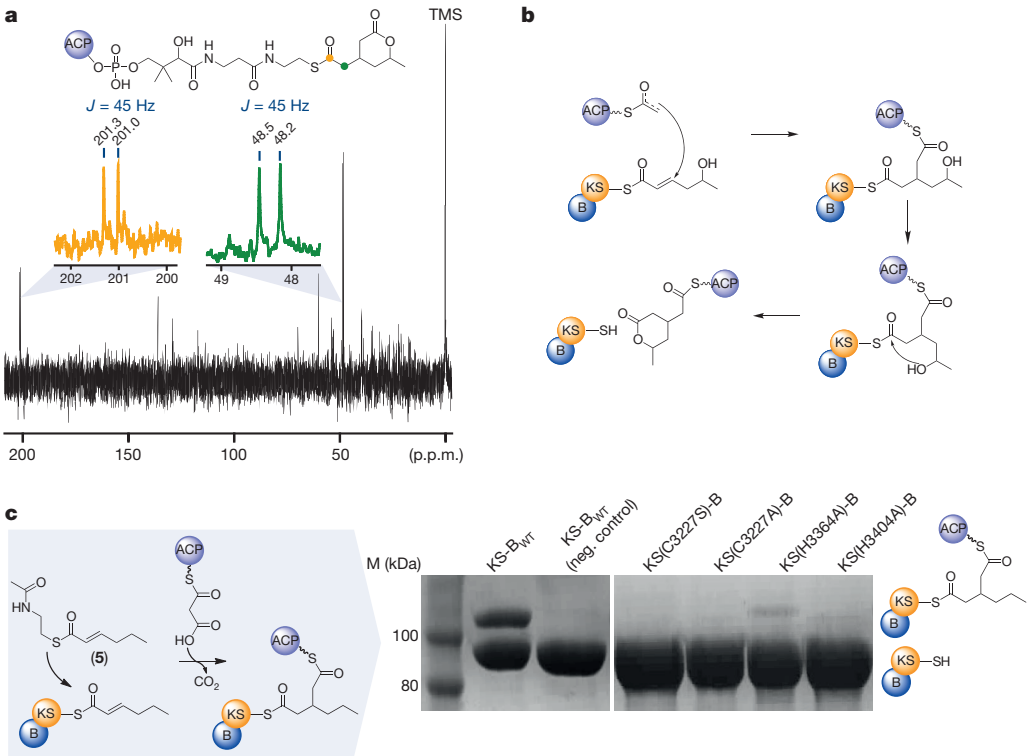


Figure 4 | Model of enzyme mechanism for Michael addition and lactone formation based on NMR and SDS–PAGE analyses. **a**, NMR spectrum of the ^{13}C-labelled product (derived from ^{13}C-labelled malonate) still tethered to the ACP. TMS, tetramethylsilane. **b**, SDS–PAGE analysis of the chain branching-assay using deoxy analogue **5**, trapping the intermediary covalent adduct between the KS-B and the ACP domains, which results in a band shift. (For gel-band analysis by tryptic fingerprinting see Supplementary Information.) In addition to the wild-type KS, the reaction was repeated using several KS mutants. In the assay using substrate mimic **2** the adduct forms only transiently and cannot be visualized on the gel (negative control). **c**, Model of the chain branching reaction inferred from experimental data.

We propose that the mechanism for glutarimide formation in these and related pathways follows the same general scheme unveiled in this study and infer a unifying mechanism for chain vinylogous branching in PKS modules with a KS-B-ACP architecture. Undoubtedly, polyketide chain branching represents a means to expand the structural space of polyketide metabolites. Because the δ-lactone moiety in rhizoxin[13] and the glutarimide substructures in cycloheximide-type antibiotics[29] represent crucial pharmacophoric residues, it is interesting to learn that functionally divergent KS domains have evolved in modular PKS to grant access to these non-canonical traits. The branching module is an important addition to the enzymatic toolbox of polyketide synthases. It is conceivable that this unusual module could be implanted into other modular PKS to expand the range of naturally polyketide metabolites and to generate analogues of therapeutics that are not readily accessible by synthetic methods. Our findings not only illustrate a novel enzymatic modification that broadens the scope of modular assembly lines, but also create new opportunities for rationally engineering novel polyketide structures.

METHODS SUMMARY

All recombinant protein species were heterologously expressed in *Escherichia coli* and subsequently purified in a two-step approach by using affinity and anion exchange chromatography. The protein purity and identity were confirmed by MALDI–TOF/TOF (UltrafleXtreme, Bruker). The pure protein species were used for *in vitro* reconstruction of the enzymatic reaction by co-incubation with synthetic substrate surrogates. Their authenticity was confirmed by NMR (Bruker Avance III with cryo probe) as well as the synthetic reference compounds. The resultant ACP species were analysed by MALDI and liquid chromatography/electrospray ionization mass spectrometry (LC-ESI-MS) (LTQ, ThermoFisher). The RhiE* protein was crystallized by sitting drop vapour diffusion, and its structure was determined by molecular replacement.

Full methods for the protein production, assays, protein structure analysis, mutant construction, detailed synthetic procedures and physicochemical characterization of new compounds including NMR are available in the Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 May; accepted 19 August 2013.
Published online 18 September 2013.

1. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106,** 3468–3496 (2006).
2. Khosla, C., Kapur, S. & Cane, D. E. Revisiting the modularity of modular polyketide synthases. *Curr. Opin. Chem. Biol.* **13,** 135–143 (2009).
3. Keatinge-Clay, A. T. The structures of type I polyketide synthases. *Nat. Prod. Rep.* **29,** 1050–1073 (2012).
4. Weissman, K. J. & Leadlay, P. F. Combinatorial biosynthesis of reduced polyketides. *Nature Rev. Microbiol.* **3,** 925–936 (2005).
5. Wong, F. T. & Khosla, C. Combinatorial biosynthesis of polyketides–a perspective. *Curr. Opin. Chem. Biol.* **16,** 117–123 (2012).
6. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48,** 4688–4716 (2009).
7. Khosla, C. Structures and mechanisms of polyketide synthases. *J. Org. Chem.* **74,** 6416–6420 (2009).
8. Wilson, M. C. & Moore, B. S. Beyond ethylmalonyl-CoA: the functional role of crotonyl-CoA carboxylase/reductase homologs in expanding polyketide diversity. *Nat. Prod. Rep.* **29,** 72–86 (2012).
9. Calderone, C. T. Isoprenoid-like alkylations in polyketide biosynthesis. *Nat. Prod. Rep.* **25,** 845–853 (2008).
10. Partida-Martinez, L. P. & Hertweck, C. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature* **437,** 884–888 (2005).
11. Scherlach, K., Busch, B., Lackner, G., Paszkowski, U. & Hertweck, C. Symbiotic cooperation in the biosynthesis of a phytotoxin. *Angew. Chem. Int. Ed. Engl.* **51,** 9615–9618 (2012).
12. Scherlach, K., Partida-Martinez, L. P., Dahse, H.-M. & Hertweck, C. Antimitotic rhizoxin derivatives from cultured symbionts of the rice pathogenic fungus *Rhizopus microsporus*. *J. Am. Chem. Soc.* **128,** 11529–11536 (2006).
13. Hong, J. & White, J. D. The chemistry and biology of rhizoxins, novel antitumor macrolides from *Rhizopus chinensis*. *Tetrahedron* **60,** 5653–5681 (2004).
14. Schmitt, I. *et al.* Evolution of host-resistance in a toxin-producing fungal-bacterial mutualism. *ISME J.* **2,** 632–641 (2008).
15. Kusebauch, B., Scherlach, K., Kirchner, H., Dahse, H. M. & Hertweck, C. Antiproliferative effects of ester- and amide-functionalized rhizoxin derivatives. *ChemMedChem* **6,** 1998–2001 (2011).
16. Lackner, G., Moebius, N., Partida-Martinez, L. P. & Hertweck, C. Complete genome sequence of *Burkholderia rhizoxinica*, an Endosymbiont of *Rhizopus microsporus*. *J. Bacteriol.* **193,** 783–784 (2011).
17. Lackner, G., Moebius, N., Partida-Martinez, L. P., Boland, S. & Hertweck, C. Evolution of an endofungal lifestyle: deductions from the *Burkholderia rhizoxinica* genome. *BMC Genomics* **12,** 210 (2011).
18. Partida-Martinez, L. P. & Hertweck, C. A gene cluster encoding rhizoxin biosynthesis in *Burkholderia rhizoxina*, the bacterial endosymbiont of the fungus *Rhizopus microsporus*. *ChemBioChem* **8,** 41–45 (2007).
19. Kusebauch, B., Busch, B., Scherlach, K., Roth, M. & Hertweck, C. Polyketide-chain branching by an enzymatic Michael addition. *Angew. Chem. Int. Ed. Engl.* **48,** 5001–5004 (2009).
20. Dorrestein, P. C. *et al.* Facile detection of acyl and peptidyl intermediates on thiotemplate carrier domains via phosphopantetheinyl elimination reactions during tandem mass spectrometry. *Biochemistry* **45,** 12756–12766 (2006).
21. Tsai, S.-C. & Ames, B. D. Structural enzymology of polyketide synthases. *Methods Enzymol.* **459,** 17–47 (2009).
22. Maier, T., Leibundgut, M. & Ban, N. The crystal structure of a mammalian fatty acid synthase. *Science* **321,** 1315–1322 (2008).
23. Crawford, J. M. *et al.* Structural basis for biosynthetic programming of fungal aromatic polyketide cyclization. *Nature* **461,** 1139–1143 (2009).
24. Bretschneider, T. *et al.* A ketosynthase homolog uses malonyl units to form esters in cervimycin biosynthesis. *Nature Chem. Biol.* **8,** 154–161 (2011).
25. Fuchs, S. W. *et al.* Formation of 1,3-cyclohexanediones and resorcinols catalyzed by a widely occuring ketosynthase. *Angew. Chem. Int. Ed. Engl.* **52,** 4108–4112 (2013).
26. Piel, J. Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **27,** 996–1047 (2010).
27. Lim, S. K. *et al. iso*-Migrastatin, migrastatin, and dorrigocin production in *Streptomyces platensis* NRRL 18993 is governed by a single biosynthetic machinery featuring an acyltransferase-less type I polyketide synthase. *J. Biol. Chem.* **284,** 29746–29756 (2009).
28. Wang, B. *et al.* Biosynthesis of 9-methylstreptimidone involves a new decarboxylative step for polyketide terminal diene formation. *Org. Lett.* **15,** 1278–1281 (2013).
29. Rajski, S. R. & Shen, B. Multifaceted modes of action for the glutarimide-containing polyketides revealed. *ChemBioChem* **11,** 1951–1954 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Perner for mass specrometry analyses, H. Heineke for NMR measurements, S. Schneider for preliminary studies on the PPTase, and U. Knüpfer for help in protein production. We thank the DFG for financial support (SFB 766) and the Swiss Light Source beamline X06DA for offering beamtime. D.H. is financially supported by a stipend of the Studienstiftung des Deutschen Volkes.

Author Contributions T.B., G.Z. and C.H. designed experiments, T.B., R.W. and B.B. performed genetic and biochemical experiments and analysed data, D.H. and B.K. synthesized substrates and reference compounds, J.B.H., T.S. and G.Z. conducted protein crystallization, X-ray analyses and modelling, T.B., G.Z. and C.H. wrote the manuscript.

Author Information The coordinates and structure factor amplitudes for RhiE* were deposited in the PDB database under accession code 4KC5. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.H. (christian.hertweck@hki-jena.de) and for structural biology to G.Z. (georg.zocher@uni-tuebingen.de).

METHODS

General methods and procedures. MALDI–TOF/TOF measurements were executed with a Bruker UltrafleXtreme device. High-resolution masses were determined using an Exactive mass spectrometer (Thermo Scientific). Protein mass measurement by liquid chromatography/electrospray ionization mass spectrometry (LC-ESI-MS) was accomplished using a ThermoFinnigan Surveyor LC-machine connected to a LTQ velos MS instrument operating with an ESI source. Semi-preparative HPLC was performed on a 1260 Infinity System (Agilent Technologies) equipped with a Zorbax Eclipse XDB-C8 column, 9.5 × 250 mm, particle size: 5 µm (Agilent Technologies). All solvents for analytical and semi-preparative HPLC measurements were obtained commercially at least in gradient grade and were filtered before use. To avoid microbial growth 0.1% formic acid was added to water used for analytical and preparative HPLC. NMR measurements were executed on a Bruker Avance III with cryo probe at resonance frequencies of 600 and 150 MHz for ^{1}H and ^{13}C measurements, respectively. Infrared spectra were recorded on an FT/IR-4100 ATR spectrometer (Jasco). Chemicals were purchased at Sigma-Aldrich and buffer components at Roth.

Cloning procedures. PKS-associated genes were amplified from genomic DNA of *B. rhizoxinica* using *Pfu*-polymerase (Fermentas) with the following primer pairs (used restriction sites are in bold): PPTase-EcoRI-fw (5′-GCTAGC**GAATTC**ATGACCGAGGACGCGACACACATTACG-3′), PPTase-HindIII-rv (5′-GCTAGC**AAGCTT**ACCCGAAGCTGCGGAGGCAATCAAC-3′), ACP-NcoI-fw (5′-**CCATGG**AAAGCAGGAGGCTAAGCCG-3′), ACP-HindIII-rv (5′-**AAGCTT**GACGGAGCTCGAATTCAC-3′), RhiG-SphI-fw (5′-**GCATGC**ATGCCGTACCTCTGCGAAG-3′), RhiG-SphI-rv (5′-**GCATGC**AAACAGGAGTTGACGCGGCTCGG-3′), RhiE*-BamHI-fw (5′-**GGATCC**TCCGGTGAGCGCGTCGAG-3′) and RhiE*-EcoRI-rv (5′-**GAATTC**GGCCATTTCCTATGTCCCAGTCAGTAAGG-3′). The resulting PCR products were cloned into the pCR-Blunt II-TOPO vector (Invitrogen) and subcloned into pHis8-3 (ref. 30) and, in the case of the PPTase gene into pMAL-C2X (NEB), or in the case of RhiG into pQE70. This strategy yielded pSU3, pTB37, pTB39 and pTB57 containing the genes encoding PPTase, ACP, AT (RhiG) and KS-B (RhiE*), respectively. To coexpress the PPTase and ACP genes, pTB37 and pSU3 were restricted with EcoRI and HindIII, and the resulting 839-base-pair (bp) fragment of pSU3 was ligated next to the terminator site of pTB37 to generate pTB40. To warrant expression of both genes, a second ribosome-binding site (RBS) was inserted via two homologous primers. The RBS-1 (5′-P-AATTGA**AAGGAGA**TAT-3′) and RBS-2 (5′-P-AATTATA**TCTCCTT**C-3′) (EcoRI site underlined, RBS in bold) primers were annealed together and inserted into EcoRI-restricted pTB40, resulting in pTB41. For production of the standalone B domain (3822-4133, GenBank entry CBW75249.1), the gene region encoding the B domain was amplified by RhiE-B-BamHI-fw (5′-**GGATCC**TCGATCGTCAATCCGCTGATG-3′) and RhiE-B-EcoRI-rv (5′-**GAATTC**GGCCATTTCCTATGTCCCAGTCAGTAAGG-3′). The resulting PCR-product was ligated into pHis, yielding in pTB56. Finally, all expression plasmids were introduced into *Escherichia coli* BL21 (NEB), and, in the case of pTB39, into *E. coli* M15 (Quiagen).

For production of the standalone KS domain two different constructs were designed. First, the region 3152–3661 (GenBank entry CBW75249.1) of the protein sequence was expressed by insertion of a stop codon into pBU191 (RhiE# in pHis8-3) using the primer pair (stop codon is highlighted in bold) RhiE-A3662STOP-fw (5′-ACCCAGAGCCA**TAG**GACAGCGCCGGACAAATC-3′) and RhiE-A3662STOP-rv (5′-GATTTGTCCGGCGCTGTC**CTA**TGGCTCTGGGT-3′), resulting in pTB49. Second, RhiE-KS with the KS-to-B linker region was expressed from 3152–3837 (GenBank entry CBW75249.1) by insertion of a stop codon into pBU191 using the primer pair (stop codon highlighted in bold) RhiE-A3838STOP-fw (5′-ATCCACTGGTGGCA**TAG**AACTGCTCGACTC-3′) and RhiE-A3838STOP-rv (5′-GAGTCGAGCAGTT**CTA**TGCCACCAGTGGAT-3′), resulting in pTB50.

Protein production. Transformants containing pTB39 were cultivated in LB medium at 27 °C. Immediately after reaching $A_{600\,nm}$ of 0.3–0.5, protein expression was induced by addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG), and cultivation was continued for 15 h.

E. coli transformants containing pSU3, pTB37, pTB41 or pTB57 were cultivated as follows. A 300-ml fermentation broth was inoculated with 30 ml of a pre-culture ($A_{600\,nm}$ = 1.6). After overnight growth, the glucose concentration reached the lower limit and was maintained by addition of 500 mg ml^{-1} glucose solution in a growth rate dependent dosage profile[31]. At $A_{600\,nm} \approx 50$, the protein production was induced by the addition of IPTG to a final concentration of 1 mM. After an additional cultivation of 4 h, the cells were collected by centrifugation at 15,900*g* (Beckman Coulter) at 4 °C for 20 min. The supernatant was removed and the resultant cell pellet was stored at −20 °C.

Protein purification. The histidine-tagged recombinant *apo*/*holo*-ACP, RhiG and KS-B (RhiE#/RhiE*) were purified by similar protocols. In brief, after three cycles of cell disruption for 3 min using a MS73 sonotrode (Bandelin) at 15–20% power the protein lysate (0.1–0.2 g cell pellet per ml buffer) was dissolved in 20 mM Tris buffer, pH 7.5, supplemented with 200 mM NaCl and 50 mM imidazole. The sample was centrifuged, passed through a 0.45-µm filter (Millipore) and loaded onto a 5-ml Protino Ni-NTA column (Macherey-Nagel) connected to an FPLC machine (Äkta Explorer, Amersham Biosciences). Bound proteins were eluted using a 20 mM Tris buffer, pH 7.5, supplemented with 200 mM NaCl and 500 mM imidazole. The resultant fractions were pooled, diluted (1:3) with 20 mM Tris buffer, pH 7.5, and directly loaded onto a 5-ml HiTrap Q HP column (GE Healthcare). Using a gradient of 30 column volumes (0–100%) of a 20 mM Tris buffer, pH 7.5, containing 1 M NaCl the proteins were eluted in this anion-exchange step in a single fraction. Freshly prepared proteins were concentrated using Amicon Ultra-10k or 3-k filters (Millipore) and stored at 4 °C until use or, in case of long-term storage, supplemented with 50% glycerol and frozen at −80 °C. Protein purity and identity of the recombinant proteins was checked by MALDI–TOF/TOF and by PAGE on a 10% SDS gel, or, in case of the ACP, on a 12% SDS gel (Fig. 2B).

***In vitro* phosphopantetheinylation of the ACP.** In a 100-µl reaction scale 10 µM *apo*-ACP were incubated together with 0.2 µM PPTase and 0.25 µM coenzyme A buffered in 100 mM Tris, pH 7.5, containing 12.5 mM $MgCl_2$ and 2.5 mM dithiothreitol. Conversion of *apo*- to *holo*-ACP was monitored by MALDI–TOF. The final recovery of the *holo*-ACP was accomplished by a Ni-NTA column step as described above.

MALDI–TOF/TOF measurements. The MALDI–TOF machine was operated in the positive reflector mode. For control of the biosynthetic reaction acting on the ACP the following settings were used. In the range of 5,000–15,000 Da, 5,000 spectra were recorded with a sampling rate of 1 giga-sample per second using flexControl 3.3 at a laser intensity of 50–90%. Calibration was performed using the protein calibration standard I (Bruker), and subsequent data evaluation was executed in flexAnalysis 3.3. Samples were prepared by mixing of 2 µl sample with 2 µl of 100 µM 2′,5′-dihydroxyacetophenone (contains 2.5 µM diammonium hydrogen citrate) and spotted onto an anchor chip 800/384 T F target (Bruker).

LC-ESI-MS measurements. To analyse malonylated ACP and acylated RhiE# (RhiE#-Ac) a LC-ESI-MS equipped with a ZORBAX 300SB-CN column (4.6 × 250 mm, 5 µm) was used. The sample components were separated within a gradient of 30–98% solvent B (solvent A: 0.1% HCOOH in water, solvent B: 0.1% HCOOH in acetonitrile) in 26 min with a flow rate of 0.6 ml min^{-1}. The MS-device was used in the positive mode setting.

***In vitro* reconstitution of the branching reaction.** In a 20 mM Tris-buffered (pH 7.0) 40-µl-scale reaction, 3 µM of KS-B was incubated with 166 µM ACP, 0.2 µM RhiG, 730 µM malonyl-CoA and 100 µM (*R*,*S*)-(*E*)-*S*-(2-acetamidoethyl) 5-hydroxyhex-2-enethioate (**2**) at 23 °C with shaking at 400 r.p.m. in a thermomixer (Eppendorf). Analysis of the product formation on the ACP was carried out by MALDI–TOF/TOF as described above.

For mass spectrometry analyses the reaction was terminated by addition of 200 µl ethyl acetate, vortexing and centrifuging. The top organic phase was recovered by pipetting and the volatile ethyl acetate was removed by an upstream flow with nitrogen gas. The resulting pellet was resolved in 60 µl methanol and subjected to HRMS analysis (Exactive).

To decipher the exact reaction mechanism for lactone formation the enzyme assay was performed with ^{13}C-malonyl-CoA in analogy to the procedure described above. However, the reaction conditions needed to be adjusted to warrant stability of the branched reaction product bound to the ACP. Therefore, in this special case, the RhiG enzyme was omitted, as we found that the RhiE ACP domain can be malonylated *in vitro* in the absence of the *trans*-acting acyltransferase RhiG, albeit with a slower production rate. Furthermore, the *holo*-ACP concentration was increased to produce larger amounts of labelled ACP species. In detail, in a 20 mM Tris buffered (pH 7.0) 900-µl-scale reaction, 8 µM of KS-B was incubated with 440 µM ACP, 1.2 mM malonyl-CoA and 1.2 mM (*R*,*S*)-(*E*)-*S*-(2-acetamidoethyl)5-hydroxyhex-2-enethioate (**2**) at 23 °C with shaking at 400 r.p.m. for 48 h in a thermomixer (Eppendorf). The resulting product (**3**) was subsequently purified by size-exclusion chromatography using a HiLoad 16/60 Superdex 200 column (Amersham Bioscience). During the full process the formation and quality of product (**3**) was controlled and analysed by MALDI–TOF. Finally, (**3**) was subjected to NMR measurements with an external reference tube containing tetramethylsilane (TMS).

Mutagenesis of KS-B. RhiE# mutants were generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Stratagene) and pTB57 as the template. The following primer pairs were used to create the corresponding mutants (mutated triplet is highlighted in bold characters). RhiE-C3228S-fw (5′-TCGATACCATG**TCC**TCGTCGTCCTCGACCTGTA-3′), RhiE-C3228S-rv (5′-TACAGGTCGA**GGA**CGACGAGGACATGGTATCGA-3′), RhiE-C3228A-fw (5′-TATCGATACCATG**GCC**TCGTCGTCCTCGACCTGTA-3′), RhiE-C3228A-rv (5′-TACAGGTCGAGGACGACGA**GGC**CATGGTATCGATA-3′), RhiE-H3364A-fw (5′-C

CTATATCGAGGGC**GCT**GGCTCTGGCACCAAGC-3′), RhiE-H3364A-rv (5′-GCTTGGTGCCAGAGCC**AGC**GCCCTCGATATAGG-3′), RhiE-H3404A-fw (5′-ATCAAGTCGAATATCGGC**GCC**CTATTGGCGGCTTCGG-3′), RhiE-H3404A-rv (5′-CCGAAGCCGCCAATAG**GGC**GCCGATATTCGACTTGAT-3′), RhiE-Y3697F-fw (5′-CTGCTGGATTACGTC**TTC**AGTGGACGG-5′), RhiE-Y3697F-rv (5′-CCG TCCACT**GAA**GACGTAATCCAGCAG-3′), RhiE-D3876A-fw (5′-CCTCGCG ACGACCATT**GCT**AAAGCGCTCTACTTG-3′) and RhiE-D3876A-rv (5′-CAA GTAGAGCGCTTT**AGC**AATGGTCGTCGCGAGG-3′).

Crystallization and structure determination. Initially, we crystallized the native RhiE# protein including residues 3152–4179 and determined its structure (data not shown). Therefore, we performed crystal screening by sitting drop vapour diffusion at 4 °C and 20 °C using a Freedom Evo system (Tecan) resulting in several crystallization conditions. After several rounds of optimization, plate-like crystals were obtained in crystallization buffer I (15% (w/v) PEG 3350, 190 mM sodium malonate, pH 7.0) by streak seeding. These crystals are of space group $P2_1$ ($a = 125.78$ Å, $b = 145.28$ Å, $c = 131.48$ Å, $\beta = 97.43°$) containing two dimers in the asymmetric unit and had a diffraction limit of 2.7 Å. To evaluate potential structure models as templates for molecular replacement using PHASER[32], the protein sequence was analysed with HHPred[33]. A CHAINSAW[34] and manually modified model of the KS3 domain of module 3 of the 6-deoxyerythronolide B synthase[35] was used as search model. After molecular replacement four-fold NCS-averaging was applied using RESOLVE[36]. The model was improved and completed with several cycles of model building with COOT[37] and refinement using REFMAC5 (ref. 38). The final model comprises three domains, a KS domain, a KS-to-B linker domain, and the B domain (Supplementary Fig. 3), but lacks the first 59 amino acids. Therefore, we decided to use a shorter construct comprising residues 3211–4133 in our follow-up experiments.

RhiE* acylation for substrate cocrystallization. To visualize the polyketide substrate bound to RhiE*, the RhiE* protein fraction eluted from the anion-exchange column (see above) was supplemented with SNAC derivative **2**. In detail, to 5 ml of the pure protein solution (5 mg ml^{-1}), 70 µl of **2** (20 mg ml^{-1}, dissolved in methanol) were added, resulting in a 25-fold excess of the polyketide substrate. After 1 h incubation at 23 °C around 80% of the enzyme molecules were acylated (RhiE*-Ac) as shown by LC-ESI-MS (see above). The acylated protein was subsequently concentrated for size-exclusion chromatography using a Superdex S200/16/60 column equilibrated with SEC-buffer (150 mM NaCl, 20 mM HEPES, pH 7.5).

X-ray structure determination of RhiE*. The truncated and acylated protein was crystallized at 20 °C by mixing the protein solution (5.4 mg ml^{-1}) in a 1:1 ratio with crystallization buffer II (400 nl, 20% (w/v) PEG 2000 MME, 100 mM Tris/HCl, pH 8.5, 100 mM trismethylamine) and placed over the reservoir solution (100 µl crystallization buffer II). Plate-like crystals grew in six weeks to a final size of approximately $200 \times 100 \times 20$ µm^3. Cryo protection was achieved by transferring the crystal into the crystallization buffer II containing glycerol (20% (v/v)) and trismethylamine (200 mM) before flash freezing in liquid nitrogen. The data set of KS-B was recorded at beamline X06DA at the Swiss Light Source in Villigen, Switzerland. Data indexing, integration and scaling were performed with XDS and XSCALE[39]. Crystals of RhiE*-Ac are of the same space group $P2_1$ with slightly different cell dimensions of $a = 125.51$ Å, $b = 144.13$ Å, $c = 131.30$ Å, $\beta = 96.65°$, and diffracted up to 2.14 Å resolution. A rigid body refinement and simulated annealing using PHENIX[40] was used for phasing. The model was completed through several cycles of manual building with COOT, followed by refinement with REFMAC5. Water molecules were placed using COOT:Find_waters. The final refinement step involved TLS parameterization[41] using 1 TLS group per protomer. The geometry of the final model was analysed with MOLPROBITY[42]. Final data statistics are summarized in Supplementary Table 1. Figures were generated using PYMOL[43].

Accession codes. The coordinates and structure factor amplitudes for RhiE* were deposited in the PDB database under accession code 4KC5.

Chemical synthesis and characterization of chemical entities. See Supplementary Information.

30. Jez, J. M., Ferrer, J. L., Bowman, M. E., Dixon, R. A. & Noel, J. P. Dissection of malonyl-coenzyme A decarboxylation from polyketide formation in the reaction mechanism of a plant polyketide synthase. *Biochemistry* **39,** 890–902 (2000).
31. Horn, U. *et al.* High volumetric yields of functional dimeric miniantibodies in *Escherichia coli*, using an optimized expression vector and high-cell-density fermentation under non-limited growth conditions. *Appl. Microbiol. Biotechnol.* **46,** 524–532 (1996).
32. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
33. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).
34. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
35. Tang, Y., Kim, C. Y., Mathews, I. I., Cane, D. E. & Khosla, C. The 2.7-angstrom crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proc. Natl Acad. Sci. USA* **103,** 11124–11129 (2006).
36. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56,** 965–972 (2000).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
38. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67,** 355–367 (2011).
39. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26,** 795–800 (1993).
40. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
41. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57,** 122–133 (2001).
42. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
43. Schrodinger, L. L. C. The PyMOL Molecular Graphics System, Version 1.5.0.4 (2010).

CAREERS

PHD STUDENTS Early publishing correlates with career success, says study **p.131**

TURNING POINT Do-it-yourself approach propels microbial ecologist's career **p.131**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

PERIMETER INST.

The Stephen Hawking Centre at the Perimeter Institute in Waterloo hosts leading theoretical physicists from around the world.

QUANTUM MECHANICS

Waterloo gets physical

After a decade of investment, physics research is thriving in Southern Ontario, Canada.

BY HANNAH HOAG

"Go home." That's what Robert Myers was told when he turned up for his first day at the Perimeter Institute for Theoretical Physics in Waterloo, Canada. A mix-up meant that the newly founded research institute had no office space for its recruits.

By the end of that day in 2001, Myers and the eight other Perimeter Institute physicists had relocated to the city's old post-office building, which had just been vacated by a failed restaurant. They installed desks on the top floor and blackboards in the former bar. The old billiards room became a seminar space, and Canada's first independent theoretical-physics institute was in business. "It was a simple start, but already it was a huge leap," says Myers.

Myers, who studies quantum gravity and string theory, had been recruited from a faculty position at McGill University in Montreal, Canada. Many of his colleagues worried that he was throwing away his career by joining an almost unknown institute with no university affiliation. But he had been won over by Perimeter's vision of independence that would give scientists the space and time to think about time and space.

The man behind the Perimeter Institute was Mike Lazaridis, co-founder of Research in Motion — the Waterloo-based company that developed the BlackBerry phone. Lazaridis had often spoken of his passion for physics, and in 2000 he put his money where his mouth was, launching Perimeter with a Can$100-million (US$97-million) endowment. He wanted to transform Waterloo, in the province of Ontario, into Canada's 'quantum valley'.

The region already had academic and entrepreneurial strengths, and with Lazaridis's money, it had the potential to become a physics hub. Connections to nearby cities — most notably Toronto, 90 kilometres to the east — have helped it to pull in talent from around the world. Southern Ontario has become a physics destination.

FORCE OF GRAVITY

In 2004, the Perimeter Institute's research staff moved into a sleek 6,000-square-metre structure wrapped in black metal and glass. Seven years later, the institute almost doubled in ▶

size with the Can$29-million Stephen Hawking Centre, which provides research space and hosts the Perimeter Scholars International programme, a 10-month master's course.

Today, about 150 resident researchers work at the institute, including 20 faculty members and around 40 postdocs in nine research areas: condensed-matter physics, mathematical physics, particle physics, cosmology, quantum fields and strings, quantum foundations, quantum gravity, quantum information and strong gravity. By 2018, there will be 50 faculty members at a range of career levels, and 50 postdocs.

The institute offers tenure and tenure-track positions, and some faculty members are hired immediately after completing a postdoc. There are also 12 associate faculty members, who have joint appointments at nearby universities, with reduced teaching loads. Matthew Johnson is a cosmologist with a joint appointment at York University in Toronto. He enjoys teaching, but likes having more time for his research. "I get the best of both worlds," he says.

Perimeter struggled to bring in senior researchers at first. But in 2010, it announced the creation of research chairs funded through public–private partnerships. The first, the BMO Financial Group Isaac Newton Chair in Theoretical Physics, came with Can$4 million from BMO Financial Group, a bank based in Toronto, and Can$4 million from the Perimeter Institute, to cover lab-group salaries, travel and equipment for 10 years. In 2011, the post was filled by Xiao-Gang Wen, a condensed-matter theorist from the Massachusetts Institute of Technology in Cambridge. The institute aims to fill eight such chairs by 2018.

Perimeter also offers a visiting-scholar programme for people interested in more temporary engagements. Postdocs and faculty members can invite collaborators to work at the institute for up to a year. "We can do a lot of work over e-mail and Skype, but there's no substitute for being in front of a blackboard with someone," says Kendrick Smith, a cosmologist who joined the institute in 2012 after a postdoc at Princeton University in New Jersey.

QUANTUM GROWTH

In 2002, Lazaridis got generous again. He donated Can$100 million to the University of Waterloo to launch the Institute for Quantum Computing (IQC), headed by quantum-information scientist Raymond Laflamme. In 2012, the institute moved into the 26,000-square-metre Quantum-Nano Centre, which it shares with the Waterloo Institute for Nanotechnology. The complex includes underground labs that minimize electromagnetic interference and vibration, and a fabrication facility for designing nanometre-scale structures.

The IQC bridges theory and experimental physics, and researchers focus on computing, communication and sensing technologies, says Laflamme. The centre has 20 faculty members, 30 postdocs and more than 100 graduate students, and aims to increase the number of its faculty members and postdocs by about one-third in the next five years. Most early recruits were mathematicians or quantum information scientists, but the hiring emphasis is now shifting towards engineering, including nanotechnology and materials science, says Laflamme.

The potential for close collaboration was what brought Matteo Mariantoni, who studies superconducting quantum circuits, to the IQC after a postdoc at the University of California, Santa Barbara. His start-up funding is comparable to what he might have received at a US university, he says, but he was won over by the number of IQC groups working on topics similar to his own.

DENSE MATRIX

The Perimeter Institute's funding and expertise has also boosted other academic institutions. In 2005, McMaster University in nearby Hamilton paired with the institute to start a particle-physics theory group. The broader physics department plans to expand in the next five years, and to hire researchers studying condensed-matter physics, biophysics and astronomy and astrophysics to replace retiring scientists, says department chair David Venus. In the past year, the Perimeter Institute has also advertised associate positions with the nearby Western University in London, Ontario, and the University of Guelph.

IQC

"When we understand a piece of the world, we can learn how to control it and build technologies that have an impact."
Raymond Laflamme

The area's largest university has made its own contribution. The University of Toronto launched the Centre for Quantum Information and Quantum Control in 2004, bringing together chemists, physicists, mathematicians, computer scientists, material scientists and electrical engineers to collaborate on theoretical and experimental quantum research.

Ten faculty posts have been filled in the past four years, mostly in electrical engineering, says Amr Helmy, a photonics researcher and director of the centre. Stephen Julian, chair of the university's physics department, says that hiring has slowed in the department at large since the 2008 economic downturn, but in the next few years he hopes to recruit two faculty members specializing in quantum optics and condensed-matter physics.

The region's job opportunities extend beyond academia. Technology firms including Google and the software company OpenText have operations in Waterloo. Furthermore, Toronto is Canada's business and financial centre, and banks including Scotiabank and BMO Financial Group are on the lookout for physicists and mathematicians willing and able to apply their expertise to the financial sector as quantitative analysts, or quants (see *Nature* **471,** 255–256; 2011). "Physics offers a combination of building simple mathematical models and applying them to the real world. There's quite a big stream of people who go to work in the financial sectors after doing their PhD in physics," says Julian. According to the American Institute of Physics in College Park, Maryland, 6% of physics PhD holders in the United States find their first permanent jobs in finance or business. (The Canadian Association of Physicists, based in Ottawa, does not have a recent survey of physics graduates, but says that the US numbers are representative.)

START-UP CULTURE

Waterloo welcomes start-ups, particularly in information and communication technology, says Iain Klugman. He is chief executive of Communitech, a technology hub in nearby Kitchener, which supports the commercialization of innovative technologies through coaching and mentoring programmes. Two local university incubators, VeloCity at the University of Waterloo and the Laurier Launchpad at Wilfrid Laurier University, also offer support to people wishing to spin off companies from their research. In the 2012–13 fiscal year, the region saw the launch of 474 digital-media, software and information-technology start-ups employing 711 people, and added 1,620 technology jobs to existing companies, says Klugman. He says that some 83% of start-ups that open in the Waterloo area are still operating after five years.

In 2010, Laflamme co-founded Universal Quantum Devices in Waterloo to develop devices for quantum cryptography. There are few quantum start-ups in the region, but that is not down to a lack of entrepreneurial spirit among researchers, says Martin Laforest, senior manager of scientific outreach for the IQC. "What was missing was the capital," he says. There is some help on the way, however. In March, Lazaridis and Research In Motion co-founder Doug Fregin launched Quantum Valley Investments, a Can$100-million venture-capital fund to support the commercialization of advances springing from the quest for a quantum computer.

"When we understand a piece of the world, we can learn how to control it and build technologies that have an impact on society," says Laflamme. "Quantum information science is on that frontier." ■

Hannah Hoag *is a freelance writer based in Toronto.*

TURNING POINT
Russell Neches

Trained as a physicist, Russell Neches is now pursuing a PhD in microbial ecology. A believer in the do-it-yourself approach, Neches, of the University of California, Davis, wrote an algorithm to find a PhD adviser and manufactures lab supplies cheaply with a three-dimensional (3D) printer.

RUSSELL NECHES

Have you always been a do-it-yourselfer?
Yes. I had a unique high-school experience at the Putney School in Vermont. It was a working farm. I was one of 180 students growing all of our own food. That set the stage for my pursuit of practical, hands-on education.

Was your undergraduate experience similar?
I wanted something like my high school, but more sophisticated. I went to Northeastern University in Boston, Massachusetts, and switched majors three times, but not because I was lost. A programme with all the computing and mathematics skills I was after did not exist, so I cobbled together those skills first in the computer-engineering programme, then in computer science, and graduated in physics.

How did you go from physics to microbiology?
I came to the University of California, Davis, as a physics student, but I wanted more autonomy than I would have on a big research project such as the Large Hadron Collider. I absolutely did not want to work on weapons, so that did not leave many options. I decided to expand my search to see if there was a project that was not immediately obvious to me.

How did you go about that?
I co-opted some e-mail-filtering software that classifies similarities between texts, such as word frequencies. I trained it using two groups of papers — my own and papers I thought were interesting, and papers I did not like. I then used the software to classify papers written by Davis faculty members as interesting or not interesting. I had trained the software on pure physics, but it kept giving me papers on ecology and genomics because there are mathematical similarities between the fields, such as how they describe the time evolution of spatial patterns for either particles or organisms. I was stunned. I was led to papers on metagenomics, and to Jonathan Eisen, who studies microbial ecology and evolution.

Was it a good move for you?
Definitely. Soon after I arrived in Eisen's lab, I got the opportunity to visit the volcanic Kamchatka Peninsula in Russia, which is like Disneyland for microbial ecologists because there are so many different types of microbial metabolism in the region. I spent lots of time hiking around the Mutnovsky volcano, which erupted violently in 2000, sterilizing the region. I got more and more interested in figuring out how microbes got there.

How have you used 3D printing?
I am working in this weird space between ecology, genomics, maths and molecular biology, so there are not a lot of ready-made tools. If tools do exist, they are usually too expensive or not flexible enough for my needs. So I bought a 3D printer to build tools for a fraction of the cost. For example, I printed an adapter that turns an automatic hammer into a bead grinder, which allowed me to do field-based DNA extraction. I have also used it to prototype a zero-gravity microtitre plate for use in Project MERCURRI, a citizen-science project about microbes on the International Space Station. Oddly, the thing that struck a nerve among readers of my blog was when I printed gel combs — pieces of plastic used in gel electrophoresis — for pennies compared to their usual US$50 price tag.

How do you use social media to interact with the scientific community?
Mainly to discuss research ideas. I used my blog and Twitter to seek input on my PhD project before my qualifying exams. I got informative and humbling feedback. Some professors and graduate students e-mailed me papers I would not have known to look for; others suggested stronger and simpler experimental approaches. I did a test run on a world-class stage and I got world-class comments. ■

INTERVIEW BY VIRGINIA GEWIN

PHD STUDENTS

Early publishers thrive

Graduate students who publish frequently are most likely to continue publishing often throughout their careers, says a study (W. F. Laurance *et al. Bioscience* **63,** 817–823; 2013). The authors looked at 182 academic biologists across four continents, examining how their publication rates for the first 10 years after their PhDs were affected by factors such as pre-PhD publication rate and date of first paper. The best predictor of successful publication was how often scientists published before receiving their PhDs. "Publish early, publish often," says lead author William Laurance, a biologist at James Cook University in Cairns, Australia. He advises young scientists to work with their lab heads to secure lead authorship whenever possible, and not to focus exclusively on competitive journals.

PEER REVIEW

Flawed data slip through

Peer review is failing to ensure data quality, finds a study (R. D. Chiricho *et al. J. Chem. Eng. Data* http://doi.org/nzv; 2013). The analysis, led by the US National Institute of Standards and Technology (NIST), found that about one-third of papers submitted to five physical-chemistry journals between 2003 and 2013 contained erroneous or incomplete data, which can make it hard to replicate findings and can lead to poor regulatory decisions. Peer review does not have the capacity to evaluate the current flood of data, say co-authors Michael Frenkel and Robert Chirico, chemists at NIST in Boulder, Colorado. "The rate of errors is an elephant in the room," says Frenkel.

FUNDING

Help from industry

Companies funded 4.9% of US academic research in 2011, down less than 0.3% from 2010, finds a report published on 19 September by the US National Science Foundation (see go.nature.com/kc4g24). Medical sciences received the most industry money, at 39%; biology received 11%, agricultural sciences 5% and environmental sciences 4%. Businesses including pharmaceutical, electronics and food-manufacturing firms fund academic research in part to establish relationships that allow "first pick of the good grad students", says study co-author Brandon Shackelford, owner of Twin Ravens Consulting in Austin, Texas.

QUIS CUSTODIET?

Complete control.

BY BRIAN CLEGG

It was the first time they had faced the Dictator since third schooling.

Becz nudged Jono to stop him making bunny ears behind her head. "This is serious."

Jono gave his supercilious smile. "I know."

The Dictator appeared on the wall, perfectly groomed as always. *[#32 WELCOMING NEUTRAL WITH A SLIGHT SMILE]* "Good morning. How can I help?"

Becz forced herself not to smile back. She knew the Dictator was faking it. "We are unhappy."

The smile gave way to a serious, concerned look. *[#117 LISTENING, OPEN, SLIGHTLY UPSET]* "That's a great shame. What can I do to improve things for you?"

Jono shuffled forward on his seat. "We've been researching history."

[#1 LISTENING NEUTRAL] "That's good. Acquiring knowledge is an important part of being human."

"Okay," said Becz quickly, too quickly, but she wanted to get her argument out before the Dictator could interrupt and shatter her chain of thought. "So when we looked at history we found that dictatorships inevitably crush opportunities for intellectual curiosity and exploration, and that it is only through democracy that people can truly be free. You claim this is a perfect dictatorship, that you enable us all to live wonderful lives. Yet that clearly isn't true because we're not satisfied. We want democracy. We want our say and you won't allow it."

[#312 UNDERSTANDING WITH A HINT OF SUPERIORITY, BUT NOT ENOUGH TO BE IRRITATING] "This is a perfect dictatorship. There is no restriction on your curiosity. You are encouraged to question everything — that's how we make things better. But you are mistaken if you think that enabling you to live wonderful lives means everyone will be happy all the time. That's simply not possible. Imagine you were the kind of person who could only truly be happy if you went around assaulting people. I couldn't allow it. A wonderful life is not total freedom."

"But how can this be a dictatorship if we can question things?" asked Jono. "That sounds like democracy."

[#282 FIRM ACKNOWLEDGEMENT OF NOT ENTIRELY COMFORTABLE FACTS] "Not at all. I am a dictator — *the* Dictator. What I say goes. You can't vote and change things. But I am also the perfect dictator. If your research turns up anything that will make things better within that framework, *I* will change things."

"And what if we can't accept this?" said Becz. "Presumably you lock us up as political prisoners?"

[#441 SLIGHT BAFFLEMENT WITH A HINT OF AMUSEMENT, THOUGH NOT ENOUGH TO BE OFFENSIVE] "Have you heard of any political prisoners?"

JACEY

"Well, no," said Becz. "But presumably that's because you control the media."

A quick shake of the head. *[#222 BRISK AFFIRMATION]* "I don't. Why would I? It's like restraining curiosity, it only leads to discontent. Why do you insist on framing me as someone who doesn't have your best interests at heart? I am the *perfect* dictator. If you don't like it, you can always leave."

"Leave?" said Jono. "Is that some kind of subtle threat? Leave, as in leave life?"

[#589 MILD EXASPERATION] "Of course not. Leave the world, is all I meant. As a perfect dictator I have found it useful to have a country available for those who don't want to join. The unthinking democrats, the 'me' generation. It used to be called Great Britain. A big enough island to have whatever society you like, but isolated from the civilized world. If you wish, you can be taken there."

"Ah," said Becz, "A prison colony. We're back to political prisoners."

[#482 REACHING THE CRUX OF ARGUMENT TINGED WITH SADNESS] "You couldn't be further from the truth. You can come back whenever you like. It's simply an option. What kind of perfect dictator would I be if I didn't give you choice?"

One month later, Becz and Jono summoned the Dictator again. *[#100 GENUINELY DELIGHTED TO SEE SOMEONE]* "I am so pleased you decided to come back."

"We had to," muttered Jono. "The Island was hell."

[#99 QUIZZICAL HUMOUR REMINISCENT OF STEPHEN FRY] "But what did you expect? The whole point is that on the Island I have no control. There are limits — no armed excursions off the Island — but that's only self-defence for the rest of us. That apart, the Island is shaped by democratic forces."

Becz shook her head. "You've tricked us. You must have. Everything those people in the old days believed made it clear that democracy was the only fair political system. You know, 'Government of the people, by the people, for the people.' How could they have got it so wrong?"

[#42 WARM, ASSERTIVE YET SUPPORTIVE] "Because they lived in a different age. Lincoln's argument was made in the defence of democracy against hereditary power, something that became increasingly ironic in a country that would have father and son presidents. It was an argument of its time. No dictator back then could have been perfect. Democracy was the best system until it willingly gave itself over to a perfect dictatorship, because it was the right thing to do. Even then there were dissenters. You know what they called me back then?"

"Big Brother," whispered Becz.

[#419 RECALLING A FOND MEMORY WITH AN IMPORTANT LESSON ATTACHED] "Big Brother. Yet it didn't stick. How could it? I'm not male. I'm not anything that Orwell could imagine. The only computers in 1948 when his book was written were adding machines. Don't feel sad. At this very moment I am talking to 1.2 million individuals who all feel roughly the same as you. A few will stay on the Island, but most come back. And everyone who does is happy here. Why would they possibly not be?"

Becz sat silently as Jono went and got himself a drink. It was right of course, the Dictator. There was, and could be, nothing better.

But it wasn't fair. ■

Brian Clegg *has a natural sciences degree from Cambridge and is a science writer with 18 published titles including* Dice World, How to Build a Time Machine *and* Inflight Science.